Análisis de Encuestas de Hogares

usando R

Andrés Gutiérrez, Stalyn Guerrero, Giovany Babativa, Cristian Téllez y Xavier Mancer

24 de agosto de 2026

Tabla de contenidos

Listado de Figuras

Listado de Tablas

Prefacio

También puede descargar la versión en PDF aquí: Descargar libro en PDF.

Este libro es el resultado de un compendio de las experiencias internacionales prácticas adquiridas por los autores en su asesoría a los países de América Latina y el Caribe desde la División de Estadísticas de la CEPAL.

Resumen

Las encuestas de hogares constituyen una fuente fundamental de información para medir indicadores sociales y económicos (como la pobreza, la desigualdad, el empleo y la educación) y para monitorear el avance hacia los Objetivos de Desarrollo Sostenible. La Comisión Económica para América Latina y el Caribe (CEPAL) gestiona repositorios de microdatos de encuestas de hogares que reúnen, armonizan y documentan encuestas de más de 18 países de la región desde el año 2000, precisamente porque la información que producen estas fuentes es de vital importancia para la toma de decisiones basada en evidencia.

Sin embargo, en el análisis de las encuestas de hogares existe una realidad técnica estructural que con frecuencia pasa inadvertida: la mayoría de estas encuestas no se basa en muestras simples. Por el contrario, están sustentadas en diseños de muestreo complejos que incorporan estratificación, múltiples etapas, probabilidades de selección desiguales y ajustes de ponderación derivados de la calibración. Cuando un analista aplica los métodos estadísticos tradicionales diseñados para muestras independientes e idénticamente distribuidas sobre datos que no poseen esas características, las consecuencias son estimaciones sesgadas, errores estándar subestimados, intervalos de confianza demasiado estrechos y pruebas de hipótesis con tasas de error infladas.

Este documento busca proporcionar al investigador los fundamentos teóricos y las herramientas computacionales necesarias para trabajar adecuadamente con encuestas provenientes de diseños complejos. A lo largo del texto se abordan de manera integrada aspectos prácticos que van desde la definición del diseño de muestreo hasta la estimación de indicadores (como totales, tamaños, medias y proporciones), la construcción de sus respectivos intervalos de confianza, la formulación de pruebas de hipótesis, la modelación mediante técnicas de regresión avanzadas y la correcta visualización de los datos.

1. Introducción

El análisis de encuestas con diseños complejos cuenta con una trayectoria consolidada en la literatura estadística, sustentada en el enfoque de inferencia basado en el diseño de muestreo. Como señalan Särndal, Swensson, y Wretman (2003) y Gutiérrez (2016), este enfoque reconoce a la distribución de probabilidad inducida por el proceso de selección de la muestra como el único mecanismo probabilístico que rige la inferencia. Bajo este paradigma, las propiedades estadísticas de los estimadores (como el insesgamiento, la precisión y la consistencia) se evalúan con respecto a la distribución de todas las posibles muestras que el diseño puede generar, sin imponer supuestos distribucionales sobre la variable de interés.

En este marco, la estimación de varianzas ha ocupado un lugar central y ha evolucionado en torno a tres grandes familias metodológicas complementarias. La primera, el método del último conglomerado (Hansen et al. 1953), simplifica la estructura multietápica de los diseños de muestreo al concentrar la variabilidad del estimador en las unidades primarias de muestreo (UPM), tratando la selección como si se realizara con reemplazo. La segunda, la linealización de Taylor (Woodruff 1971; David A. Binder 1983), extiende este principio a parámetros no lineales (como medias, proporciones y razones) mediante una aproximación de primer orden que reduce el problema de estimación de la varianza al de un total. Finalmente, los métodos de replicación (Rust y Rao 1996) estiman la varianza a partir de la dispersión entre estimaciones obtenidas mediante réplicas sistemáticas de la muestra.

Paralelamente, en el ámbito de la modelización estadística, contribuciones fundamentales como las de Kish y Frankel (1974), Fuller (1975) y David A. Binder (1983) sentaron las bases de la regresión ponderada y de la estimación de varianzas de coeficientes de regresión en diseños estratificados y de múltiples etapas. Sobre este desarrollo, el método de máxima pseudo-verosimilitud (MPV) (Skinner, Holt, y Smith 1989; Pfeffermann 1993) incorporó los pesos muestrales en la función de log-verosimilitud para producir estimadores consistentes bajo el diseño de muestreo, extendiendo así los modelos lineales generalizados de Nelder y Wedderburn (1972) al contexto de encuestas complejas.

Desde una perspectiva socioeconómica, una referencia fundamental para el análisis de encuestas de hogares es el trabajo de Deaton (1997), que establece un puente entre la teoría estadística, el análisis económico y las necesidades de información para el diseño y la evaluación de políticas públicas. En su libro, el autor articula los principios del muestreo con las herramientas de la microeconometría aplicada, explica cómo deben incorporarse las características del diseño muestral en la estimación y la inferencia, y desarrolla métodos para estudiar, a partir de microdatos, el bienestar, la pobreza, la desigualdad, los patrones de consumo, la demanda de los hogares y las decisiones de ahorro.

Heeringa, West, y Berglund (2017) y Valliant, Dever, y Kreuter (2018) han sintetizado estas contribuciones en un marco analítico unificado que integra de manera coherente el diseño muestral, la ponderación y la modelización, constituyéndose en referencias estándar para la práctica contemporánea del análisis de encuestas complejas.

Este documento está dirigido a lectores con formación en estadística o ciencia de datos, o con experiencia profesional equivalente. Está especialmente pensado para quienes cuentan con conocimientos de los fundamentos de la teoría de muestreo, de modelos de regresión lineal y logística, y con un manejo introductorio del lenguaje de programación R. En este contexto, el documento está orientado a profundizar y consolidar estos conocimientos, facilitando su aplicación rigurosa en contextos complejos y de alta exigencia técnica. Su propósito es acompañar al investigador en la transición desde los fundamentos teóricos hacia su implementación adecuada en la producción de estadísticas oficiales.

1.1. Justificación

A pesar de la riqueza metodológica disponible en la literatura especializada, persiste una brecha significativa entre el estado del arte y la práctica cotidiana del análisis de encuestas en la región. Una de las principales manifestaciones de esta brecha se relaciona con el acceso y uso del software estadístico. Algunos textos de referencia ampliamente reconocidos en el análisis de encuestas, como Cochran (1977) y Särndal, Swensson, y Wretman (2003), constituyen fuentes fundamentales desde el punto de vista teórico; sin embargo, no presentan ejemplos de aplicación real con software estadístico. Otros documentos están orientados principalmente hacia software con licencia, como `SAS`, `Stata` o `SPSS`, o bien no logran integrarse plenamente con un ecosistema computacional de software libre como `R`. En contraste, muchos tutoriales disponibles en línea sobre análisis de encuestas suelen privilegiar los aspectos operativos del software, pero rara vez alcanzan el nivel de profundidad conceptual necesario para un uso adecuado y responsable en la producción de estadísticas oficiales.

Adicionalmente, existe una fragmentación metodológica importante entre los distintos componentes que intervienen en el análisis de encuestas complejas. Con frecuencia, los fundamentos del diseño muestral, la gestión computacional de los datos, la estimación de parámetros descriptivos e inferenciales con corrección por diseño y el tratamiento de la no respuesta son abordados de manera separada, dificultando la construcción de flujos de trabajo integrales y reproducibles. En este contexto, el presente texto propone una integración coherente de estos elementos dentro de un único marco analítico implementado completamente en `R`.

Otro aspecto relevante corresponde a la contextualización regional de los ejemplos y aplicaciones. Una proporción considerable de los textos de referencia utiliza bases de datos foraneas, lo que limita parcialmente su aplicabilidad inmediata en América Latina. Con el propósito de acercar los contenidos a las realidades estadísticas de la región, este documento emplea una base de datos nativa de `R`, diseñada para replicar la estructura tradicional de las encuestas de hogares latinoamericanas y, por tanto, proporcionar ejemplos más cercanos a los problemas enfrentados por las oficinas nacionales de estadística y los equipos técnicos de la región.

Finalmente, este documento constituye en sí mismo un ejercicio de investigación reproducible. Todo su contenido, incluyendo tablas, figuras, ecuaciones y resultados numéricos, es generado íntegramente mediante código en `R` incorporado directamente en los capítulos. De esta manera, el lector no solo puede estudiar los métodos presentados, sino también replicarlos, modificarlos y adaptarlos a sus propios contextos y bases de datos, fortaleciendo así las capacidades analíticas y la transparencia en la producción estadística.

1.2. Uso del software R para el análisis de encuestas

La elección de `R` como entorno computacional para este documento no es arbitraria. `R` reúne al menos tres características que lo convierten en una opción idónea para el análisis de encuestas de hogares en el contexto latinoamericano.

En primer lugar, su condición de acceso libre. A diferencia de otros programas con costos de licencia que pueden resultar prohibitivos para muchas oficinas nacionales de estadística en países en desarrollo, `R` se distribuye bajo licencia GNU, lo que garantiza que los métodos presentados en este documento sean plenamente accesibles y replicables, independientemente de las restricciones presupuestarias que algunas veces aquejan a las instituciones públicas.

En segundo lugar, destaca su ecosistema especializado. El paquete `survey` (Lumley et al. 2026), en constante evolución, implementa una amplia gama de métodos de estimación para diseños complejos: estimadores de Horvitz–Thompson, técnicas de calibración, linealización de Taylor, métodos de replicación, modelos de regresión ponderados y pruebas de hipótesis ajustadas al diseño, entre otros. A su vez, el paquete `srvyr` (Freedman Ellis y Schneider 2026) extiende estas capacidades al incorporar la sintaxis del ecosistema `tidyverse`, facilitando una transición fluida entre el análisis exploratorio y el inferencial. Paquetes complementarios como `TeachingSampling` (Gutiérrez 2020) y `convey` (Jacob, Pessoa, y Damico 2024) consolidan un entorno analítico difícil de igualar en otros lenguajes.

Finalmente, `R` ofrece una sólida integración con herramientas de reproducibilidad científica. En particular, su combinación con `Quarto` y con el ecosistema de `R Markdown`, que incluye herramientas como `bookdown`, permite generar documentos reproducibles en los que los resultados, las tablas y las figuras se generan directamente a partir del código fuente. Esta capacidad resulta especialmente relevante en el ámbito de las estadísticas oficiales, donde la trazabilidad y la transparencia metodológica son requisitos fundamentales.

1.3. Estructura del documento

Este documento está organizado siguiendo una progresión lógica que va desde los fundamentos conceptuales del muestreo hasta aplicaciones más avanzadas del modelamiento estadístico. Esta secuencia no es arbitraria: cada nivel de conocimiento constituye un requisito necesario para comprender el siguiente, de modo que el lector pueda transitar de los principios básicos hacia herramientas analíticas más sofisticadas de manera coherente.

En primer lugar, se abordan los fundamentos del diseño muestral. El análisis adecuado de una encuesta comienza necesariamente con la comprensión de su diseño probabilístico. Antes de estimar cualquier parámetro, el analista debe familiarizarse con tres elementos constitutivos: la población objetivo, entendida como el conjunto de unidades sobre las cuales se desean hacer inferencias; el marco de muestreo, que corresponde al dispositivo que permite ubicar e identificar dichas unidades; y las unidades de muestreo en cada etapa del diseño. Estos componentes determinan tanto la validez como el alcance de las inferencias posteriores. A continuación, se introducen los elementos computacionales que permiten materializar estos conceptos en un entorno aplicado. El documento introduce el manejo y transformación de variables mediante un ambiente de programación adecuado para la incorporación del diseño de muestreo, que constituye el eje central del flujo de análisis y garantiza que todas las operaciones estadísticas respeten dicha estructura.

Con el diseño de muestreo correctamente especificado, el documento avanza hacia la estimación de parámetros descriptivos. Esta sección cubre los estimadores más utilizados en el análisis de encuestas para variables continuas. Asimismo, se presentan sus correspondientes medidas de precisión (varianzas, errores estándar e intervalos de confianza) ajustadas a la complejidad del diseño muestral.

Posteriormente, se desarrolla el análisis de relaciones entre variables a través del modelamiento estadístico, yendo más allá de la estadística descriptiva hacia un enfoque analítico e inferencial. En particular, se abordan los modelos de regresión lineal bajo diseños complejos, donde la incorporación explícita de los pesos muestrales resulta fundamental para obtener estimadores insesgados respecto al diseño. Este contexto pone de manifiesto la discusión entre el enfoque basado en el diseño y el enfoque basado en los modelos. Adicionalmente, se introducen los modelos lineales generalizados y los modelos multinivel, que amplían el alcance del análisis a variables de respuesta no normales, como variables binarias y conteos.

La organización de los capítulos sigue la arquitectura conceptual descrita anteriormente, avanzando desde los fundamentos hacia las aplicaciones más complejas. De esta forma, el capítulo 2 presenta los conceptos básicos para el análisis de encuestas de hogares, se explican los elementos centrales de un diseño de muestreo (estratificación, conglomeración y pesos de muestreo desiguales) y se introduce los principales estimadores de muestreo para totales, desarrollando los fundamentos para estimar su varianza y obtener intervalos de confianza.

En el capítulo 3 se aborda la estimación rigurosa de parámetros descriptivos para variables numéricas bajo diseños de muestreo complejos y se desarrolla la teoría para estimar totales, medias y razones. El capítulo también cubre la estimación de percentiles, y medidas de desigualdad mediante indicadores como el índice de Gini y la curva de Lorenz. Se presentan los elemento principales para realizar pruebas de hipótesis y contrastes, así como la correcta visualización de variables continuas.

En el capítulo 4 se aborda el análisis de variables categóricas bajo diseños complejos. Se desarrollan técnicas para estimar tamaños poblacionales y proporciones. Asimismo, se presentan tablas de contingencia ponderadas con sus respectivas frecuencias y proporciones, así como pruebas de independencia que ajustan la estadística clásica para reflejar el efecto del diseño. El capítulo también cubre la estimación e interpretación de medidas de asociación como la razón de Odds, el análisis de diferencias de proporciones entre subpoblaciones mediante contrastes, y diversas estrategias de visualización, incluyendo gráficos de barras con intervalos de confianza y mapas.

El capítulo 5 se enfoca en analizar relaciones entre variables y construir modelos estadísticos válidos bajo diseños complejos, cuando los supuestos tradicionales de independencia e idéntica distribución no se cumplen. Se presenta la evolución del problema desde los aportes iniciales hasta los desarrollos más recientes, y discute las implicaciones de trabajar con datos de encuestas en el contexto de los modelos de regresión. El capítulo también aborda la realización de pruebas de hipótesis, el diagnóstico de residuales y la identificación de observaciones influyentes.

En el capítulo 6 se extiende la modelación estadística a variables que no siguen una distribución normal (como variables binarias, multinomiales y estrictamente positivas). Para ello, se presenta un marco unificado basado en los modelos lineales generalizados, y se contrasta el enfoque clásico de máxima verosimilitud con el de máxima pseudo-verosimilitud, basado en ecuaciones de estimación ponderadas que incorporan los pesos muestrales.

En el capítulo 7 se revisan los modelos multinivel (también llamados jerárquicos), para analizar datos con estructura anidada, incorporando simultáneamente variables a nivel individual y contextual. A

través de ejemplos de regresión (con interceptos y pendientes que varían por estrato), se muestra que ignorar la estructura jerárquica puede llevar a inferencias incorrectas, mientras que modelarla explícitamente permite capturar la heterogeneidad entre grupos.

Adicionalmente, el primer apéndice introduce el manejo de bases de datos en `R` mediante el entorno `tidyverse`, con énfasis en el paquete `dplyr`. A partir de la base de datos `BigCity`, se presenta un flujo básico de trabajo que incluye la carga de librerías, la inspección inicial de la base, el uso del operador de encadenamiento `%>%` (*pipeline*) y la aplicación de los principales verbos de `dplyr`: `filter()`, `select()`, `arrange()`, `mutate()`, `summarise()` y `group_by()`. El capítulo muestra cómo seleccionar registros y variables, ordenar datos, crear nuevas variables y producir resúmenes descriptivos, explicando tanto la lógica de los códigos como la interpretación de sus salidas.

Finalmente, el segundo apéndice aborda la inferencia en poblaciones finitas y la diferencia entre la inferencia basada en modelos y la inferencia basada en el diseño muestral. A partir de ejemplos de simulación, muestra que en las encuestas complejas la aleatoriedad proviene principalmente de la selección de la muestra y que, por tanto, las estimaciones deben incorporar las probabilidades de inclusión y los pesos muestrales. Se presenta el método de máxima pseudo-verosimilitud como una extensión adecuada para datos de encuestas, donde cada contribución individual se pondera según sus probabilidades de inclusión.

El lector que recorra en orden los capítulos estará en la capacidad de analizar una encuesta de hogares, al entender su diseño, procesar sus datos, estimar los indicadores de interés con la precisión estadística correcta y modelar las relaciones entre variables de manera válida bajo el diseño complejo. Esa capacidad es, en esencia, lo que la producción rigurosa de estadísticas sociales exige.

2. Conceptos básicos en encuestas de hogares

Las encuestas de hogares constituyen una de las principales herramientas para comprender la realidad social y económica de una población. A partir de ellas se obtiene información esencial sobre las condiciones de vida, el empleo, la educación, la salud y otros aspectos que orientan la formulación de políticas públicas y la toma de decisiones basadas en evidencia. El análisis riguroso de estas encuestas es fundamental para estudiar la realidad social, económica y demográfica de una población, así como para monitorear el progreso de los países en el cumplimiento de la Agenda 2030 y los Objetivos de Desarrollo Sostenible (Naciones Unidas 2021). Sin embargo, la validez de los análisis no depende únicamente del tamaño de la muestra o de la calidad de la recolección de datos, sino también de la manera en que la encuesta fue diseñada y de cómo dicho diseño se incorpora en el análisis estadístico. Ignorar el diseño de muestreo y aplicar métodos de análisis tradicionales (que asumen una muestra aleatoria simple) puede conducir a estimaciones sesgadas, errores en la inferencia y por ende conclusiones equivocadas.

En la práctica, las encuestas de hogares suelen utilizar diseños de muestreo complejos que combinan la estratificación, junto con la selección de informantes en varias etapas y ajustes a los factores de expansión que generan ponderadores desiguales. Este tipo de diseños busca optimizar recursos y mejorar la precisión de las estimaciones, e ignorar estas características invalida la representatividad de la información producida. Por ende, para obtener conclusiones válidas sobre la población, es necesario reconocer que la muestra no proviene de una selección aleatoria cualquiera, sino de un plan probabilístico cuidadosamente definido.

2.1. Población objetivo y unidades de muestreo

Gambino y Nascimento Silva (2009) menciona que, desde la popularización de las encuestas de hogares en la década de 1940, se ha evidenciado diversas tendencias asociadas a los avances tecnológicos tanto en las oficinas e institutos nacionales de estadística como en la sociedad en general, las cuales se aceleraron con la introducción de las herramientas computacionales. En este contexto, el muestreo surge como una respuesta a la necesidad de obtener información estadística precisa sobre una población objetivo sin recurrir a la realización de un censo completo. Como señala Gutiérrez (2016), una encuesta por muestreo estudia una parte de la población con el propósito de realizar inferencias sobre su conjunto.

En las últimas décadas, la realización de encuestas se ha consolidado en distintos campos, especialmente en el sector gubernamental, mediante la producción de estadísticas oficiales que permiten el seguimiento de las políticas públicas. Asimismo, el uso del muestreo se ha extendido al ámbito académico, al sector privado y a los medios de comunicación, donde constituye una herramienta fundamental para la generación y el análisis de información (Groves et al. 2009).

Toda encuesta se encuentra asociada a una población finita compuesta por individuos o elementos sobre los cuales se desea obtener información. El conjunto de unidades sobre las que se producirán

estimaciones y resultados recibe el nombre de población objetivo. Asimismo, las encuestas definen unidades de análisis, que corresponden a los distintos niveles de desagregación para los cuales se presentan estadísticas de interés, como personas, hogares o viviendas.

2.2. Diseños de muestreo

Las encuestas de hogares suelen basarse en diseños de muestreo complejos, caracterizados por tres elementos principales: factores de expansión desiguales, selección de unidades mediante conglomerados y múltiples etapas, y estratificación de la población. Estas características determinan la manera en que cada observación contribuye a las estimaciones y afectan tanto la estimación puntual como su precisión. Por esta razón, deben incorporarse explícitamente en el análisis estadístico de las encuestas.

El principio de representatividad (Gutiérrez 2016) indica que cada unidad de la población posee una probabilidad conocida y distinta de cero de ser incluida en la muestra. Este paradigma constituye la principal garantía de que los resultados puedan generalizarse a la población de referencia. Bajo este enfoque, los estimadores de muestreo son insesgados (o aproximadamente insesgados) con respecto al diseño de muestreo, y no es necesario asumir ningún supuesto sobre distribuciones específicas para la variable de interés. Esta característica confiere a la inferencia basada en el diseño un carácter robusto y ampliamente aceptado en el análisis de encuestas de hogares.

2.2.1. Factores de expansión

Los factores de expansión (ponderadores, o pesos de muestreo) indican cuántas unidades de la población representa cada unidad seleccionada. En su forma más básica, se obtienen como el inverso multiplicativo de la probabilidad de inclusión de un elemento. Por consiguiente, cuando las unidades tienen probabilidades de selección diferentes, también reciben factores de expansión desiguales. Al especificar correctamente estos factores en el análisis, las estimaciones tienen en cuenta las particularidades del diseño y reflejan adecuadamente la estructura de la población.

No obstante, los pesos básicos de muestreo suelen requerir ajustes adicionales con el fin de mejorar la precisión y la robustez de las estimaciones. Uno de los ajustes más comunes corresponde al tratamiento de la no respuesta, mediante el cual los pesos de las unidades que sí respondieron se incrementan para representar también a aquellas unidades seleccionadas que no participaron en la encuesta, contribuyendo así a reducir posibles sesgos. Otro ajuste ampliamente utilizado es la calibración (Deville y Särndal 1992), que consiste en modificar los pesos para garantizar que las sumas ponderadas de ciertas variables auxiliares, como la edad o el sexo, coincidan con totales poblacionales conocidos provenientes de censos o proyecciones demográficas.

Los ajustes de los pesos no solo corrigen las imperfecciones derivadas de la no respuesta o la cobertura incompleta, sino que también fortalecen la coherencia entre la muestra y la población objetivo, garantizando que las inferencias sean estadísticamente válidas y comparables con otras fuentes oficiales (Kalton y Flores-Cervantes 2003). Los pesos calibrados tienden a reducir la varianza de las estimaciones, mejorando su precisión sin alterar los totales poblacionales conocidos (Särndal y Lundström 2005). En consecuencia, la calibración no solo corrige, sino que también optimiza el uso de la información disponible, permitiendo obtener resultados más precisos.

Valliant y Dever (2017) presentan un marco sistemático para construir, ajustar y evaluar los pesos de una encuesta compleja, que abarca desde el cálculo de los pesos iniciales hasta los ajustes por falta de respuesta, la calibración con información auxiliar y su consideración en la estimación de varianzas. En consonancia con este enfoque, el paquete `weightflow` permite implementar estas etapas mediante un flujo de trabajo reproducible (Ferreira 2026).

Como muestran Edward L. Korn y Graubard (1995), las estimaciones ponderadas y no ponderadas pueden diferir sustancialmente, lo que evidencia la importancia de utilizar métodos de análisis coherentes con el diseño de la encuesta. Naciones Unidas (2026, cap. 9) presenta el siguiente ejemplo. Supóngase un país conformado por dos regiones: la Región A, con 100 habitantes y un ingreso promedio de \$10.000, y la Región B, con 900 habitantes y un ingreso promedio de \$2.000. El ingreso promedio verdadero de la población es:

$$\theta = \frac{(100 \times 10.000) + (900 \times 2.000)}{100 + 900} = 2.800$$

Si se seleccionan 50 personas en cada región y se ignora el diseño de muestreo, asignando el mismo peso a todas las observaciones, el promedio estimado sería:

$$\hat{\theta} = \frac{(50 \times 10.000) + (50 \times 2.000)}{100} = 6.000$$

En este caso, la estimación sobrestima considerablemente el ingreso nacional, ya que la Región A, que representa únicamente el 10 % de la población, termina influyendo tanto como la Región B, que concentra el 90 %. En cambio, al aplicar pesos proporcionales a los tamaños poblacionales (1 para la Región A y 9 para la Región B), se obtiene la siguiente estimación:

$$\hat{\theta} = \frac{(1 \times 50 \times 10.000) + (9 \times 50 \times 2.000)}{(1 \times 50) + (9 \times 50)} = 2.800$$

Este resultado coincide con el valor verdadero de la población y muestra cómo el uso adecuado de los pesos corrige el sesgo introducido por un análisis que ignora el diseño de muestreo.

2.2.2. Conglomeración y selección en múltiples etapas

En las encuestas de hogares con diseños complejos, los marcos de muestreo suelen estar basados en áreas geográficas. Es decir, se construyen a partir de estructuras espaciales que vinculan a los hogares o personas con zonas delimitadas del territorio. Este tipo de marco de áreas facilita la organización territorial en unidades más manejables y permite implementar procesos de selección en varias etapas, manteniendo los principios del muestreo probabilístico. En estos diseños, las unidades primarias de muestreo (UPM) suelen corresponder a sectores censales, áreas de enumeración u otras divisiones territoriales, dentro de las cuales se construyen o actualizan listados de viviendas para las etapas posteriores de selección (Naciones Unidas 2005, cap. 4).

En la práctica, este tipo de diseños permite reducir los costos operativos y facilitar el trabajo de campo; sin embargo, también tiene implicaciones importantes sobre la precisión de las estimaciones. Cuando los hogares pertenecientes a un mismo conglomerado comparten características similares, la información adicional que aporta cada nueva observación dentro del conglomerado disminuye.

Por ejemplo, supóngase una encuesta que selecciona 100 conglomerados y, dentro de cada uno, 10 hogares, obteniendo así una muestra total de 1.000 hogares. Si los hogares de un mismo conglomerado presentan comportamientos muy parecidos (por ejemplo, niveles de ingreso cercanos, o todos cuentan con acceso a electricidad), la variabilidad efectiva de los datos en la muestra se reducirá considerablemente. En términos prácticos, la precisión obtenida podría ser equivalente a la de una muestra aleatoria simple de apenas 100 hogares (Kish y Frankel 1974). Analizar los 1.000 hogares como si fueran observaciones completamente independientes, ignorando la estructura de conglomeración, conduce a una subestimación de los errores estándar y, en consecuencia, a intervalos de confianza artificialmente estrechos y conclusiones potencialmente erróneas.

2.2.3. Estratificación

La estratificación consiste en dividir previamente la población en grupos mutuamente excluyentes, denominados estratos, y seleccionar una muestra independiente dentro de cada uno de ellos (Särndal, Swensson, y Wretman 2003). Los estratos pueden definirse a partir de criterios geográficos, administrativos, demográficos o socioeconómicos. En las encuestas de hogares, es frecuente utilizar combinaciones de región, área urbana o rural y características socioeconómicas para formar los estratos del diseño (CEPAL 2023).

Este procedimiento permite garantizar la presencia en la muestra de grupos relevantes que podrían quedar insuficientemente representados bajo una selección no estratificada. También puede mejorar la precisión de las estimaciones cuando las unidades pertenecientes a un mismo estrato son relativamente homogéneas y existen diferencias importantes entre los estratos. La asignación de la muestra entre ellos puede realizarse de manera proporcional a su tamaño o mediante asignaciones no proporcionales orientadas a satisfacer objetivos específicos de precisión (Cochran 1977; Gutiérrez 2016).

La estratificación también puede producir probabilidades de selección diferentes entre las unidades. Por ejemplo, una región pequeña puede recibir una fracción de muestreo mayor para asegurar un número suficiente de observaciones, mientras que una región más poblada puede recibir una fracción menor. Esta asignación desigual genera factores de expansión diferentes y refuerza la necesidad de considerar conjuntamente los estratos y los pesos de muestreo durante el análisis (Naciones Unidas 2005).

2.3. Estimadores de muestreo

Al analizar datos provenientes de encuestas complejas, es fundamental definir con precisión el parámetro de interés. Este corresponde a una cantidad fija, aunque generalmente desconocida, que se obtiene a partir de los valores de todas las unidades que conforman la población finita, denotada por U, y que resume alguna de sus características. Entre los parámetros más comunes se encuentran las proporciones, los totales, las medias, las razones, entre otros. Dado que, en la práctica, no es posible observar a toda la población, las encuestas por muestreo permiten inferir estos parámetros a partir de una muestra, denotada como s.

En el muestreo probabilístico, cada unidad de la población tiene una probabilidad de inclusión conocida y mayor que cero de ser seleccionada en la muestra. Estas probabilidades constituyen la base para el cálculo de los pesos básicos de muestreo, los cuales se utilizan posteriormente para estimar

los parámetros poblacionales mediante sumas ponderadas de los datos recolectados en la encuesta. Cuando el diseño y la selección se implementan correctamente, los estimadores resultantes son insesgados respecto del diseño; es decir, su valor esperado, calculado sobre todas las muestras posibles, coincide exactamente con el parámetro poblacional (Särndal, Swensson, y Wretman 2003).

La estimación de totales constituye un paso central en el análisis estadístico aplicado a poblaciones finitas. Gran parte de los indicadores de interés para la formulación de políticas públicas (como por ejemplo, el número de personas en situación de pobreza, el total de ocupados o el gasto agregado de los hogares) se derivan de un total poblacional. Por esta razón, comprender cómo se definen y estiman los totales resulta fundamental para garantizar la calidad y pertinencia de la información producida.

En términos formales, si y_k denota el valor de una variable de interés para la unidad $k \in U$, el total poblacional se define como

$$t_y = \sum_U y_k$$

Denotando a N como el tamano de la población, la media poblacional se define como $\bar{y} = \frac{t_y}{N}$. Dado que en la práctica solo se observa una muestra $s \subset U$, es necesario recurrir a estimadores que incorporen el diseño de muestreo. Por ejemplo, el estimador de Horvitz–Thompson (Horvitz y Thompson 1952) se expresa como

$$\hat{t}_y = \sum_{k \in s} d_k y_k \tag{2.1}$$

donde $d_k = 1/\pi_k$ corresponde al peso básico proveniente del diseño de muestreo y $\pi_k = \Pr(k \in s)$ es la probabilidad de inclusión de primer orden asociada al individuo k. En la práctica, los pesos básicos d_k suelen ajustarse para reflejar procesos adicionales como el ajuste por no respuesta o la calibración a totales poblacionales conocidos, obteniendo así nuevos pesos ajustados, denotados como w_k para todo $k \in s$. El uso de los pesos w_k permite mejorar la precisión y reducir sesgos en las estimaciones, especialmente cuando existen fuentes auxiliares de información confiables (Deville y Särndal 1992).

Toda estimación a partir de una muestra conlleva incertidumbre, puesto que las estimaciones varían de una muestra a otra debido a la propiedad de aleatoriedad del diseño. Esta incertidumbre se cuantifica mediante la varianza del estimador con la cual se puede calcular el error estándar (se) o el coeficiente de variación (cv). Estos indicadores son herramientas indispensables para evaluar la precisión de los totales estimados y, por tanto, para interpretar de manera adecuada la información estadística. Por ejemplo, un estimador insesgado para la varianza del estimador de Horvitz–Thompson (ecuación 2.1) puede expresarse como:

$$\widehat{Var}_p(\hat{t}_y) = \sum_{k \in s} \sum_{l \in s} (d_k d_l - d_{kl}) y_k y_l \tag{2.2}$$

En donde $d_{kl} = 1/\pi_{kl}$ y $\pi_{kl} = \Pr(k, l \in s)$ representan probabilidades de inclusión conjuntas. Para garantizar el insesgamiento en ecuación 2.2, el diseño muestral debe satisfacer $\pi_{kl} > 0$ para todo par de unidades $k, l \in U$.

Para comprender de manera más concreta la relevancia de considerar el diseño muestral en la estimación de totales y de sus varianzas, analicemos un ejemplo de Naciones Unidas (2026) que permite visualizar cómo el método de estimación se ajusta al esquema de selección adoptado. Suponga una población finita de tamaño $N = 6$, de la cual se selecciona una muestra aleatoria simple (MAS) sin reemplazo de tamaño $n = 3$. En dicha muestra se observan los valores $(y_1 = 10, y_2 = 14, y_3 = 18)$. Bajo este diseño, el estimador de Horvitz–Thompson del total poblacional se define como la suma de los valores observados divididos por sus probabilidades de inclusión, y su varianza estimada se obtiene a partir de las covarianzas inducidas por el proceso de selección. De esta forma, la estimación del total poblacional es $\hat{t}_y = 84$.

Como señala Gutiérrez (2016), para este diseño de muestreo particular, la expresión general de la varianza presentada en ecuación 2.2 se reduce a:

$$\widehat{Var}_{MAS}(\hat{t}_y) = \frac{N^2}{n}\left(1 - \frac{n}{N}\right) S^2_{y_s}$$

donde $S^2_{y_s}$ corresponde a la varianza muestral de los valores observados. Sustituyendo‘los valores del ejemplo en la expresión anterior, se obtiene $\widehat{Var}_{MAS}(\hat{t}_y) = 96$. En contraste, si se ignoraran las características del diseño muestral y los datos se analizaran como si provinieran de una muestra aleatoria simple, la varianza se estimaría incorrectamente mediante la expresión $\frac{N^2}{n} S^2_{y_s} = 192$. Este valor duplica la varianza estimada bajo el diseño de muestreo, y conduce, por tanto, a una sobreestimación de la incertidumbre asociada al estimador. En consecuencia, el intervalo de confianza resultante sería innecesariamente más amplio, lo que reduciría la precisión aparente de la estimación y podría conducir a conclusiones equivocadas sobre la incertidumbre de los resultados.

2.4. El efecto de diseño (DEFF)

De acuerdo con Kish (1965, 258), el efecto del diseño (*design effect*, DEFF) se define como el cociente entre la varianza de un estimador $\hat{\theta}$ bajo un diseño de muestreo complejo y la varianza que tendría ese mismo estimador bajo un muestreo aleatorio simple (MAS) con igual tamaño de muestra. Su estimación se expresa como:

$$\widehat{DEFF} = \frac{\widehat{Var}_p(\hat{\theta})}{\widehat{Var}_{MAS}(\hat{\theta})}$$

En donde, $\widehat{Var}_p(\hat{\theta})$ corresponde a la varianza estimada de $\hat{\theta}$ bajo el diseño complejo $p(\cdot)$, mientras que $\widehat{Var}_{MAS}(\hat{\theta})$ representa la varianza estimada del mismo estimador bajo un diseño MAS con igual tamaño muestral.

Este indicador permite cuantificar cuánto aumenta la varianza debido a la conglomeración y otras características propias de los diseños complejos en comparación con un muestreo simple. Según Naciones Unidas (2009, 40), el DEFF puede entenderse de tres maneras: como el factor de incremento de la varianza frente al MAS, como una medida de la pérdida relativa de precisión o como una indicación del aumento en el tamaño de la muestra que sería necesario en un diseño complejo para alcanzar el mismo nivel de varianza que en un MAS. Según Park et al. (2003), el efecto del diseño de una encuesta está supeditado a los siguientes tres factores:

- Ponderadores desiguales: la presencia de pesos muestrales no uniformes suele incrementar ligeramente la varianza; por ello, el uso de pesos uniformes resulta ventajoso y explica por qué los diseños auto-ponderados son preferidos en las encuestas de hogares.
- Estratificación: cuando se aplica correctamente, puede disminuir la varianza, aunque en la práctica su efecto reductor suele ser moderado.
- Muestreo en varias etapas: generalmente incrementa la varianza, ya que las unidades dentro de un mismo conglomerado tienden a ser más homogéneas entre sí que en comparación con las de otros conglomerados.

En el análisis de encuestas, el DEFF constituye un indicador esencial para evaluar la precisión y eficiencia de las estimaciones, así como para orientar la planificación de futuros estudios. Un valor elevado evidencia que el diseño complejo introduce un incremento de la varianza, reduciendo la precisión de los resultados. Por el contrario, un valor cercano a uno indica que el diseño tiene un efecto mínimo sobre la varianza. Esta información permite a los investigadores identificar si es necesario ajustar la ponderación, optimizar la estratificación o modificar el tamaño de muestra para aumentar la eficiencia en levantamientos posteriores.

La interpretación de un DEFF elevado debe realizarse con cautela, pues no necesariamente indica que el diseño de muestreo sea inadecuado. Su magnitud debe evaluarse en función del contexto y de las condiciones operativas de la encuesta. Aunque Naciones Unidas (2005) afirma que un valor superior a 2.5 o 3 podría considerarse inicialmente preocupante, este puede responder a restricciones presupuestarias, dificultades logísticas o decisiones orientadas a facilitar la participación de los informantes.

Por ejemplo, en algunas encuestas de hogares puede ser necesario seleccionar únicamente a una fracción de las personas elegibles dentro de cada vivienda. Asimismo, los problemas de cobertura y las tasas elevadas de no respuesta pueden incrementar la variabilidad de los pesos muestrales y, por esta vía, aumentar el DEFF. Gutiérrez y Babativa-Márquez (2023) presentan un análisis detallado de los efectos del diseño en las encuestas de hogares de América Latina a partir de la información disponible en BADEHOG.

2.5. Inferencia en encuestas complejas

Tal como destacan Heeringa, West, y Berglund (2017), el cálculo de totales y medias poblacionales, junto con sus varianzas, ha sido esencial para el desarrollo de la teoría del muestreo probabilístico y la interpretación adecuada de los resultados de encuestas de hogares. La determinación de los totales poblacionales constituye uno de los pilares fundamentales del análisis de encuestas. Tanto las medias como las proporciones y las razones se derivan de estos totales. Un total se define como la suma de una variable específica (por ejemplo, ingreso o gasto) a nivel de toda la población.

En el análisis de encuestas de hogares, es habitual estimar estadísticas descriptivas como totales, medias y razones, las cuales permiten sintetizar las principales características de la población y proporcionan información relevante para la toma de decisiones y monitoreo de indicadores. En este apartado se presentan los fundamentos teóricos de la inferencia basada en el diseño de muestreo, necesarios para estimar adecuadamente los distintos parámetros de interés y evaluar la incertidumbre asociada a dichas estimaciones.

2.5.1. Estimación puntual

Una vez definidos la población objetivo, las variables de interés y el diseño muestral, se procede a estimar los parámetros poblacionales correspondientes. Para establecer la notación de libro, se utilizará la letra h para identificar los estratos; la letra i para representar las unidades primarias de muestreo (UPM) contenidas en cada estrato; y la letra k para denotar las unidades finales observadas, como hogares o personas. Además, con el propósito de simplificar la escritura a lo largo de este documento, a menos que se indique lo contrario, $\sum_h$ denotará la suma sobre todos los estratos de la población; $\sum_i$ representará la suma sobre todas las UPM seleccionadas en la muestra dentro del estrato h; y $\sum_k$ corresponderá a la suma sobre todos los elementos observados en la muestra de la UPM i, perteneciente al estrato h de la población de interés. Con esta notación, el estimador de un total poblacional puede expresarse como:

$$\hat{t}_y = \sum_h \sum_i \sum_k w_{hik}\, y_{hik}$$

En esta expresión, w_{hik} representa el factor de expansión de la unidad k, perteneciente a la UPM i del estrato h, mientras que y_{hik} corresponde al valor observado de la variable de interés para esa misma unidad.

2.5.2. Estimación de la varianza

Algunos parámetros de interés pueden expresarse como combinaciones lineales de varios parámetros poblacionales. Considérese, en particular, la función:

$$f(\theta_1, \dots, \theta_J) = \sum_{j=1}^{J} a_j \theta_j,$$

donde a_j son constantes conocidas y θ_j representa el j-ésimo parámetro poblacional. Si $\hat{\theta}_j$ es un estimador de θ_j, entonces el estimador muestral de $f(\theta_1, \dots, \theta_J)$ se define como:

$$\hat{f} = \sum_{j=1}^{J} a_j \hat{\theta}_j$$

De esta manera, su varianza puede expresarse como:

$$Var(\hat{f}) = \sum_{j=1}^{J} a_j^2 Var(\hat{\theta}_j) \; + \; 2 \sum_{j=1}^{J-1} \sum_{k>j}^{J} a_j a_k \, Cov(\hat{\theta}_j, \hat{\theta}_k).$$

Este resultado es especialmente relevante, pues abarca numerosos casos de interés práctico que se desarrollarán a lo largo de este documento. No obstante, la estimación de la varianza bajo diseños de muestreo complejos puede resultar difícil, ya que depende de las características particulares de cada diseño. Para abordar esta dificultad, se han desarrollado diversos enfoques metodológicos que permiten cuantificar adecuadamente la incertidumbre y evaluar la precisión de las estimaciones. Con el apoyo de softwares estadísticos especializado, estos métodos pueden implementarse de manera eficiente y contribuir a la obtención de resultados rigurosos y confiables.

2.5.2.1. Último conglomerado

El método del último conglomerado es un enfoque ampliamente utilizado para estimar la varianza de totales en encuestas con diseños de muestreo estratificados y multietápicos. Desarrollado por Hansen et al. (1953), este método simplifica el tratamiento de las distintas etapas de selección al concentrar la estimación de la varianza en las unidades primarias de muestreo (UPM). Para su aplicación, se supone que, dentro de cada estrato, las UPM se seleccionan de manera independiente y con reemplazo, posiblemente con probabilidades desiguales. La aplicación de este método requiere disponer de estimadores insesgados del total de la variable de interés para cada UPM seleccionada. Asimismo, cuando el diseño incorpora estratificación en la primera etapa, es necesario contar con al menos dos UPM seleccionadas en cada estrato, para que sea posible estimar la variabilidad entre conglomerados dentro de este.

Considere un diseño de muestreo en múltiples etapas donde se seleccionan n_h UPM en el estrato h, $(h = 1, \dots, H)$. Una estimación del total de la variable de interés en la UPM i del estrato h es:

$$\hat{t}_{y_i} = \sum_{k \in s_{hi}} w_{hik}\, y_{hik}$$

En donde s_{hi} corresponde a la muestra de unidades en la UPM i del estrato h; por tanto, un estimador insesgado del total poblacional se expresaría como

$$\hat{t}_y = \sum_h \sum_i \hat{t}_{y_i}$$

Según el método del último conglomerado, suponiendo que $\hat{\bar{t}}_{y_h} = (1/n_h) \sum_i \hat{t}_{y_i}$ es la media muestral de los totales estimados de las UPM en el estrato h, un estimador aproximado de la varianza de $\hat{t}_y$ se obtiene mediante al siguiente expresión:

$$\widehat{Var}(\hat{t}_y) = \sum_h \frac{n_h}{n_h - 1} \sum_i (\hat{t}_{y_i} - \hat{\bar{t}}_{y_h})^2$$

Para más detalles, véase Hansen et al. (1953, 258) o Wolter (2007). Aunque este método se propuso originalmente para calcular varianzas de estimadores de totales, el método puede combinarse con otras técnicas para derivar varianzas de otros parámetros poblacionales más complejos. Esta flexibilidad hace que el método sea aplicable a diversos contextos de análisis de encuestas de hogares.

Aunque los métodos más sofisticados que consideran todas las etapas del diseño pueden ofrecer estimaciones de varianza ligeramente más precisas, su aplicación requiere información más detallada y mayor complejidad computacional. Por el contrario, el método del último conglomerado proporciona una aproximación confiable y eficiente, especialmente útil al estimar totales o medias en encuestas de hogares.

Un supuesto fundamental de esta técnica es que, dentro de cada estrato, las UPM se eligen de forma independiente y con reemplazo. En la práctica, la mayoría de las encuestas selecciona sus UPM sin reemplazo, generando diseños más eficientes. En consecuencia, las varianzas calculadas bajo la hipótesis de independencia constituyen aproximaciones a las verdaderas varianzas de muestreo. Cuando la fracción muestral es pequeña, dichas aproximaciones suelen ser suficientemente precisas para su utilización por parte de las oficinas nacionales de estadística o de otros analistas.

2.5.2.2. Ecuaciones de estimación y linealización de Taylor

Muchos parámetros poblacionales pueden expresarse como soluciones de ecuaciones de estimación que involucran totales poblacionales. Aunque los detalles técnicos pueden ser complejos, la idea fundamental es que los mismos principios utilizados para estimar totales pueden aplicarse también para la estimación de varianzas. Este marco general hace que el método sea sencillo y flexible, facilitando su implementación en softwares estadísticos como `R`. Una ecuación de estimación poblacional genérica puede expresarse de la siguiente manera:

$$\sum_{k \in U} z_k(\theta) = 0,$$

donde $z_k(\cdot)$ es una función de estimación evaluada para la unidad k y θ representa el parámetro poblacional de interés. Estas ecuaciones proporcionan un marco general para definir y calcular diversos parámetros de la población, como totales, medias y razones. Las siguientes formulaciones muestran cómo distintos parámetros de interés pueden expresarse mediante ecuaciones de estimación que comparten una estructura común.

- Para el total poblacional, se define $z_k(\theta) = y_k - \theta/N$, de manera que la ecuación de estimación es $\sum_{k \in U}(y_k - \theta/N) = 0$. La solución de esta ecuación conduce a $\theta = \sum_{k \in U} y_k = t_y$, es decir, al total poblacional.
- De forma similar, para la media poblacional se utiliza $z_k(\theta) = y_k - \theta$, y la ecuación $\sum_{k \in U}(y_k - \theta) = 0$ tiene como solución $\theta = \left(\sum_{k \in U} y_k\right)/N = \bar{Y}$, correspondiente a la media de la población.

La idea de definir parámetros poblacionales como soluciones de ecuaciones de estimación conduce naturalmente a un método general para obtener los estimadores de muestreo. Bajo un diseño de muestreo probabilístico, la suma poblacional $\sum_{k \in U} z_k(\theta)$ puede estimarse mediante su versión ponderada en la muestra. En particular, si w_k corresponde al factor de expansión de la unidad k, se tiene que $\sum_{k \in s} w_k z_k(\theta)$ es un estimador insesgado, bajo el diseño, de la ecuación poblacional. En consecuencia, el estimador de muestreo $\hat{\theta}$ se define como la solución de la ecuación muestral ponderada

$$\sum_{k \in s} w_k z_k(\hat{\theta}) = 0.$$

Bajo condiciones de regularidad, esta sustitución permite obtener estimadores consistentes de una amplia variedad de parámetros poblacionales (David A. Binder 1983).

Asimismo, mediante la linealización de Taylor es posible aproximar un estimador no lineal por una expresión lineal cuya varianza puede calcularse con los métodos habituales para estimadores de totales. Este enfoque, introducido tempranamente para el análisis de estadísticas complejas (Woodruff 1952), constituye la base de numerosos procedimientos utilizados actualmente para estimar las varianzas de medias, razones, proporciones y otros estimadores no lineales. El procedimiento consiste en aplicar una expansión de Taylor de primer orden a la ecuación de estimación, de modo que el estimador original se sustituya por una aproximación lineal. Para un parámetro definido como la solución de una ecuación de estimación muestral, un estimador consistente de su varianza puede expresarse como:

$$\widehat{Var}_p(\hat{\theta}) \approx [\hat{J}(\hat{\theta})]^{-1}\, \widehat{Var}_p\left[\sum_{k\in s} w_k\, z_k(\hat{\theta})\right] [\hat{J}(\hat{\theta})]^{-1} \tag{2.3}$$

donde

$$\hat{J}(\hat{\theta}) = \sum_{k\in s} w_k \left[\frac{\partial z_k(\theta)}{\partial \theta}\right]_{\theta=\hat{\theta}}$$

Este resultado muestra cómo la linealización de Taylor convierte la estimación de varianzas de parámetros complejos en un problema de estimación de totales. Por ejemplo, para el caso de la razón poblacional $\theta = \frac{t_y}{t_x}$ entre los totales t_y y t_x, se define $z_k(\theta) = y_k - \theta x_k$. Así, la ecuación de estimación muestral toma la siguiente forma

$$\sum_{k\in s} w_k z_k(\theta) = \sum_{k\in s} w_k\,(y_k - \theta x_k) = 0$$

La solución de esta ecuación corresponde al estimador de razón

$$\hat{\theta} = \frac{\sum_{k\in s} w_k y_k}{\sum_{k\in s} w_k x_k} = \frac{\hat{t}_y}{\hat{t}_x}$$

Dado que la derivada de la función de estimación con respecto a θ es $\frac{\partial z_k(\theta)}{\partial \theta} = -x_k$, entonces $\hat{J}(\hat{\theta}) = -\sum_{k\in s} w_k x_k = -\hat{t}_x$. Al sustituir este resultado en la expresión general (ecuación 2.3), se obtiene:

$$\widehat{Var}_p(\hat{\theta}) \approx \frac{1}{\hat{t}_x^2}\widehat{Var}_p\left[\sum_{k\in s} w_k\left(y_k - \hat{\theta} x_k\right)\right]$$

Esta expresión muestra que la estimación de la varianza de un estimador no lineal puede reducirse a la estimación de la varianza de un total ponderado de su variable linealizada. En consecuencia, pueden aplicarse los procedimientos habituales para estimar la varianza de totales, incorporando adecuadamente las características del diseño muestral.

2.5.2.3. Bootstrap

Como alternativa a los métodos analíticos, se desarrollaron los métodos de replicación, que generan múltiples conjuntos de pesos o réplicas de la muestra, calculan el estimador de interés en cada una de ellas y utilizan la variabilidad entre las estimaciones replicadas para aproximar la varianza del estimador original. Su principal ventaja es que pueden aplicarse de manera relativamente uniforme a una amplia variedad de parámetros, incluso cuando la derivación analítica de la varianza resulta difícil (Rust y Rao 1996; Wolter 2007).

Por ejemplo, el Bootstrap es una herramienta de replicación robusta y versátil. Originalmente introducido por Efron (1979) para datos que no provenían de encuestas, su adaptación más usada para encuestas de hogares es el Bootstrap de Reescalamiento propuesto por J. N. K. Rao, Wu, y

Yue (1992). Este método se ajusta de manera óptima a diseños de muestreo estratificados y multietápicos, y es ampliamente empleado para la estimación de varianzas en encuestas complejas. El procedimiento consiste en generar muchas réplicas de la muestra original, simulando extracciones repetidas de la población. Cada réplica se construye mediante la creación de columnas adicionales de pesos de replicación en la base de datos, siguiendo este proceso:

- Para cada estrato, se seleccionan aleatoriamente las UPM con reemplazo; algunas pueden repetirse y otras no aparecer. Cada UPM elegida se incorpora con todas sus observaciones. Si el tamaño de la muestra de primera etapa en el estrato h es mayor que dos ($n_h > 2$), el número de UPM seleccionadas por réplica es $n_h - 1$.
- Este proceso se repite muchas veces, habitualmente cientos, generando un gran número de réplicas. La cantidad de veces que una UPM i del estrato h aparece en la réplica r se denota $n_{hi}^{(r)}$, variando entre 0 y $n_h - 1$.
- A partir de cada réplica se calculan nuevos pesos bootstrap para todas las unidades, reflejando cuántas veces fue seleccionada su UPM. El peso de la unidad k en la réplica r se calcula como:

$$w_{hik}^{(r)} = w_{hik} \times \frac{n_h}{n_h - 1} \times n_{hi}^{(r)}$$

  Si los pesos originales incluyen ajustes por no respuesta o calibración, estos deben aplicarse también a cada conjunto de pesos bootstrap.
- Para cada réplica r, se calcula el parámetro de interés $\hat{\theta}^{(r)}$ usando los pesos bootstrap $w_{hik}^{(r)}$. Asumiendo que $\tilde{\theta} = \frac{1}{R}\sum_{r=1}^{R} \hat{\theta}^{(r)}$, entonces la varianza del estimador se aproxima mediante la variabilidad entre todas las réplicas:

$$\widehat{Var}_B(\hat{\theta}) = \frac{1}{R}\sum_{r=1}^{R}\left(\hat{\theta}^{(r)} - \tilde{\theta}\right)^2$$

Este enfoque asegura que la dispersión entre réplicas capture fielmente la incertidumbre del parámetro. El Bootstrap ofrece múltiples ventajas, a pesar de requerir un mayor procesamiento computacional, es eficaz para diseños de encuesta complejos y permite estimar parámetros difíciles de calcular con métodos tradicionales, como medianas u otras estadísticas no lineales. La simplicidad del método facilita su aplicación incluso sin software estadístico especializado. No obstante, su uso no es recomendable en encuestas repetidas con muestras superpuestas ni en situaciones con fracciones de muestreo grandes y tamaños de muestra pequeños (Bruch, Münnich, y Zins 2011).

2.5.3. Intervalos de confianza

Además de producir estimaciones puntuales, uno de los objetivos fundamentales de una encuesta es cuantificar la incertidumbre asociada a dichas estimaciones. En este contexto, los intervalos de confianza permiten construir un rango de valores plausibles para el parámetro poblacional de interés, incorporando explícitamente la variabilidad introducida por el diseño muestral. El intervalo de confianza de nivel $(1-\alpha)100\,\%$ para el total poblacional Y se calcula como:

$$\hat{\theta} \pm t_{1-\alpha/2,df} \times \sqrt{\widehat{Var}(\hat{\theta})}$$

donde $\hat{\theta}$ corresponde al estimador por muestreo del parámetro poblacional θ; $\widehat{Var}(\hat{\theta})$ representa el estimador de varianza bajo el diseño complejo de la encuesta; y $t_{1-\alpha/2,df}$ es el cuantil de la

distribución t de Student con df grados de libertad (Fuller 2009). En encuestas complejas, los grados de libertad suelen estar relacionados con la cantidad de unidades primarias de muestreo (UPM) y los estratos considerados en el diseño. Una aproximación común consiste en calcularlos como:

$$df = m - H$$

donde m representa el número total de UPM observadas en la muestra y H el número de estratos. Esta aproximación refleja la cantidad efectiva de información independiente disponible para estimar la variabilidad muestral. A medida que los grados de libertad aumentan, la distribución t de Student converge hacia la distribución normal estándar. Esta propiedad explica por qué muchas veces se utilizan aproximaciones normales para reportar intervalos de confianza, especialmente en encuestas de hogares de gran escala. Sin embargo, cuando el número de conglomerados es reducido o existen estratos con pocas UPM, la aproximación normal puede subestimar la incertidumbre y producir intervalos excesivamente estrechos.

3. Análisis de las variables continuas

El análisis estadístico de encuestas ha evolucionado de manera constante, impulsado por el desarrollo de nuevas metodologías y enfoques que buscan mejorar la precisión y la representatividad de los resultados (Lumley 2010; Heeringa, West, y Berglund 2017). En general, estos avances surgen en el ámbito académico y, con el tiempo, son adoptados por instituciones públicas, empresas privadas y organismos estadísticos, que demandan su incorporación en los principales programas de análisis de datos.

Este capítulo presenta los principales procedimientos para el análisis estadístico de variables continuas utilizando `R`, incorporando adecuadamente las características del diseño muestral, tales como los factores de expansión, así como la estratificación y la conglomeración. A lo largo de este capítulo se presentan métodos para estimar parámetros poblacionales, evaluar la precisión de las estimaciones y realizar inferencias válidas a partir de datos provenientes de encuestas complejas. Asimismo, tanto en este capítulo como en el resto del documento, se introducen de manera gradual las principales funciones y herramientas del paquete `survey`, una de las alternativas más completas y ampliamente utilizadas en `R` para el análisis adecuado de encuestas de hogares.

De acuerdo con R Core Team (2026b), las bases de datos utilizadas en el análisis de encuestas pueden encontrarse en una amplia variedad de formatos, como `xlsx`, `dat`, `csv`, `parquet`, `sas7bdat`, `sav` y `txt`, entre otros. Sin embargo, una buena práctica consiste en importar la información desde cualquiera de estos formatos y guardarla posteriormente en un archivo con extensión `.rds`, un formato nativo de `R` que permite almacenar distintos tipos de objetos, como bases de datos, vectores, matrices o listas. Otro formato nativo de `R` es `.RData`, que permite guardar varios objetos en un mismo archivo. En contraste, los archivos `.rds` almacenan un único objeto, lo que facilita un manejo más controlado, explícito y reproducible de la información. Por esta razón, se recomienda utilizar preferentemente el formato `.rds` al documentar proyectos y desarrollar análisis reproducibles.

Para ilustrar la sintaxis computacional que se utilizará a lo largo del capítulo, se empleará una base de datos proveniente de una encuesta con un diseño de muestreo complejo. A continuación, se muestra cómo cargar el archivo, almacenado en formato `.rds`:

```
survey_data <- readRDS("../Data/encuesta.rds")
```

El objeto `survey_data` carga la infomración desde el archivo `encuesta.rds`, el cual está organizado de manera que cada fila corresponde a una unidad observada, mientras que las columnas contienen tanto las variables de interés como la información necesaria para representar el diseño muestral. Para examinar rápidamente su estructura, puede utilizarse la función `glimpse()`, que presenta el número de observaciones y variables, sus nombres, sus tipos de datos y una muestra de los valores registrados.

```r
library(tidyverse)

dplyr::glimpse(survey_data)
```

```
Rows: 2,605
Columns: 20
$ HHID        <chr> "idHH00031", "idHH00031", "idHH00031", "idHH00031", "idHH0~
$ Stratum     <chr> "idStrt001", "idStrt001", "idStrt001", "idStrt001", "idStr~
$ NIh         <int> 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9~
$ nIh         <dbl> 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2~
$ dI          <dbl> 4.5, 4.5, 4.5, 4.5, 4.5, 4.5, 4.5, 4.5, 4.5, 4.5, 4.5, 4.5~
$ PersonID    <chr> "idPer01", "idPer02", "idPer03", "idPer04", "idPer05", "id~
$ PSU         <chr> "PSU0003", "PSU0003", "PSU0003", "PSU0003", "PSU0003", "PS~
$ Zone        <chr> "Rural", "Rural", "Rural", "Rural", "Rural", "Rural", "Rur~
$ Sex         <chr> "Male", "Female", "Female", "Male", "Female", "Female", "F~
$ Age         <int> 68, 56, 24, 26, 3, 61, 23, 5, 1, 59, 50, 33, 81, 76, 45, 4~
$ MaritalST   <fct> Married, Married, Married, Married, NA, Widowed, Separated~
$ Income      <dbl> 409.9, 409.9, 409.9, 409.9, 409.9, 823.8, 823.8, 823.8, 82~
$ Expenditure <dbl> 346.3, 346.3, 346.3, 346.3, 346.3, 392.2, 392.2, 392.2, 39~
$ Employment  <fct> Employed, Employed, Employed, Employed, NA, Employed, Inac~
$ Poverty     <fct> NotPoor, NotPoor, NotPoor, NotPoor, NotPoor, NotPoor, NotP~
$ dki         <dbl> 8.00, 8.00, 8.00, 8.00, 8.00, 8.00, 8.00, 8.00, 8.00, 8.00~
$ dk          <dbl> 36, 36, 36, 36, 36, 36, 36, 36, 36, 36, 36, 36, 42, 42, 42~
$ wk          <dbl> 34.5, 33.6, 33.6, 34.5, 33.6, 33.6, 33.6, 34.5, 33.6, 34.5~
$ Region      <fct> Norte, Norte, Norte, Norte, Norte, Norte, Norte, Norte, No~
$ CatAge      <ord> Más de 60, 46-60, 16-30, 16-30, 0-5, Más de 60, 16-30, 0-5~
```

La base contiene 2.605 registros individuales y 20 variables. Cada fila corresponde a una persona, identificada mediante `PersonID`, y se encuentra vinculada con su hogar a través de `HHID`. Además de variables sociodemográficas y económicas, como sexo, edad, estado civil, condición de empleo, ingreso, gasto y situación de pobreza, la base incluye información necesaria para representar el diseño de muestreo complejo, como el estrato (`Stratum`), la unidad primaria de muestreo (`PSU`) y los factores de expansión (`wk`). También incorpora variables de clasificación geográfica y demográfica, como zona, región y grupo de edad.

Según Naciones Unidas (2009, sec. 7.8), es fundamental que la estructura del diseño muestral sea tenida en cuenta en el proceso de inferencia al estimar estadísticas oficiales basadas en encuestas de hogares. Como se ha insistido en este documento, ignorar este aspecto puede generar estimaciones sesgadas y errores de muestreo subestimados. En este sentido los programas estadísticos incorporan funcionalidades específicas para manejar datos provenientes de encuestas con diseños complejos.

Una vez cargada la encuesta de hogares en `R`, el siguiente paso es definir el diseño muestral del cual proviene dicha muestra. Para ello se utilizará el paquete `srvyr`, que surge como un complemento del paquete `survey`. Estas librerías permiten definir objetos tipo `survey.design`, a los que se aplican las funciones de estimación y análisis de encuestas, y que pueden ser combinadas con la programación del paquete `tidyverse`. A continuación, se muestra un ejemplo en el que se define un objeto de tipo `survey_design` para la encuesta:

```r
options(survey.lonely.psu = "adjust")

library(survey)
library(srvyr)

survey_design <- survey_data %>%
  as_survey_design(
    strata = Stratum,
    ids = PSU,
    weights = wk,
    nest = TRUE
  )
```

Como explica Lumley et al. (2026), la opción `options(survey.lonely.psu = "adjust")` define cómo el paquete `survey` debe tratar los estratos que contienen una sola UPM en la muestra. En estos estratos no es posible calcular directamente la variabilidad entre UPM, porque no existe una segunda unidad con la cual realizar la comparación. Con el método `"adjust"`, la contribución de la UPM solitaria a la varianza se calcula centrando su estimación en la media general de las UPM de la muestra, en lugar de utilizar la media del propio estrato. Este ajuste evita que la contribución del estrato sea omitida o quede indefinida y suele producir una estimación conservadora de la varianza. La opción no modifica las estimaciones puntuales, sino únicamente sus varianzas, errores estándar e intervalos de confianza. Su aplicación debe distinguirse del caso en el que una UPM sea seleccionada con probabilidad de inclusión uno. En cualquier caso, la contribución a la varianza debe especificarse mediante la información apropiada del diseño.

En la implementación descrita por Freedman Ellis y Schneider (2026), `survey_data` corresponde a la base de datos que contiene la muestra seleccionada. A partir de estos datos, la función `as_survey_design()` convierte la información en un objeto de clase `tbl_svy`, el cual almacena todos los elementos necesarios para describir el diseño muestral de la encuesta. Entre ellos se incluyen la variable de estratificación (`strata`), que identifica los estratos definidos en el diseño, junto con la variable que representa las unidades primarias de muestreo (UPM) o conglomerados seleccionados en la primera etapa del muestreo (`ids`). Asimismo, el argumento `weights` define la variable que contiene los factores de expansión asociados a cada observación. Finalmente, la opción `nest = TRUE` vuelve únicos dentro de cada estrato los identificadores de conglomerado que pudieran repetirse entre estratos.

3.1. Estimación de parámetros descriptivos

En las encuestas de hogares es común estimar parámetros descriptivos asociados a variables numéricas, tales como totales, medias, razones y desviaciones estándar. Estas medidas permiten caracterizar distintos fenómenos sociales y económicos de la población. En esta sección se presenta la lógica detrás de la sintaxis utilizada para este tipo de estimación, junto con sus correspondientes medidas de precisión.

3.1.1. Totales

Como se ha señalado anteriormente, el total poblacional de una variable de interés puede estimarse de manera aproximadamente insesgada mediante el estimador de calibración (Deville y Särndal 1992), que pondera el valor observado en cada unidad por su correspondiente factor de expansión calibrado. En un diseño de muestreo con estratificación y selección en varias etapas, este estimador se expresa como:

$$\hat{t}_y = \sum_h \sum_i \sum_k w_{hik}\, y_{hik},$$

donde h identifica el estrato, i la unidad primaria de muestreo y k la unidad final observada. El término y_{hik} representa el valor de la variable de interés para la unidad k perteneciente a la unidad primaria i del estrato h, mientras que w_{hik} corresponde a su factor de expansión. De este modo, cada unidad de la muestra representa a un determinado número de unidades de la población, y la suma ponderada permite reconstruir el total poblacional.

En `R`, la función `survey_total()` se utiliza para estimar totales a partir del objeto de diseño. Esta función incorpora automáticamente los factores de expansión y la estructura del diseño, garantizando que las estimaciones reflejen correctamente el modelo de inferencia.

```
survey_total(
  x,
  na.rm = FALSE,
  vartype = c("se", "ci", "var", "cv"),
  level = 0.95,
  deff = FALSE
)
```

Los argumentos principales de la función `survey_total()` son `x`, `na.rm`, `vartype`, `level` y `deff`. El argumento `x` corresponde a la variable sobre la cual se desea calcular el total; también puede dejarse vacío cuando se busca obtener únicamente el total ponderado de casos de la muestra. Por su parte, `na.rm` indica si los valores faltantes (`NA`) deben eliminarse antes de realizar el cálculo, siendo `FALSE` el valor predeterminado.

El argumento `vartype` permite especificar el tipo de medida de incertidumbre que se desea estimar, pudiendo solicitarse uno o varios resultados simultáneamente, tales como el error estándar (`"se"`), el intervalo de confianza (`"ci"`), la varianza (`"var"`) o el coeficiente de variación (`"cv"`). Asimismo, `level` define el nivel de confianza utilizado para el cálculo de los intervalos de confianza, cuyo valor por defecto es 0.95. Por otro lado, el intervalo de confianza puede obtenerse directamente al incluir el argumento `vartype = "ci"` dentro de `survey_total()`, o bien utilizando la función `confint()` sobre el objeto resultante, especificando el nivel de confianza requerido. Finalmente, `deff` es un valor lógico que indica si se debe retornar el efecto del diseño (DEFF).

El siguiente ejemplo presenta cómo estimar totales y sus intervalos de confianza para diferentes variables de interés en `R`, utilizando la función `survey_total()`, cuyos resultados se presentan en la tabla 3.1:

```
survey_design %>%
  summarise(
    total = survey_total(Income,
                         vartype = c("se", "ci"),
                         deff = TRUE))
```

Tabla 3.1.: Estimación del total de ingresos de los hogares con intervalo de confianza al 95 %

total	total_se	total_low	total_upp	total_deff
85793667	4778674	76331414	95255920	11

De la tabla, se obtiene que la estimación del total poblacional para el ingreso total es de 85.793.667. Este valor corresponde al estimador puntual obtenido a partir de la encuesta, incorporando los factores de expansión y las características del diseño muestral. El error estándar asociado a esta estimación es de 4.778.674, lo que refleja la variabilidad muestral del estimador. A partir de este valor se construye un intervalo de confianza del 95 %, cuyos límites inferior y superior son 76.331.414 y 95.255.920, respectivamente. El efecto de diseño (`deff`), en el sentido desarrollado por Kish (1965), es igual a 11, lo cual sugiere una ligera pérdida de precisión asociada principalmente a las características del diseño complejo. En consecuencia, aunque el tamaño muestral nominal pueda ser grande, el tamaño muestral efectivo resultaría un poco menor debido al efecto combinado de la ponderación, la estratificación y la conglomeración.

Para continuar ilustrando el uso de la función `survey_total()`, estimemos el total de gastos de los hogares, pero ahora calculando el intervalo de confianza al 90 %. El siguiente código realiza esta estimación, presentada en la tabla 3.2:

```
survey_design %>%
  summarise(total = survey_total(
    Expenditure,
    vartype = "ci",
    level = 0.9,
    deff = TRUE
  ))
```

Tabla 3.2.: Estimación del total de gastos de los hogares con intervalo de confianza al 90 %

total	total_low	total_upp	total_deff
55677504	51360469	59994539	10.2

Si el objetivo es estimar el total de ingresos de los hogares, pero desagregado por sexo, se puede utilizar la función `group_by()` para agrupar por la variable de interés, junto con `cascade()` de la librería `srvyr`, que permite agregar una fila con el total general al final de la tabla. De esta manera, se pueden obtener fácilmente los totales por categoría y el total global dentro del mismo marco de resumen, como se muestra en la tabla 3.3.

```
survey_design %>%
  group_by(Sex) %>%
  cascade(
    total = survey_total(Income,
                         vartype = c("se", "ci"),
                         level = 0.95),
    .fill = "Total income")
```

Tabla 3.3.: Total de ingresos de los hogares desagregado por sexo

Sex	total	total_se	total_low	total_upp
Female	44153820	2324452	39551172	48756467
Male	41639847	2870194	35956576	47323118
Total income	85793667	4778674	76331414	95255920

Como se observa en los códigos anteriores, una forma práctica de obtener las estimaciones del total, su error estándar y su intervalo de confianza es mediante el argumento `vartype`, especificando las opciones `"se"` y `"ci"` respectivamente.

3.1.2. Medias

Según Gutiérrez (2016), la estimación de la media poblacional ocupa un lugar central en las encuestas de hogares, ya que muchos de los indicadores utilizados para el análisis socioeconómico corresponden a valores promedio, como el ingreso medio del hogar, el gasto per cápita o las horas promedio trabajadas. Este tipo de parámetros permite resumir el comportamiento general de las variables de interés y describir las tendencias centrales de la población objetivo. El estimador de la media poblacional puede expresarse como una razón no lineal entre dos totales poblacionales, los cuales se estiman de la siguiente manera:

$$\hat{\bar{y}} = \frac{\hat{t}_y}{\hat{N}} = \frac{\sum_h \sum_i \sum_k w_{hik}\ y_{hik}}{\sum_h \sum_i \sum_k w_{hik}}$$

De acuerdo con Särndal, Swensson, y Wretman (2003), dado que $\hat{\bar{y}}$ no es una estadística lineal, su varianza suele estimarse mediante una aproximación. Por esta razón, es necesario recurrir a métodos de remuestreo o a la linealización de Taylor para aproximar su varianza. En este caso particular, utilizando series de Taylor, la varianza se define como:

$$\widehat{Var}\left(\hat{\bar{y}}\right) \approx \frac{\widehat{Var}\left(\hat{t}_y\right) + \hat{\bar{y}}^2\ \widehat{Var}\left(\hat{N}\right) - 2\ \hat{\bar{y}}\ \widehat{Cov}\left(\hat{t}_y, \hat{N}\right)}{\hat{N}^2}$$

Como se puede observar, el cálculo de la varianza de la media poblacional implica componentes analíticos complejos, como la covarianza entre el total estimado y el tamaño poblacional estimado. Sin embargo, el paquete `srvyr` en `R` facilita estos cálculos al incorporar funciones que estiman directamente la media, su error estándar, el intervalo de confianza y el efecto del diseño. A continuación,

se presenta la sintaxis para estimar la media de los ingresos de los hogares, junto con su intervalo de confianza al 95 %, cuyos resultados se presentan en la tabla 3.4:

```
survey_design %>%
  summarise(mean = survey_mean(
    Income,
    vartype = c("se", "ci"),
    level = 0.95,
    deff = TRUE
  ))
```

Tabla 3.4.: Estimación de la media de ingresos de los hogares

mean	mean_se	mean_low	mean_upp	mean_deff
571	28.5	515	627	8.82

Como se aprecia, los argumentos utilizados en la función `survey_mean()` son similares a los de `survey_total()`. En este caso, `vartype` permite obtener el error estándar (`"se"`) y el intervalo de confianza (`"ci"`), mientras que el argumento `deff = TRUE` solicita el cálculo del efecto del diseño. La estimación indica que el ingreso medio de los hogares en la población es de aproximadamente 571 dólares. El error estándar asociado es de 28,5, lo que refleja la incertidumbre muestral de la estimación. El intervalo de confianza, cuyos límites son 515 y 627, para un nivel de confianza del 95 %. Además, el efecto de diseño estimado es de 8,82, lo que significa que la varianza de la media bajo el diseño muestral complejo es 8,82 veces mayor que la que se obtendría con un muestreo aleatorio simple de igual tamaño. Este valor evidencia una pérdida considerable de eficiencia, posiblemente asociada con la magnitud de la variabilidad del ingreso entre las unidades primarias de muestreo.

De forma análoga, es posible estimar la media de los gastos de los hogares, empleando la misma estructura de código, presentada en la tabla 3.5:

```
survey_design %>%
  summarise(mean = survey_mean(
    Expenditure,
    vartype = c("se", "ci"),
    level = 0.95,
    deff = TRUE
  ))
```

Tabla 3.5.: Estimación de la media de gastos de los hogares

mean	mean_se	mean_low	mean_upp	mean_deff
371	13.3	344	397	6.02

La media estimada del gasto es de 371 dólares, con un error estándar de 13,3. El intervalo de confianza se extiende entre 344 y 397, mientras que el efecto de diseño de 6,02 indica una alta

heterogeneidad entre las UPM de la muestra con respecto al gasto promedio. También se pueden realizar estimaciones de la media por subgrupos siguiendo el mismo esquema mostrado previamente. De manera análoga, se pueden calcular las medias del gasto por zona, las cuales son presentadas en la tabla 3.6:

```
survey_design %>%
  group_by(Zone) %>%
  cascade(
    mean = survey_mean(Expenditure,
                       level = 0.95,
                       vartype = c("se", "ci")),
    .fill = "Mean expenditure")
```

Tabla 3.6.: Media de gastos de los hogares desagregada por zona

Zone	mean	mean_se	mean_low	mean_upp
Mean expenditure	371	13.3	344	397
Rural	274	10.3	254	294
Urban	460	22.2	416	504

Al desagregar por zona, se observa que los hogares urbanos registran un gasto medio de 460, frente a 274 en los hogares rurales. Aunque los intervalos de confianza muestran mayor incertidumbre en la zona urbana, no se superponen, lo que sugiere una diferencia importante entre ambas zonas. Asimismo, es posible realizar una estimación combinada por zona y región, presentada en la tabla 3.7:

```
survey_design %>%
  group_by(Zone, Region) %>%
  cascade(
    mean = survey_mean(Expenditure,
                       level = 0.95,
                       vartype = c("se", "ci")),
    .fill = "Mean expenditure")
```

Tabla 3.7.: Media de gastos de los hogares desagregada por zona y región

Zone	Region	mean	mean_se	mean_low	mean_upp
Mean expenditure	Mean expenditure	371	13.3	344	397
Rural	Norte	307	18.4	271	343
Rural	Sur	317	14.6	288	346
Rural	Centro	337	42.8	252	422
Rural	Occidente	242	11.1	220	264
Rural	Oriente	242	22.8	197	287
Rural	Mean expenditure	274	10.3	254	294
Urban	Norte	402	48.0	307	497
Urban	Sur	497	62.5	373	621

Tabla 3.7.: Media de gastos de los hogares desagregada por zona y región

Zone	Region	mean	mean_se	mean_low	mean_upp
Urban	Centro	531	53.1	426	636
Urban	Occidente	376	20.9	334	417
Urban	Oriente	466	45.3	377	556
Urban	Mean expenditure	460	22.2	416	504

El gasto medio varía de forma importante según la zona de residencia y la región. En el área rural, los mayores promedios se observan en las regiones Centro, Sur y Norte, mientras que Occidente y Oriente presentan los valores más bajos. En el área urbana, el gasto medio es superior en todas las regiones y alcanza sus niveles más altos en Centro, Sur y Oriente. No obstante, algunas estimaciones regionales, especialmente las de Centro rural y Sur urbano, presentan errores estándar relativamente elevados, por lo que deben interpretarse con mayor cautela.

3.1.3. Razones

La razón poblacional es un parámetro no lineal definido como el cociente entre los totales de dos variables de interés. Permite cuantificar la relación entre ambas variables y resulta especialmente útil para construir indicadores relativos, comparar grupos y analizar su evolución a lo largo del tiempo. En las encuestas de hogares, este parámetro se utiliza ampliamente para calcular indicadores como el número de hombres por cada mujer, la proporción de personas ocupadas respecto de la población en edad de trabajar y la prevalencia de la subalimentación (FAO 2026).

Dado que la razón corresponde al cociente entre dos parámetros (totales) poblacionales desconocidos, tanto el numerador como el denominador deben ser estimados a partir de la muestra. Siendo t_y el total de la variable y y t_x el total de la variable x, la razón poblacional se define como $R = \frac{t_y}{t_x}$ y su estimador puntual se expresa como:

$$\hat{R} = \frac{\hat{t}_y}{\hat{t}_x} = \frac{\sum_h \sum_i \sum_k w_{hik}\, y_{hik}}{\sum_h \sum_i \sum_k w_{hik}\, x_{hik}}$$

Sin embargo, la estimación de la varianza de una razón requiere procedimientos específicos, entre los que destacan la linealización de Taylor y los métodos de remuestreo. En el contexto de encuestas complejas, la función `survey_ratio()` permite estimar razones poblacionales y obtener sus correspondientes medidas de precisión, incorporando adecuadamente las características del diseño muestral.

Como primer ejemplo de aplicación, se estima la razón entre el número de mujeres y el número de hombres en la población representada por la base de datos. Este indicador permite examinar la composición de la población por sexo y detectar posibles diferencias en la presencia relativa de ambos grupos. Aunque está estrechamente relacionado con el índice de feminidad, en este caso se expresa directamente como una razón y, por tanto, no se multiplica por 100.

Un valor igual a 1 indica que existe el mismo número de mujeres y hombres; un valor superior a 1 refleja una mayor cantidad de mujeres, mientras que un valor inferior a 1 señala una mayor cantidad de hombres. Por ejemplo, una razón de 1,05 indica que existen aproximadamente 1,05 mujeres por

cada hombre, lo que equivale a una presencia ligeramente mayor de mujeres en la población. El procedimiento de estimación y los resultados obtenidos se presentan en la tabla 3.8:

```
survey_design %>%
  summarise(ratio = survey_ratio(
    numerator = (Sex == "Female"),
    denominator = (Sex == "Male"),
    level = 0.95,
    vartype = c("se", "ci")
  ))
```

Tabla 3.8.: Razón de mujeres por hombre en la población

ratio	ratio_se	ratio_low	ratio_upp
1.11	0.035	1.04	1.18

Dado que la variable sexo en la base de datos es categórica, fue necesario acudir a variables indicadoras para realizar el cálculo de la razón, utilizando `Sex == "Female"` para identificar a las mujeres y `Sex == "Male"` para los hombres. Los resultados muestran que en la población analizada hay una mayor cantidad de mujeres que de hombres, obteniéndose una razón estimada de 1.11. Esto indica que, por cada 100 hombres, existen aproximadamente 111 mujeres. Además, el intervalo de confianza al 95 % sugiere que esta razón podría variar entre 1.04 y 1.18. La tabla 3.9 presenta la estimación de esta razón en la zona rural. Esta salida se obtuvo con la ejecución del siguiente código:

```
rural_subset <- survey_design %>%
  filter(Zone == "Rural")

rural_subset %>%
  summarise(ratio = survey_ratio(
    numerator = (Sex == "Female"),
    denominator = (Sex == "Male"),
    level = 0.95,
    vartype = c("se", "ci")
  ))
```

Tabla 3.9.: Razón de mujeres por hombre en zona rural

ratio	ratio_se	ratio_low	ratio_upp
1.07	0.035	0.997	1.14

La razón estimada entre el número de mujeres y el número de hombres en la zona rural es de 1,07, lo que indica que el número de mujeres es cerca de un 7 % mayor que el de hombres. El intervalo de confianza se extiende de 0,997 a 1,14. Dado que el intervalo de confianza incluye el valor 1, no existe

evidencia suficiente para concluir que la cantidad de mujeres difiere de la cantidad de hombres en la zona rural.

Otro ejemplo frecuente de estimación de razones en las encuestas de hogares es la razón entre el gasto y el ingreso, la cual permite analizar patrones de consumo y aproximarse a la capacidad de los hogares para sostener sus niveles de gasto (OECD 2013). A continuación, se muestra cómo estimar este indicador, cuyos resultados se presentan en la tabla 3.10:

```
survey_design %>%
  summarise(ratio = survey_ratio(
    numerator = Expenditure,
    denominator = Income,
    level = 0.95,
    vartype = c("se", "ci")
  ))
```

Tabla 3.10.: Razón gasto/ingreso de los hogares

ratio	ratio_se	ratio_low	ratio_upp
0.649	0.023	0.603	0.695

En este caso, se especificaron la variable de gasto como numerador (`numerator`), la variable de ingreso como denominador (`denominator`), el nivel de confianza (`level`) utilizado para la construcción de los intervalos de confianza y las medidas de precisión requeridas mediante el argumento (`vartype`). Como se puede observar, la razón entre el gasto y el ingreso es cercana a 0.65. Esto implica que, por cada 100 dólares que ingresan al hogar, se gastan aproximadamente 65; el intervalo de confianza al 95 % va de 0.60 a 0.69.

Ahora bien, otro análisis de interés es estimar la razón de gastos en cada una de las regiones. A continuación, se presentan los códigos computacionales, los cuales dan origen a las estimaciones presentadas en la tabla 3.11.

```
survey_design %>%
  group_by(Region) %>%
  summarise(ratio = survey_ratio(
    numerator = Expenditure,
    denominator = Income,
    level = 0.95,
    vartype = c("se", "ci")
  ))
```

Tabla 3.11.: Razón gasto/ingreso para las regiones

Region	ratio	ratio_se	ratio_low	ratio_upp
Norte	0.627	0.034	0.559	0.695
Sur	0.674	0.062	0.553	0.796

Tabla 3.11.: Razón gasto/ingreso para las regiones

Region	ratio	ratio_se	ratio_low	ratio_upp
Centro	0.752	0.043	0.667	0.837
Occidente	0.606	0.042	0.522	0.690
Oriente	0.600	0.054	0.493	0.708

La razón entre el gasto y el ingreso presenta diferencias entre las regiones. No obstante, los intervalos de confianza se superponen ampliamente, por lo que las diferencias observadas entre las regiones no necesariamente representan diferencias estadísticamente significativas. En conjunto, los resultados indican que los hogares destinan, en promedio, entre el 60 % y el 75 % de sus ingresos al gasto, con una mayor proporción en la región Centro.

3.2. Estimación de parámetros de distribución y desigualdad

El análisis de encuestas de hogares no se limita a la estimación de medidas descriptivas, pues también resulta fundamental describir la dispersión, la posición relativa de las observaciones y el grado de desigualdad de la distribución de las variables. Debido a su naturaleza no lineal, la forma funcional de los estimadores y la cuantificación de su incertidumbre deben tratarse usando procedimientos de linealización de Taylor o los métodos de replicación.

3.2.1. Varianza y desviación estándar

En las encuestas de hogares, la estimación de medidas de dispersión es fundamental para evaluar el grado de heterogeneidad de la población respecto de las variables analizadas. Ya sea que una encuesta mida el consumo de alcohol en la población adulta o el índice de masa corporal de los niños en edad escolar, una desviación estándar poblacional pequeña indica una mayor homogeneidad en la población, mientras que una desviación estándar grande indica una población más heterogénea (SAS Institute Inc. 2012).

En particular, estudiar la variabilidad de los ingresos permite identificar diferencias económicas entre los hogares y aporta información relevante para el diagnóstico de la desigualdad y el diseño de políticas públicas. Una de las medidas de dispersión más utilizadas es la desviación estándar, que cuantifica cuánto se alejan, en promedio, los valores respecto de la media. Una desviación estándar pequeña indica que los ingresos se encuentran relativamente concentrados alrededor del promedio y, por tanto, que la población es más homogénea en términos económicos. Por el contrario, una desviación estándar elevada refleja una mayor dispersión y, en consecuencia, diferencias más marcadas entre los ingresos de los hogares.

De acuerdo con Heeringa, West, y Berglund (2017), la desviación estándar poblacional se estima mediante una transformación no lineal del estimador de la varianza poblacional, $\widehat{S}_y^2$. Esta transformación incorpora, además, una corrección por grados de libertad basada en el número de observaciones válidas, denotado por n. En consecuencia, el estimador de la desviación estándar poblacional se expresa como:

$$\hat{s}_y = \sqrt{\frac{n}{n-1} \cdot \hat{S}_y^2} = \sqrt{\frac{n}{n-1} \frac{\sum_h \sum_i \sum_k w_{hik} \left(y_{hik} - \hat{\bar{y}}\right)^2}{\sum_h \sum_i \sum_k w_{hik}}}$$

El error estándar de este estimador puede aproximarse mediante linealización de Taylor o, equivalentemente, mediante el método delta [sarndal2003model]. Para estimar la desviación estándar de una variable numérica en encuestas de hogares mediante `R`, se estima primero la varianza poblacional con la función `survey_var()` y, posteriormente, se obtiene su raíz cuadrada. A continuación, se ilustra este procedimiento mediante la estimación de la desviación estándar del ingreso desagregada por zona.

```
survey_design %>%
  group_by(Zone) %>%
  summarise(
    Var = survey_var(
      Income,
      level = 0.95,
      vartype = c("se", "ci"),
      na.rm = TRUE
    )
  ) %>%
  mutate(
    Sd     = sqrt(Var),
    Sd_low = sqrt(Var_low),
    Sd_upp = sqrt(Var_upp)
  )
```

Tabla 3.12.: Desviación estándar de los ingresos desagregada por zona

Zone	Var	Var_se	Var_low	Var_upp	Sd	Sd_low	Sd_upp
Rural	96274	13794	68960	123588	310	263	352
Urban	338628	81241	177763	499494	582	422	707

Los argumentos de la función `survey_var()` son equivalentes a los utilizados previamente para la estimación de medias y totales. Los resultados de la tabla 3.12 muestran que los ingresos presentan una dispersión considerablemente mayor en la zona urbana que en la rural. En la zona rural, la desviación estándar alcanza 310 dólares, mientras que en la zona rural se eleva a 582 dólares. Además, los intervalos de confianza de las desviaciones estándar no se superponen, lo que proporciona evidencia de que la variabilidad de los ingresos es efectivamente mayor en las áreas urbanas.

3.2.2. Percentiles y medianas

Los percentiles permiten identificar puntos específicos de la distribución de una variable de interés y determinar los valores por debajo de los cuales se encuentra una proporción determinada de la población. En las encuestas de hogares, estas medidas resultan especialmente útiles para caracterizar

la distribución de variables económicas y sociales, como el ingreso, el gasto o las horas trabajadas. Por ejemplo, los percentiles de ingreso pueden utilizarse para identificar a la población perteneciente al 10 % superior de la distribución con fines tributarios (UNECE 2011), o para focalizar subsidios dirigidos a los hogares ubicados en los percentiles más bajos. La mediana, correspondinte al valor que divide a la población en dos partes iguales es poco sensible a valores extremos, razón por la cual suele considerarse una medida robusta de tendencia central.

Los percentiles constituyen un caso particular de los cuantiles. En términos generales, el cuantil de orden q (con $0 < q < 1$) es el valor que deja por debajo una proporción q de la población. Cuando esta proporción se expresa en una escala de 0 a 100, el cuantil se denomina percentil. La estimación de cuantiles en encuestas complejas se basa en el uso de estimadores ponderados y en la estimación de la función de distribución acumulada (FDA) de la población (Schulze Waltrup y Kauermann 2016). Para una población finita de tamaño N, un estimador de esta distribución está dado por:

$$\hat{F}(x) = \frac{\sum_h \sum_i \sum_k w_{hik}\, I(y_k \leq x)}{\sum_h \sum_i \sum_k w_{hik}}$$

en donde la función $I(y_k \leq x)$ es una variable indicadora que toma el valor uno si $y_k \leq x$ y cero en otro caso. Una vez estimada la FDA utilizando los pesos del diseño muestral, el cuantil q-ésimo de una variable y se define como el menor valor de y para el cual la FDA es mayor o igual que q. En particular, la mediana estimada es el valor donde la FDA estimada es mayor o igual a 0.5. Para la estimación de cuantiles en encuestas complejas, se ordenan las observaciones en orden ascendente (estadísticas de orden) de forma que $y_{(1)} \leq y_{(2)} \leq ... \leq y_{(n)}$ y se identifica el valor de j ($j = 1, ..., n$) tal que:

$$\hat{F}\left(y_{(j)}\right) \leq q \leq \hat{F}\left(y_{(j+1)}\right)$$

Bajo esta condición, el estimador del cuantil q-ésimo se obtiene mediante interpolación lineal, la cual permite obtener estimaciones más estables y precisas de los cuantiles, especialmente cuando la FDA presenta saltos importantes asociados al uso de pesos muestrales.

Woodruff (1952) propone el método de inversión de la función de distribución acumulada como procedimiento para construir intervalos de confianza de cuantiles. Por su parte, Kovar, Rao, y Wu (1988) analiza la estimación de la varianza y la construcción de intervalos de confianza para percentiles y medianas, y muestra que los métodos de replicación presentan un desempeño satisfactorio para estos parámetros no lineales en encuestas con diseños muestrales complejos.

Tanto los estimadores de cuantiles como los métodos para estimar sus varianzas se encuentran implementados en `R`. En particular, la función `survey_median()` permite estimar la mediana y obtener, de manera conjunta, su error estándar y su intervalo de confianza. Al igual que otros parámetros poblacionales, la mediana puede estimarse para diferentes dominios de estudio, como la zona geográfica, el sexo o los grupos de edad. A continuación, se presenta la sintaxis empleada para estimar la mediana del gasto de los hogares por zona a partir de la base de datos de ejemplo.

```
survey_design %>%
  group_by(Zone) %>%
  summarise(
    median = survey_median(Expenditure,
```

```
                        level = 0.95,
                        vartype = c("se", "ci")))
```

Tabla 3.13.: Estimación de la mediana de los gastos desagregada por zona

Zone	median	median_se	median_low	median_upp
Rural	241	11.0	214	258
Urban	381	19.8	337	416

A partir de la tabla 3.13 se concluye que el ingreso mediano presenta una diferencia importante entre zonas. En el área rural, la mediana estimada es de 241 dólares, lo que indica que la mitad de la población rural tiene ingresos inferiores a este valor y la otra mitad registra ingresos superiores. En la zona urbana, la mediana asciende a 381 dólares, es decir, aproximadamente un 58 % más que en la zona rural. Los intervalos de confianza no se superponen, lo que sugiere una diferencia clara en los niveles centrales de ingreso entre ambas zonas. Aunque la estimación urbana presenta un error estándar mayor, los resultados muestran consistentemente que el ingreso mediano es más elevado en las áreas urbanas.

Cuando el objetivo es estimar un percentil distinto de la mediana, puede utilizarse la función `survey_quantile()`. De acuerdo con Dorfman y Valliant (1993), algunos procedimientos para construir intervalos de confianza de cuantiles pueden presentar un desempeño menos satisfactorio que otras alternativas disponibles. En particular, el argumento `interval_type = "score"` emplea un método de inversión para determinar los límites del intervalo, en lugar de calcularlos directamente a partir de la función de influencia del percentil.

A modo de ejemplo, a continuación se presenta la sintaxis para estimar el percentil 25 del gasto. Los resultados obtenidos se muestran en la tabla 3.14.

```
survey_design %>%
  summarise(
    quantile = survey_quantile(
      Expenditure,
      quantiles = 0.25,
      level = 0.95,
      vartype = c("se", "ci"),
      interval_type = "score"
    )
  )
```

Tabla 3.14.: Estimación del primer cuartil de los gastos de los hogares

quantile_q25	quantile_q25_se	quantile_q25_low	quantile_q25_upp
200	13.8	163	218

De los resultados anteriores, el percentil 25 del gasto se estima en 200 dólares, lo que indica que aproximadamente una cuarta parte de la población presenta niveles de gasto iguales o inferiores a

este valor, mientras que el 75 % restante registra gastos superiores. El error estándar de 13,8 dólares refleja la incertidumbre asociada a la estimación.

3.2.3. El coeficiente de Gini y la curva de Lorenz

El tema sobre desigualdad económica trasciende el ámbito estrictamente estadístico y constituye uno de los temas centrales del análisis social contemporáneo. La desigualdad económica se relaciona con la persistencia de brechas en el acceso a oportunidades, la baja movilidad social y el debilitamiento de la cohesión social. En este contexto, su medición rigurosa resulta fundamental para producir diagnósticos sólidos y orientar el diseño, la implementación y la evaluación de políticas públicas integrales dirigidas a reducir las desigualdades y promover el desarrollo social inclusivo (CEPAL 2025, 32-33).

Entre los indicadores más ampliamente utilizados para cuantificar la desigualdad económica se encuentra el coeficiente de Gini (G), el cual mide el grado de concentración del ingreso comparando la distribución observada con una situación hipotética de igualdad perfecta. Este coeficiente toma valores entre 0 y 1, donde $G = 0$ representa igualdad perfecta y $G = 1$ corresponde al máximo nivel de desigualdad posible. En consecuencia, cuanto más cercano a uno sea el coeficiente, mayor será la concentración del ingreso en una proporción reducida de la población.

En el contexto de las encuestas de hogares, la estimación del coeficiente de Gini debe incorporar adecuadamente las características del diseño muestral, incluyendo los factores de expansión, la estratificación y la conglomeración. Siguiendo a David A. Binder y Kovačević (1995) y Langel y Tillé (2013), el estimador ponderado del coeficiente de Gini puede expresarse en función de la posición acumulada de cada observación en la distribución del ingreso, así:

$$\hat{G} = \frac{2 \sum_h \sum_i \sum_k w^*_{hik} \, \hat{F}_{hik} \, y_{hik}}{\hat{\bar{y}}} - 1$$

donde $w^*_{hik} = \dfrac{w_{hik}}{\sum_h \sum_i \sum_k w_{hik}}$ corresponde al peso muestral normalizado, $\hat{F}_{hik}$ representa la función de distribución acumulada (FDA) estimada para la observación k dentro del conglomerado i del estrato h, y $\hat{\bar{y}}$ denota la media ponderada de la variable de ingreso en la población.

Con respecto a la estimación de la varianza, Osier (2009) desarrolla un procedimiento de linealización que aproxima el estimador no lineal mediante una variable linealizada. A partir de este desarrollo, Jacob, Pessoa, y Damico (2024) presenta su implementación en `R` mediante la función `svygini()`. Para ello, primero se prepara el diseño muestral mediante `convey_prep()`. Luego, la función `svygini()` calcula el índice de Gini utilizando la variable de ingreso y el diseño de encuesta complejo especificado.

A continuación, se presenta la sintaxis utilizada para estimar el coeficiente de Gini en la base de datos de ejemplo, cuyos resultados se muestran en la tabla 3.15.

```
library(convey)
gini_design <- convey_prep(survey_design)
gini <- svygini( ~ Income, design = gini_design)

data.frame(
```

```
  Gini = coef(gini),
  standard_error = SE(gini),
  ci_lower = confint(gini)[1],
  ci_upper = confint(gini)[2]
)
```

Tabla 3.15.: Estimación del coeficiente de Gini para el ingreso de los hogares

	Gini	Income	ci_lower	ci_upper
Income	0.413	0.019	0.377	0.45

El coeficiente de Gini estimado indica un grado apreciable de desigualdad en la distribución del ingreso y evidencia una concentración relevante de los ingresos en una parte de la población.

Por su parte, la curva de Lorenz representa gráficamente la distribución del ingreso mediante la relación entre la proporción acumulada de la población, ordenada de menor a mayor ingreso, y la proporción acumulada del ingreso total que le corresponde (Lorenz 1905). La diagonal de 45 grados representa una situación hipotética de igualdad perfecta, en la que cada proporción acumulada de la población concentra una proporción equivalente del ingreso total.

El coeficiente de Gini se obtiene como la razón entre el área comprendida entre la curva de Lorenz y la línea de igualdad y el área total situada bajo dicha línea (Deaton 1997). Dado que esta última área es igual a un medio, el coeficiente también puede expresarse como el doble del área situada entre ambas curvas. En consecuencia, cuanto más próxima se encuentre la curva de Lorenz a la línea de igualdad, más equitativa será la distribución del ingreso; por el contrario, una mayor distancia entre ambas indica un grado más elevado de desigualdad.

En encuestas con diseños muestrales complejos, la estimación de las ordenadas de la curva de Lorenz y de sus varianzas debe incorporar las ponderaciones y las demás características del diseño muestral (Kovačević y Binder 1997). Para realizar la curva de Lorenz en `R` se utiliza la función `svylorenz()`:

```
clorenz <- svylorenz(
  formula = ~Income,
  design = gini_design,
  quantiles = seq(0, 1, .05),
  alpha = .01,
  plot = TRUE
)
```

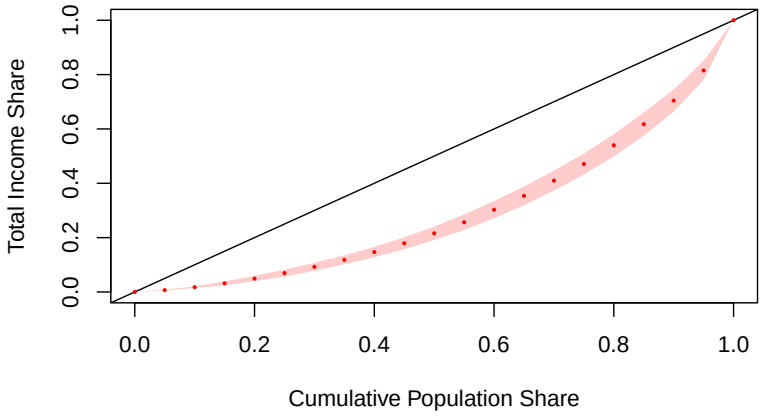

Figura 3.1.: Curva de Lorenz estimada para el ingreso de los hogares

El argumento `formula = ~Income` especifica la variable de ingreso sobre la cual se calculará la distribución acumulada. El argumento `design = gini_design` indica el objeto de diseño muestral previamente definido, el cual contiene la información relacionada con pesos, estratos y conglomerados de la encuesta. Por su parte, `quantiles = seq(0, 1, .05)` define los puntos de la distribución en los cuales se evaluará la curva de Lorenz, generando percentiles desde 0 hasta 1 en incrementos de 0.05. Finalmente, el argumento `alpha = .01` establece un nivel de significación del 1 %, lo que equivale a construir intervalos de confianza del 99 % para las estimaciones asociadas a la curva.

A partir de la figura 3.1, se observa un grado considerable de desigualdad en la distribución del ingreso per cápita, dado que la curva de Lorenz se sitúa claramente por debajo de la línea de igualdad perfecta. Su marcada curvatura muestra que los grupos ubicados en la parte inferior de la distribución concentran una proporción reducida del ingreso total, mientras que una participación considerable se acumula entre las personas de mayores ingresos. La separación entre ambas líneas se amplía progresivamente hacia los tramos superiores, lo que refleja una mayor concentración del ingreso en estos grupos. Por su parte, la banda sombreada representa la incertidumbre asociada a la estimación de la curva.

3.2.4. Correlaciones

En el estudio de encuestas de hogares, además de describir variables de forma individual, es importante analizar cómo se relacionan dos variables entre sí. Una de las herramientas más utilizadas para este propósito es el coeficiente de correlación de Pearson, el cual mide la fuerza y la dirección de la relación lineal entre dos variables numéricas. Este coeficiente toma valores entre –1 y 1. Un valor positivo indica que ambas variables tienden a aumentar simultáneamente, mientras que un valor negativo señala que cuando una variable aumenta, la otra tiende a disminuir. Por su parte, valores cercanos a cero sugieren la ausencia de una relación lineal fuerte entre las variables analizadas. Por ejemplo, en una encuesta de hogares puede ser de interés estudiar qué tan fuerte es la asociación entre el ingreso de los hogares y su nivel de gasto. Este tipo de análisis permite comprender mejor los patrones económicos y sociales en la población.

Supóngase que se dispone de dos variables numéricas de interés, denotadas por x y y, observadas en la encuesta compleja. De acuerdo con Heeringa, West, y Berglund (2017), la asociación lineal entre ambas variables puede estimarse mediante el coeficiente de correlación de Pearson ponderado, definido como la covarianza ponderada entre x y y dividida por el producto de sus desviaciones estándar ponderadas. Esta estandarización elimina la influencia de las unidades de medida y permite evaluar tanto la dirección como la intensidad de la relación lineal. Si $\hat{\bar{x}}$ y $\hat{\bar{y}}$ representan las estimaciones de las medias poblacionales de x y y, respectivamente, el coeficiente de correlación ponderado se expresa como:

$$\hat{\rho}_{xy} = \frac{\sum_h \sum_i \sum_k w_{hik}(y_{hik} - \hat{\bar{y}})(x_{hik} - \hat{\bar{x}})}{\sqrt{\sum_h \sum_i \sum_k w_{hik}(y_{hik} - \hat{\bar{y}})^2} \sqrt{\sum_h \sum_i \sum_k w_{hik}(x_{hik} - \hat{\bar{x}})^2}}$$

El paquete `survey` cuenta con la función `svyvar()` que permite obtener matrices de covarianzas ponderadas, a partir de las cuales se puede calcular la correlación. La tabla 3.16 presenta un ejemplo de estimación para la matriz de varianzas y covarianzas entre el ingreso y el gasto de los hogares.

```
cov_matrix <- svyvar(~Income + Expenditure, design = survey_design)
```

Tabla 3.16.: Matriz de varianzas y covarianzas ponderada entre ingreso y gasto

	Variable	Income	Expenditure
Income	Income	243719	98337
Expenditure	Expenditure	98337	77887

La función `svyvar` estima la matriz de varianzas y covarianzas considerando el diseño muestral. A partir de dicha matriz puede obtenerse la correlación aplicando las siguientes instrucciones:

```
cov_xy <- cov_matrix[1, 2]
sd_x <- sqrt(cov_matrix[1, 1])
sd_y <- sqrt(cov_matrix[2, 2])

weighted_cor <- cov_xy / (sd_x * sd_y)
weighted_cor
```

```
[1] 0.714
```

Si además se desea realizar inferencia estadística sobre la correlación, por ejemplo obtener errores estándar o intervalos de confianza, pueden utilizarse métodos de replicación compatibles con el paquete `survey`. Otra opción es utilizar la función `svycor()` del paquete `jtools` (Long 2026) para estimar directamente la correlación:

```r
library(jtools)

svycor(~Income + Expenditure, design = survey_design)
```

```
             Income Expenditure
Income         1.00        0.71
Expenditure    0.71        1.00
```

La correlación entre el ingreso y el gasto es relativamente alta, con un coeficiente de Pearson estimado de 0,71. Esto indica que, en términos generales, los hogares con mayores ingresos tienden a registrar también niveles de gasto más elevados. Los valores iguales a 1 en la diagonal corresponden a la correlación de cada variable consigo misma. Finalmente, este resultado refleja una asociación lineal y no debe interpretarse como evidencia de una relación causal.

3.3. Pruebas de hipótesis

En el análisis de datos provenientes de encuestas de hogares no basta con estudiar cada variable de manera aislada, sino que también es fundamental comparar grupos y evaluar si las diferencias observadas entre ellos reflejan desigualdades reales en la población o si podrían explicarse simplemente por el error de muestreo. Por ejemplo, una pregunta de interés consiste en determinar si existen diferencias estadísticamente significativas en el ingreso medio entre hogares dirigidos por hombres y hogares dirigidos por mujeres. Este tipo de análisis permite identificar brechas sociales y económicas, además de generar evidencia útil para el diseño y evaluación de políticas públicas.

Para responder este tipo de interrogantes se utilizan pruebas de hipótesis; es decir, procedimientos estadísticos que permiten contrastar afirmaciones sobre parámetros poblacionales utilizando la información observada en la muestra. En el contexto de las encuestas de hogares, estas pruebas deben incorporar adecuadamente las características del diseño muestral con el fin de garantizar inferencias válidas y resultados confiables.

En términos generales, toda prueba de hipótesis se basa en el contraste entre dos proposiciones opuestas: la hipótesis nula, denotada por H_0, y la hipótesis alternativa, denotada por H_1. En el caso más común de una prueba bilateral, estas hipótesis se expresan como:

$$\begin{cases} H_0: & \theta = \theta_0 \\ H_1: & \theta \neq \theta_0 \end{cases}$$

donde θ representa el parámetro poblacional de interés y θ_0 un valor de referencia especificado bajo la hipótesis nula. Dependiendo del objetivo del análisis, la hipótesis alternativa también puede formularse de manera unilateral, por ejemplo, cuando se desea evaluar si $\theta > \theta_0$ o si $\theta < \theta_0$. En cualquier caso, el propósito de la prueba es determinar si la evidencia contenida en la muestra resulta suficientemente fuerte para rechazar la hipótesis nula en favor de la hipótesis alternativa.

Por ejemplo, supóngase que $\bar{y}_1$ y $\bar{y}_2$ representan las medias poblacionales de una variable y en dos dominios distintos, como los hogares con jefatura masculina y femenina. En este caso, el parámetro de interés corresponde a la diferencia entre ambas medias:

$$\Delta = \bar{y}_1 - \bar{y}_2$$

Su estimador muestral se define como:

$$\hat{\Delta} = \hat{\bar{y}}_1 - \hat{\bar{y}}_2$$

Su error estándar se estima mediante la siguiente expresión:

$$\widehat{se}(\hat{\Delta}) = \sqrt{\widehat{Var}(\hat{\bar{y}}_1) + \widehat{Var}(\hat{\bar{y}}_2) - 2\widehat{Cov}(\hat{\bar{y}}_1, \hat{\bar{y}}_2)}$$

La presencia del término de covarianza es importante porque las estimaciones de ambos dominios suelen provenir de la misma muestra y, por tanto, no necesariamente son independientes. Una vez calculado el estimador y su error estándar, el contraste de hipótesis se realiza utilizando el siguiente estadístico de prueba:

$$t = \frac{\hat{\Delta}}{\widehat{se}(\hat{\Delta})} \sim t_{df}$$

Este estadístico sigue aproximadamente una distribución t de Student con (df) grados de libertad. En la práctica, estos grados de libertad suelen aproximarse mediante $df = n_I - H$, donde n_I representa el número de unidades primarias de muestreo y H el número de estratos. Bajo la hipótesis nula, valores absolutos elevados del estadístico t constituyen evidencia en contra de H_0. De manera complementaria, también puede construirse un intervalo de confianza para el parámetro Δ. Para un nivel de confianza de $(1-\alpha)100\,\%$, dicho intervalo está dado por:

$$\hat{\Delta} \pm t_{(1-\alpha/2,df)}\, \hat{se}(\hat{\Delta}).$$

Este intervalo proporciona un rango de valores plausibles para la diferencia poblacional y constituye una herramienta útil para evaluar tanto la magnitud como la significación estadística de las diferencias observadas entre grupos.

En `R`, estas pruebas pueden implementarse mediante la función `svyttest()` del paquete `survey`, la cual incorpora automáticamente los ajustes asociados al diseño muestral. El siguiente código crea primero el objeto `survey_design_heads`, que conserva únicamente las observaciones correspondientes a las personas jefas de hogar, identificadas mediante `PersonID == "idPer01"`, sin perder las características del diseño muestral. Posteriormente, la función `svyttest()` compara el ingreso medio entre las categorías de la variable `Sex`. El argumento `level = 0.95` establece un nivel de confianza del 95 % para los intervalos.

```
survey_design_heads  <- survey_design %>%
  filter(PersonID == "idPer01")

ttest_sex <- svyttest(Income ~ Sex,
                      design = survey_design_heads,
                      level = 0.95)
```

Tabla 3.17.: Prueba de diferencia de ingresos medios por sexo (población total)

statistic	p_value	gl	ci_lower	ci_upper	estimated_difference
2.76	0.007	118	34.7	210	122

A partir de los resultados presentados en la tabla 3.17, se estima que el ingreso medio de los hombres jefes de hogar es 122 dólares mayor al de las muejeres jefes de hogar. Dado que el valor p obtenido es 0,007, inferior al nivel de significación establecido de $\alpha = 0,05$, se rechaza la hipótesis nula de igualdad de medias y se concluye que la diferencia observada es estadísticamente significativa. Esta conclusión es consistente con el intervalo de confianza del 95 %, que no incluye el valor cero. En conjunto, los resultados aportan evidencia de que, en la población representada por la encuesta, las mujeres jefas de hogar presentan un ingreso medio significativamente menor al de los hombres jefes de hogar.

El procedimiento descrito no se limita a las medias, sino que también puede aplicarse a proporciones, totales, razones o cualquier función diferenciable de totales. En todos los casos, el contraste se fundamenta en la estimación puntual, su varianza (incluyendo covarianzas cuando corresponde) y la comparación con la distribución t ajustada al diseño muestral.

3.4. Estimación de contrastes

En el análisis de encuestas de hogares es frecuente que el interés no se limite a comparar únicamente dos subpoblaciones, sino varias de manera simultánea. Por ejemplo, puede ser necesario comparar los ingresos medios de los hogares entre distintas regiones, áreas geográficas o grupos poblacionales, con el propósito de identificar diferencias y establecer patrones de desigualdad entre ellos. En este tipo de situaciones, las pruebas de diferencia de medias presentadas anteriormente resultan limitadas, ya que están diseñadas para comparar únicamente pares de poblaciones.

Según Heeringa, West, y Berglund (2017), para abordar comparaciones más generales se utilizan los contrastes, que constituyen una herramienta flexible para evaluar combinaciones lineales de parámetros poblacionales. En términos generales, un contraste se define como:

$$f\left(\theta_1, \theta_2, \ldots, \theta_J\right) = \sum_{j=1}^{J} a_j \theta_j,$$

donde los coeficientes a_j son constantes conocidas y θ_j representan los parámetros de interés. Este enfoque permite formular y evaluar una amplia variedad de comparaciones entre grupos, incluyendo diferencias entre medias, comparaciones múltiples y combinaciones más complejas de parámetros.

Por ejemplo, si se quiere comparar los ingresos medios de dos subpoblaciones particulares (las regiones Norte y Sur) se puede usar el contraste $\bar{y}_{Norte} - \bar{y}_{Sur}$. Dado que la muestra cubre cinco regiones en total, el contraste debe construirse asignando coeficientes únicamente a las regiones involucradas en la comparación y dejando coeficientes iguales a cero para las demás regiones. De esta manera, el contraste se define como:

$$1 \times \hat{\bar{y}}_{Norte} + (-1) \times \hat{\bar{y}}_{Sur} + 0 \times \hat{\bar{y}}_{Centro} + 0 \times \hat{\bar{y}}_{Occidente} + 0 \times \hat{\bar{y}}_{Oriente}.$$

Esta expresión indica que el contraste corresponde a la diferencia entre las medias estimadas de las regiones Norte y Sur, mientras que las demás regiones no participan en la comparación. De forma matricial, el contraste puede escribirse como:

$$[1, -1, 0, 0, 0] \times \begin{bmatrix} \hat{\bar{y}}_{Norte} \\ \hat{\bar{y}}_{Sur} \\ \hat{\bar{y}}_{Centro} \\ \hat{\bar{y}}_{Occidente} \\ \hat{\bar{y}}_{Oriente} \end{bmatrix}.$$

En consecuencia, el vector de contraste asociado a esta comparación es $[1, -1, 0, 0, 0]$, donde los coeficientes positivos y negativos indican las poblaciones que se desean comparar y los ceros representan las poblaciones excluidas del contraste. En R, estos procedimientos se encuentran disponibles mediante la función `svycontrast()` del paquete `survey`. Para esto, es necesario crear un objeto con las medias de los ingresos desagregados por región, como se presenta en la tabla 3.18:

```
region_mean_contrast <- svyby(
  formula = ~ Income,
  by = ~ Region,
  design = survey_design,
  FUN = svymean,
  na.rm = TRUE,
  covmat = TRUE,
  vartype = c("se", "ci")
)
```

Tabla 3.18.: Ingreso medio estimado por región

	Region	Income	se	ci_l	ci_u
Norte	Norte	552	55.4	444	661
Sur	Sur	626	62.4	503	748
Centro	Centro	651	61.5	530	771
Occidente	Occidente	517	46.2	426	608
Oriente	Oriente	542	71.7	401	682

La función `svyby()` estima el ingreso medio por región, de acuerdo con el diseño de muestreo `survey_design`. El argumento `formula = ~Income` especifica la variable numérica de interés, mientras que `by = ~Region` define los dominios para los cuales se realizará la estimación. La opción `FUN = svymean` indica que debe calcularse la media del ingreso en cada región y `na.rm = TRUE` excluye las observaciones con valores faltantes. Por su parte, `vartype = c("se", "ci")` solicita el error estándar y el intervalo de confianza de cada media estimada. Finalmente, `covmat = TRUE` calcula la matriz de varianzas y covarianzas entre las estimaciones regionales, lo que permite realizar posteriormente contrastes o comparaciones entre ellas.

Luego, la función `svycontrast()` devuelve el contraste estimado y su error estándar, como se muestra en la tabla 3.19. Los argumentos de esta función son los promedios de los ingresos estimados (`stat`) y las constantes de contraste (`contrasts`).

```
contrast_region_ns <- svycontrast(
  stat = region_mean_contrast,
  contrasts = list(north_south_difference = c(1, -1, 0, 0, 0))
)
```

Tabla 3.19.: Contraste de ingresos medios: región Norte menos región Sur

contrast	estimate	variance	standard_error
north_south_difference	-73.4	6959	83.4

La estimación de la diferencia de medias indica que el ingreso promedio en la región Norte es aproximadamente 73,4 dólares inferior al observado en la región Sur. Sin embargo, el error estándar de 83,4 es mayor que la diferencia estimada, lo que refleja una incertidumbre considerable. Por tanto, no hay evidencia suficiente para concluir que exista una diferencia estadísticamente significativa entre ambas regiones. Nótese que estos mismos resultados podrían obtenerse realizando explícitamente el proceso de construcción del estimador del contraste y de su varianza estimada. Si se consideran únicamente los ingresos medios estimados de las regiones Norte y Sur, su diferencia es:

```
# Paso 1: diferencia de las estimaciones (Norte - Sur)
region_income_mean <- setNames(
  as.numeric(region_mean_contrast[["Income"]]),
  region_mean_contrast[["Region"]]
)

region_income_mean["Norte"] - region_income_mean["Sur"]
```

```
Norte
-73.4
```

El paso siguiente consiste en calcular la matriz de varianzas y covarianzas y extraer de ella los elementos correspondientes a las regiones Norte y Sur, los cuales se presentan en la tabla 3.20. Las covarianzas estimadas entre las regiones son todas nulas, lo que resulta coherente con el diseño muestral, dado que la estratificación se realizó por región y la selección de la muestra se efectuó de manera independiente en cada estrato. En consecuencia, los estimadores regionales no comparten unidades muestrales y su covarianza bajo el diseño es igual a cero.

```
# Paso 2: matriz de varianzas y covarianzas
region_vcov <- vcov(region_mean_contrast)
```

Tabla 3.20.: Matriz de varianzas y covarianzas de los ingresos medios por región

Region	Norte	Sur	Centro	Occidente	Oriente
Norte	3065	0	0	0	0

Tabla 3.20.: Matriz de varianzas y covarianzas de los ingresos medios por región

Region	Norte	Sur	Centro	Occidente	Oriente
Sur	0	3894	0	0	0
Centro	0	0	3778	0	0
Occidente	0	0	0	2136	0
Oriente	0	0	0	0	5136

Para calcular el error estándar de la diferencia (contraste) se usarán las propiedades de la varianza, de la siguiente forma:

$$\widehat{se}\left(\hat{\bar{y}}_{Norte} - \hat{\bar{y}}_{Sur}\right) = \sqrt{\widehat{Var}\left(\hat{\bar{y}}_{Norte}\right) + \widehat{Var}\left(\hat{\bar{y}}_{Sur}\right) - 2\,\widehat{Cov}\left(\hat{\bar{y}}_{Norte}, \hat{\bar{y}}_{Sur}\right)}$$

Para evitar inconsistencias con valores redondeados, el cálculo se realiza directamente a partir de la matriz estimada:

```
sqrt(region_vcov["Norte", "Norte"] +
       region_vcov["Sur", "Sur"] -
       2 * region_vcov["Norte", "Sur"])
```

```
[1] 83.4
```

Este resultado corresponde al mismo obtenido anteriormente con la función `svycontrast()`.

Además de comparar únicamente dos poblaciones, los contrastes permiten evaluar simultáneamente varias comparaciones de interés entre diferentes grupos. Por ejemplo, supóngase que se desea comparar los ingresos medios entre distintas regiones geográficas. En particular, podrían considerarse contrastes como $\bar{y}_{Norte} - \bar{y}_{Centro}$, $\bar{y}_{Sur} - \bar{y}_{Centro}$ y $\bar{y}_{Occidente} - \bar{y}_{Oriente}$. Cada una de estas expresiones representa un contraste lineal específico entre medias regionales. De manera conjunta, dichos contrastes pueden organizarse mediante la siguiente matriz de contrastes:

$$\begin{bmatrix} 1 & 0 & -1 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1 \end{bmatrix}$$

En esta matriz, cada fila define un contraste distinto y cada columna corresponde a una de las regiones consideradas en el análisis. Los coeficientes positivos y negativos indican las medias que participan en cada comparación, mientras que los ceros representan las regiones que no intervienen en el contraste correspondiente. A continuación, se muestra la implementación de los contrastes en `R`, cuyos resultados se presentan en la tabla 3.21.

```
contrast_regions <- svycontrast(
  stat = region_mean_contrast,
  contrasts = list(
    north_south = c(1, 0, -1, 0, 0),
    south_center = c(0, 1, -1, 0, 0),
    west_east = c(0, 0, 0, 1, -1)
  )
)
```

Tabla 3.21.: Contrastes de ingresos medios entre regiones

contrast	estimate	variance	standard_error
north_south	-98.4	6843	82.7
south_center	-25.0	7673	87.6
west_east	-24.7	7272	85.3

También es posible construir contrastes entre variables asociadas a diferentes grupos poblacionales, como ocurre al comparar el ingreso medio según el sexo del jefe de hogar. Siguiendo el mismo procedimiento del ejemplo anterior, y haciendo uso del objeto `survey_design_heads`, el primer paso consiste en estimar el ingreso medio para cada grupo, en este caso hombres y mujeres, como se presenta en la tabla 3.22.

```
sex_mean_contrast <- svyby(
  formula = ~ Income,
  by = ~ Sex,
  design = survey_design_heads,
  FUN = svymean,
  na.rm = TRUE,
  covmat = TRUE,
  vartype = c("se", "ci")
)
```

Tabla 3.22.: Ingreso medio estimado por sexo

Sex	Income	se	ci_l	ci_u
Female	442	23.5	396	488
Male	564	34.0	498	631

En este caso, el contraste de interés está dado por $\bar{y}_F - \bar{y}_M$, el cual permite cuantificar la diferencia entre el ingreso medio de las mujeres y el correspondiente a los hombres. A continuación, utilizando la función `svycontrast()`, se obtiene la estimación del contraste, cuyos resultados se presentan en la tabla 3.23. A partir de estos resultados, se concluye que, en promedio, los hombres jefes de hogar perciben 122 dólares más que las mujeres jefes de hogar, con un error estándar estimado de 44,3 dólares.

```
contrast_sex <- svycontrast(
  stat = sex_mean_contrast,
  contrasts = list(sex_difference = c(1, -1))
)
```

Tabla 3.23.: Contraste de ingresos medios entre mujeres y hombres

contrast	estimate	variance	standard_error
sex_difference	-122	1958	44.3

Dado que los contrastes corresponden a funciones lineales de parámetros, también es posible aplicarlos a estimadores de razón. Un ejemplo de ello consiste en comparar la relación entre ingresos y gastos según el sexo del jefe de hogar. En este caso, primero se estima la razón ingreso-gasto para cada grupo y posteriormente se construye el contraste entre dichas razones. A continuación, se presentan los códigos computacionales utilizados para realizar este análisis, cuyos resultados se muestran en la tabla 3.24:

```
sex_ratio_contrast <- svyby(
  formula = ~ Income,
  by = ~ Sex,
  denominator = ~ Expenditure,
  design = survey_design_heads,
  FUN = svyratio,
  na.rm = TRUE,
  covmat = TRUE,
  vartype = c("se", "ci")
)
```

Tabla 3.24.: Razón ingreso/gasto desagregada por sexo

	Sex	Income/Expenditure	se.Income/Expenditure	ci_l	ci_u
Female	Female	1.42	0.051	1.31	1.52
Male	Male	1.59	0.072	1.45	1.74

La función `svycontrast()` permite construir el contraste entre estas razones estimadas a partir de diseños de muestreo complejos. A continuación, se presentan los códigos computacionales utilizados para realizar este análisis, cuyos resultados se muestran en la tabla 3.25. A partir de dichos resultados, puede concluirse que la diferencia entre las razones estimadas es de 0.178 a favor de los hombres.

```
contrast_sex_ratio <- svycontrast(
  stat = sex_ratio_contrast,
  contrasts = list(sex_difference = c(1, -1))
)
```

Tabla 3.25.: Contraste de la razón ingreso/gasto entre sexos

contrast	estimate	standard_error	variance
sex_difference	-0.178	0.088	0.00774

3.5. Visualización de variables continuas

La representación gráfica de variables continuas constituye un componente fundamental del análisis de encuestas de hogares. Los gráficos bien diseñados permiten examinar la forma de las distribuciones, identificar tendencias y explorar relaciones entre variables, lo que facilita tanto la interpretación de los resultados como su comunicación a distintos públicos. En este contexto, es importante que las representaciones gráficas incorporen, cuando corresponda, las características del diseño muestral (Lumley 2010). En particular, el uso de los factores de expansión permite que la distribución representada refleje adecuadamente a la población de interés, en lugar de limitarse a describir únicamente las observaciones incluidas en la muestra.

La construcción de los gráficos presentados en este docuemnto se basa en el enfoque conocido como la gramática de los gráficos, el cual concibe cada visualización como la combinación de distintos componentes que se agregan: los datos, las correspondencias entre variables y atributos visuales, los elementos geométricos, las escalas, las facetas y el sistema de coordenadas (Wilkinson 2005). Este enfoque se implementa en `R` mediante el paquete `ggplot2`, que permite construir representaciones gráficas de manera sistemática y flexible (Wickham, Chang, et al. 2026). Las estimaciones obtenidas con los paquetes `survey` y `srvyr`, incluidos los factores de expansión y las medidas de incertidumbre, se utilizan como insumos para elaborar los gráficos. Cuando es necesario presentar varias visualizaciones de manera conjunta, el paquete `patchwork` permite organizarlas y combinarlas en una composición gráfica coherente (Pedersen 2025).

```
library(ggplot2)
library(patchwork)
```

3.5.1. Histogramas

Un histograma es una representación gráfica de la distribución de una variable numérica continua, construida a partir de rectángulos cuya altura es proporcional a la frecuencia de observaciones dentro de determinados intervalos. En el contexto de las encuestas de hogares, los histogramas ponderados permiten aproximar la distribución de la variable de interés en la población, incorporando los factores de expansión asociados al diseño muestral.

A continuación, se presenta el histograma ponderado de la variable `Income`, construido incorporando los pesos muestrales (`wk`) con el fin de representar la distribución estimada del ingreso en la población. Los resultados obtenidos se muestran en la figura 3.2:

```
weighted_income_hist <- ggplot(
  data = survey_data,
  aes(x = Income, weight = wk)) +
```

```
  geom_histogram(aes(y = ..density..)) +
  ylab("") +
  ggtitle("Histograma ponderado")

weighted_income_hist
```

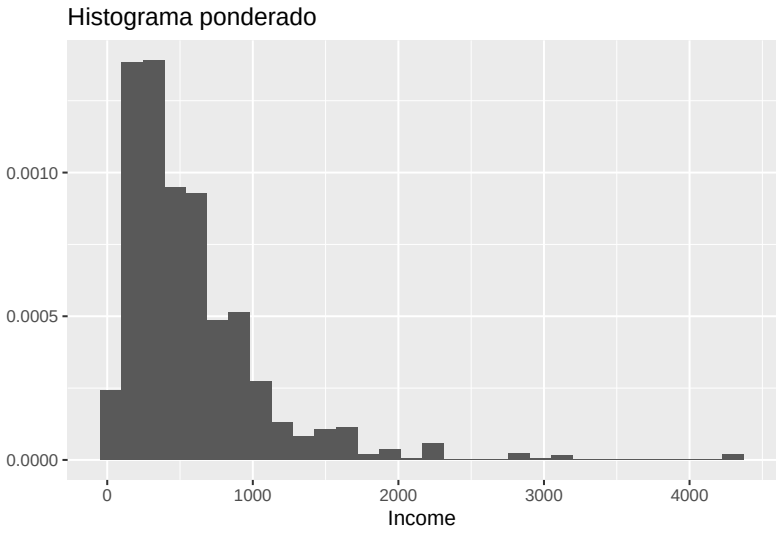

Figura 3.2.: Histograma ponderado del ingreso de los hogares

En este código, se inicia la construcción del gráfico utilizando la base de datos `survey_data`. Dentro de `aes()`, el argumento `x = Income` indica que la variable representada en el eje horizontal corresponde al ingreso, mientras que `weight = wk` incorpora los pesos muestrales para construir un histograma ponderado que refleje la distribución estimada en la población. La función `geom_histogram()` genera el histograma y, mediante `aes(y = ..density..)`, la altura de las barras se expresa en términos de densidad en lugar de frecuencias absolutas. Posteriormente, `ylab("")` elimina la etiqueta del eje vertical, y `ggtitle("Histograma ponderado")` agrega un título descriptivo al gráfico.

Los histogramas también permiten comparar visualmente la distribución de una variable entre distintos subgrupos de la población mediante el argumento `fill`, el cual asigna un color diferente a cada categoría. Por ejemplo, es posible comparar la distribución del ingreso entre zonas urbanas y rurales, como se muestra en la figura 3.3:

```
zone_colors <- c(Urban = "#48C9B0", Rural = "#117864")

weighted_zone_hist <- ggplot(survey_data, aes(x = Income, weight = wk)) +
  geom_histogram(
    aes(y = ..density.., fill = Zone),
    alpha = 0.5, position = "identity") +
  scale_fill_manual(values = zone_colors) +
  ylab("") +
  ggtitle("Weighted")
```

```
weighted_zone_hist
```

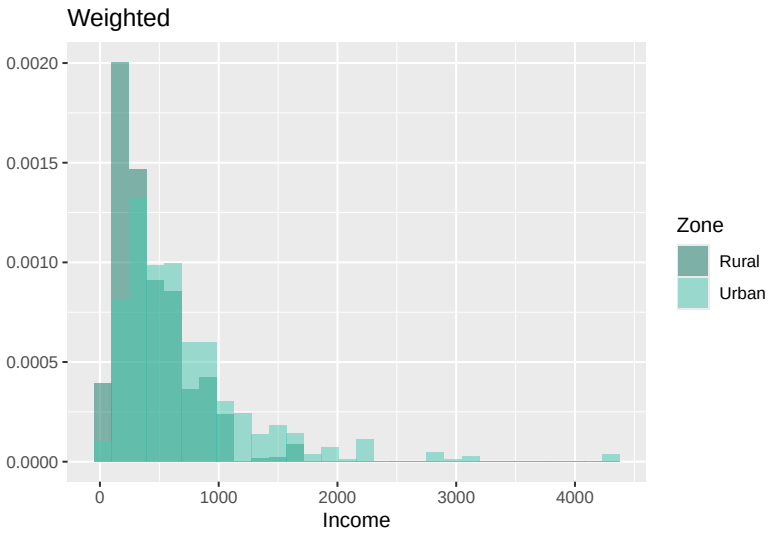

Figura 3.3.: Histograma ponderado del ingreso por zona geográfica

En este código, se construye un histograma ponderado utilizando la base de datos `survey_data`, donde `x = Income` define la variable representada en el eje horizontal y `weight = wk` incorpora los pesos muestrales para reflejar la distribución poblacional del ingreso. La función `geom_histogram()` genera el histograma y, mediante `aes(y = ..density.., fill = Zone)`, representa las barras en términos de densidad y asigna colores diferentes según la variable `Zone`, permitiendo comparar visualmente la distribución del ingreso entre zonas. El argumento `alpha = 0.5` agrega transparencia a las barras para facilitar la superposición de los histogramas, mientras que `position = "identity"` permite que las distribuciones se dibujen una encima de otra en lugar de apilarse. Posteriormente, `scale_fill_manual(values = zone_colors)` define manualmente los colores asociados a cada zona. Finalmente, `ylab("")` elimina la etiqueta del eje vertical, `ggtitle("Weighted")` agrega un título al gráfico.

Además, es posible superponer curvas de densidad suavizadas mediante la función `geom_density()`, lo que permite representar de manera continua la forma de la distribución de la variable analizada. Este tipo de gráfico facilita la identificación de características como la asimetría, la concentración de los valores y la posible presencia de múltiples modas. En la figura 3.4 se presentan los histogramas del ingreso y el gasto junto con sus respectivas curvas de densidad estimadas. Para disponer ambos gráficos uno al lado del otro, se utiliza el operador `|` del paquete `patchwork`, que permite combinar distintas visualizaciones en una misma composición gráfica.

```
weighted_income_hist +
  geom_density(fill = "blue", alpha = 0.3) |
  weighted_zone_hist +
  geom_density(aes(fill = Zone), alpha = 0.3) +
  theme(legend.position = "top")
```

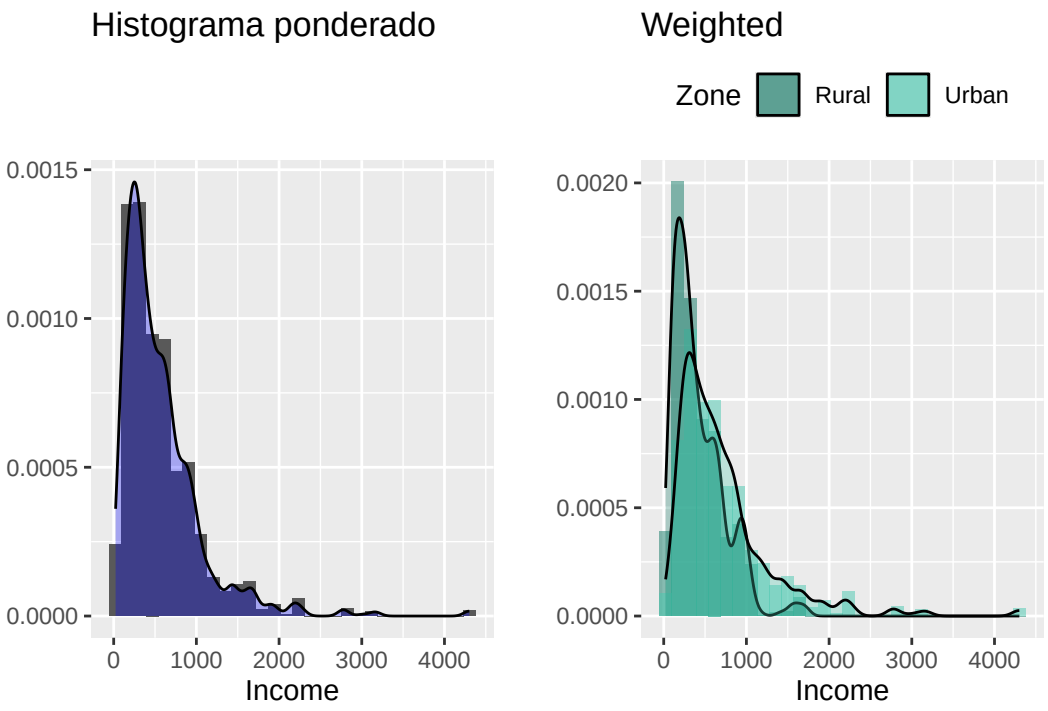

Figura 3.4.: Histogramas con curvas de densidad superpuestas para ingreso y gasto

3.5.2. Diagramas de caja (*boxplots*)

Introducidos por Tukey (1977), los diagramas de caja (*boxplots*) permiten identificar valores atípicos y evaluar la dispersión de la distribución. Son una representación gráfica que resume una variable mediante estadísticas basadas en cuartiles, permitiendo identificar la mediana, la dispersión de los datos y la presencia de valores atípicos. En el contexto de encuestas de hogares, este tipo de gráfico resulta útil para comparar distribuciones entre distintos grupos poblacionales de manera compacta y visualmente intuitiva.

Para construir *boxplots* ponderados en `ggplot2` se utiliza la función `geom_boxplot()`, como se muestra en la figura 3.5. En este caso, la función `ggplot()` inicia la construcción del gráfico utilizando la base de datos `survey_data`, donde `x = Income` define la variable numérica analizada y `weight = wk` incorpora los pesos muestrales para representar la distribución estimada en la población. Posteriormente, `geom_boxplot()` genera los diagramas de caja y, mediante `aes(fill = Zone)` o `aes(fill = Sex)`, asigna colores diferentes a cada categoría de las variables `Zone` y `Sex`, respectivamente, facilitando la comparación visual entre grupos. La función `coord_flip()` invierte los ejes para presentar los *boxplots* de forma horizontal, mejorando así su legibilidad. Por su parte, `scale_fill_manual()` permite definir manualmente los colores asociados a cada categoría mediante los vectores `zone_colors` y `sex_colors`.

```
weighted_zone_boxplot <-
  ggplot(survey_data, aes(x = Income, weight = wk)) +
  geom_boxplot(aes(fill = Zone)) +
  ggtitle("Weighted") +
```

```
  coord_flip() +
  scale_fill_manual(values = zone_colors)

sex_colors <- c(Male = "#5DADE2", Female = "#2874A6")

weighted_sex_boxplot <-
  ggplot(survey_data, aes(x = Income, weight = wk)) +
  geom_boxplot(aes(fill = Sex)) +
  ggtitle("Weighted") +
  coord_flip() +
  scale_fill_manual(values = sex_colors)

weighted_zone_boxplot | weighted_sex_boxplot
```

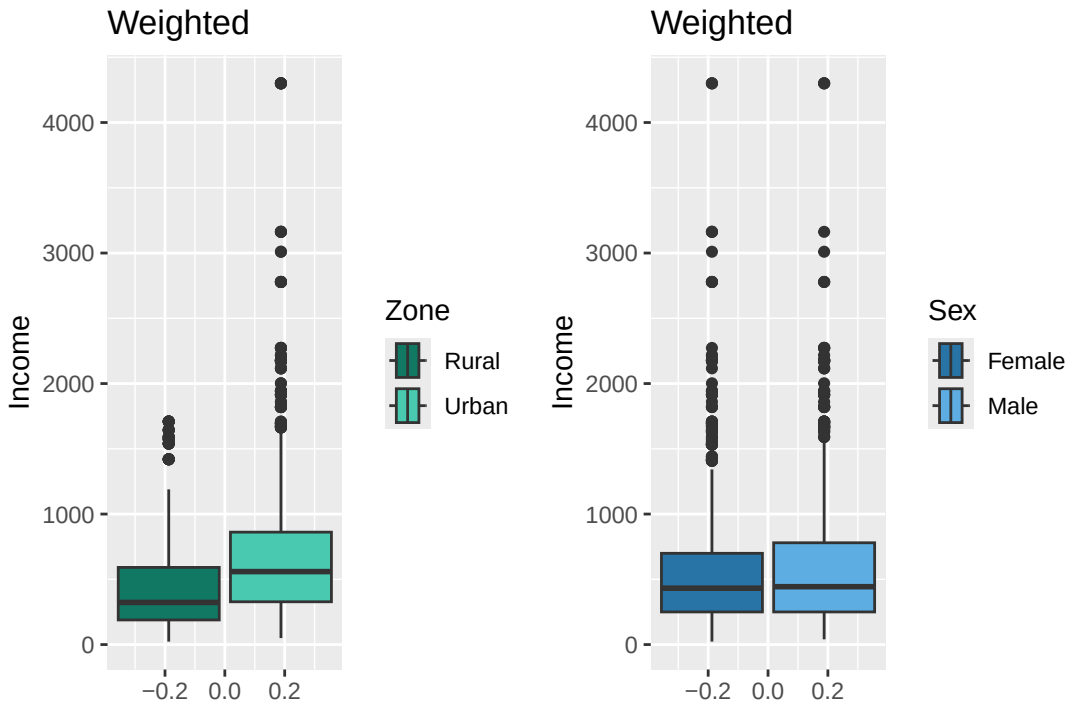

Figura 3.5.: Boxplots ponderados del ingreso por zona y sexo

Además de las herramientas disponibles en `ggplot2`, las funciones especializadas `svyhist()` y `svyboxplot()` del paquete `survey` incorporan directamente las características del diseño muestral complejo en la construcción de los gráficos. Estas funciones permiten obtener representaciones gráficas coherentes con el diseño de la encuesta, facilitando el análisis exploratorio de las variables numéricas. Un ejemplo de su aplicación se presenta en la figura 3.6.

```
old_par <- par(no.readonly = TRUE)
par(mfrow = c(1, 2))
svyhist(
```

```
  ~ Income,
  survey_design,
  main = "Household income",
  col  = "grey80",
  xlab = "Income",
  probability = FALSE
)

svyboxplot(
  Income ~ Zone,
  survey_design,
  col  = "grey80",
  ylab = "Income",
  xlab = "Zone"
)
par(old_par)
```

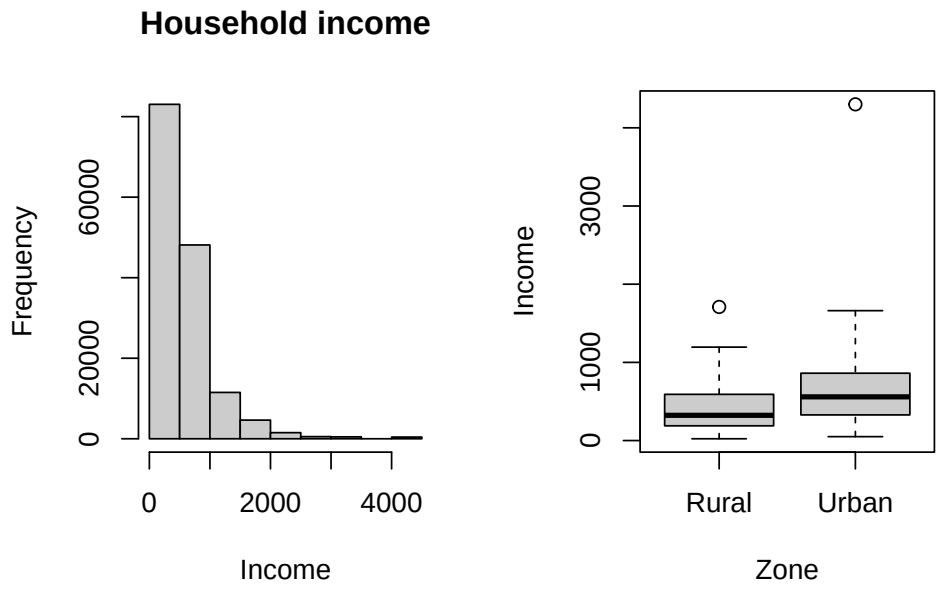

Figura 3.6.: Histograma y boxplot del ingreso ajustados al diseño muestral complejo

3.5.3. Diagramas de dispersión

Los diagramas de dispersión permiten explorar visualmente la relación entre dos variables continuas, facilitando la identificación de patrones de asociación, tendencias y posibles valores atípicos. En el contexto de encuestas con diseños muestrales complejos, resulta importante incorporar la información de los pesos de muestreo para que la representación gráfica refleje adecuadamente la contribución relativa de cada observación en la población.

Cuando se trabaja con conjuntos de datos de gran tamaño, representar todos los puntos simultáneamente puede generar gráficos difíciles de interpretar debido a la sobreposición de observaciones. Sin embargo, en muestras de tamaño moderado, una estrategia sencilla y efectiva consiste en representar los pesos mediante el tamaño de los símbolos en el gráfico de dispersión. De esta manera, las observaciones con mayor peso poblacional aparecerán resaltadas visualmente. Un ejemplo de este tipo de representación se presenta en la figura 3.7:

```
weighted_basic_scatter <- ggplot(
  data = survey_data,
  aes(y = Income, x = Expenditure)) +
  geom_point(aes(size = wk), alpha = 0.3)

weighted_basic_scatter
```

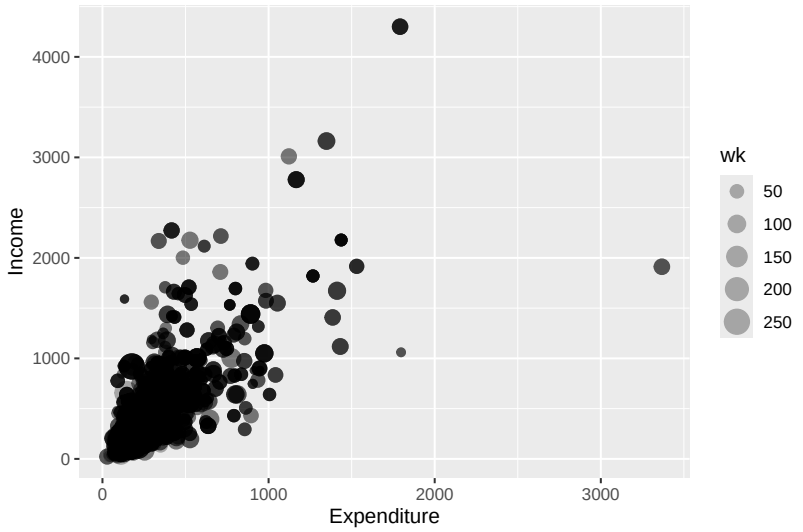

Figura 3.7.: Diagrama de dispersión entre ingreso y gasto de los hogares

En este código, se construye un diagrama de dispersión entre las variables `x = Expenditure` y `y = Income`. La función `geom_point()` agrega los puntos del gráfico y, mediante `aes(size = wk)`, utiliza los pesos muestrales (`wk`) para controlar el tamaño de cada punto, de modo que las observaciones con mayor representación poblacional aparezcan visualmente más destacadas. El argumento `alpha = 0.3` incorpora transparencia a los puntos, reduciendo el problema de sobreposición y facilitando la visualización de áreas con alta concentración de observaciones.

Adicionalmente, es posible incorporar variables de agrupamiento en los diagramas de dispersión mediante los argumentos `shape` y `color`, los cuales permiten diferenciar visualmente las observaciones según categorías específicas de la población. Esto facilita la identificación de patrones diferenciados entre grupos y mejora la interpretación de la relación entre las variables analizadas. Un ejemplo de este tipo de representación se presenta en la figura 3.8.

```
weighted_zone_scatter <- ggplot(
  data = survey_data,
```

```
  aes(y = Income, x = Expenditure, shape = Zone)) +
  geom_point(aes(size = wk, color = Zone), alpha = 0.3) +
  labs(size = "Peso") +
  scale_color_manual(values = zone_colors)

weighted_zone_scatter
```

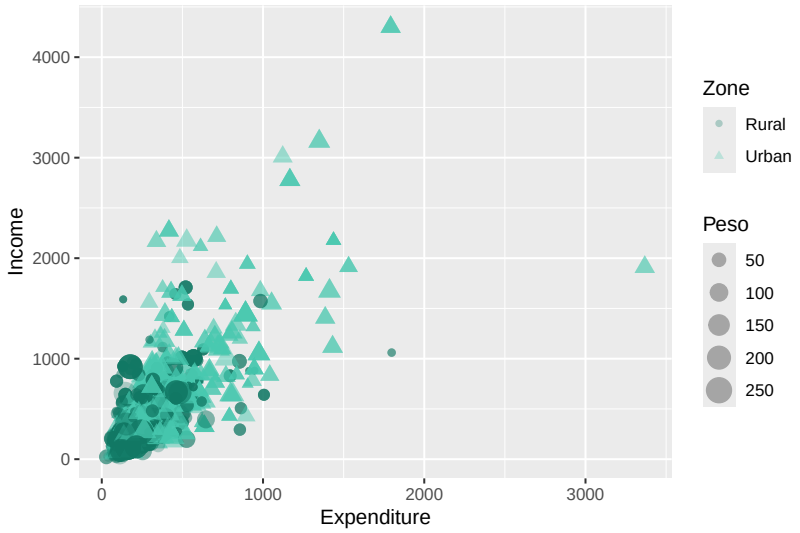

Figura 3.8.: Diagrama de dispersión entre ingreso y gasto por zona geográfica

4. Análisis de variables categóricas

En el análisis de encuestas de hogares, uno de los productos más habituales es la estimación de parámetros descriptivos asociados a variables categóricas, los cuales permiten sintetizar las principales características de la población. Estas medidas facilitan una representación clara y comprensible de fenómenos sociales a partir de información recolectada mediante muestras probabilísticas (CEPAL 2023).

Como explica Agresti (2019), las frecuencias y las proporciones se encuentran entre los indicadores descriptivos más utilizados para analizar variables categóricas. Las frecuencias representan el número de personas u hogares que pertenecen a cada categoría, como la cantidad de personas en condición de pobreza, mientras que las proporciones expresan el peso relativo de cada categoría dentro de la población, como el porcentaje de hogares en situación de hacinamiento. Aunque el análisis descriptivo de este tipo de variables se apoya principalmente en estos parámetros, también puede complementarse con medidas derivadas de variables numéricas, como los cuantiles y los indicadores de desigualdad abordados en el capítulo anterior.

Para estimar correctamente estas medidas a partir de una encuesta de hogares, es necesario comenzar por definir el diseño de muestreo. Este paso permite incorporar las características de la encuesta, incluidos los factores de expansión, las unidades primarias de muestreo y los estratos de selección. La adecuada especificación de estos elementos es fundamental para que las frecuencias, las proporciones y las demás estadísticas estimadas representen apropiadamente a la población objetivo y para que sus medidas de precisión reflejen la variabilidad introducida por el diseño muestral.

```
library(tidyverse)
library(survey)
library(srvyr)

survey_data <- readRDS("../Data/encuesta.rds")
options(survey.lonely.psu = "adjust")

survey_design <- survey_data %>%
  as_survey_design(
    strata = Stratum,
    ids = PSU,
    weights = wk,
    nest = TRUE
  )
```

A partir de la información original de la encuesta, se construyen algunas variables categóricas que serán utilizadas en los ejercicios desarrollados a lo largo de este capítulo. En particular, se define una variable dicotómica que identifica si una persona se encuentra o no en condición de pobreza, considerando como pobres a quienes pertenecen a cualquiera de las categorías de pobreza registradas

en la encuesta. Asimismo, se crea una variable que permite distinguir a las personas desempleadas del resto de la población y otra que clasifica a las personas en dos grupos de edad: menores de 18 años y personas de 18 años o más. Estas variables facilitarán la desagregación de los resultados y la comparación de indicadores entre distintos grupos de la población:

```
survey_design <- survey_design %>%
  mutate(
    poor = ifelse(Poverty != "NotPoor", 1, 0),
    unemployed = ifelse(Employment == "Unemployed", 1, 0),
    age_18 = case_when(
      Age < 18 ~ "< 18 years",
      TRUE ~ ">= 18 years"
    )
  )
```

En el código anterior se crean nuevas variables derivadas utilizando la función `mutate()` del paquete `dplyr`. Las variables `poor` y `unemployed` se construyen mediante la función `ifelse()`, que evalúa una condición lógica y asigna un valor dependiendo de si esta se cumple o no. En este caso, las variables toman el valor `1` cuando la persona se encuentra en condición de pobreza o desempleo, respectivamente, y `0` en caso contrario. Por su parte, la variable `age_18` se genera mediante la función `case_when()`, la cual permite definir categorías a partir de una o varias condiciones de forma más clara y flexible que `ifelse()`. Esta función resulta especialmente útil cuando se requiere construir variables categóricas con múltiples niveles. En el ejemplo, las personas menores de 18 años se clasifican en la categoría `"< 18 years"`, mientras que el resto se agrupa en la categoría `">= 18 years"`.

Además de las variables categóricas definidas previamente, es necesario establecer subpoblaciones que permitan obtener estimaciones desagregadas y comparar los indicadores entre distintos grupos. Para los ejemplos desarrollados en este capítulo, se consideran cuatro subpoblaciones definidas según la zona de residencia y el sexo: población urbana, población rural, mujeres y hombres.

```
urban_subset <- survey_design %>%
  filter(Zone == "Urban")

rural_subset <- survey_design %>%
  filter(Zone == "Rural")

female_subset <- survey_design %>%
  filter(Sex == "Female")

male_subset <- survey_design %>%
  filter(Sex == "Male")
```

4.1. Estimación del tamaño poblacional

En el análisis de encuestas de hogares, es fundamental estimar el tamaño de las distintas subpoblaciones y la proporción que cada una representa dentro de la población total (CEPAL 2023). Estas

estimaciones permiten cuantificar grupos definidos por características demográficas o socioeconómicas y describir su importancia relativa. Por ejemplo, conocer el número y la proporción de personas que se encuentran por debajo de la línea de pobreza, están desempleadas o han alcanzado un determinado nivel educativo contribuye a identificar brechas entre grupos, evaluar sus condiciones de bienestar y generar información relevante para el diseño de políticas públicas.

Las subpoblaciones se definen mediante variables categóricas que clasifican a las unidades de la población finita en grupos mutuamente excluyentes, como las distintas condiciones de actividad económica o los niveles educativos alcanzados. El tamaño de una subpoblación corresponde al número total de personas u hogares de la población que pertenecen a una categoría determinada. Para estimarlo a partir de una encuesta, se suman los pesos muestrales de las unidades clasificadas en esa categoría, dado que cada peso indica cuántas personas u hogares de la población representa la unidad observada en la muestra. A partir de estas estimaciones también puede calcularse la proporción que cada subpoblación representa respecto del total poblacional. Por ende, el estimador del tamaño de la población (Särndal, Swensson, y Wretman 2003), se define como:

$$\hat{N} = \sum_h \sum_i \sum_k w_{hik}$$

En donde, w_{hik} es el peso o factor de expansión de la unidad k en la UPM i del estrato h.

Asimismo, la estimación del tamaño de una subpoblación sigue el mismo principio que la estimación del tamaño total, pero se restringe a las unidades que presentan una característica determinada. Para estimar cuántas personas u hogares pertenecen a una categoría específica d, se identifica en la muestra a las unidades que cumplen dicha condición y se suman sus pesos muestrales. De esta manera, cada unidad contribuye a la estimación de acuerdo con el número de unidades de la población que representa:

$$\hat{N}_d = \sum_h \sum_i \sum_k w_{hik}\ I(y_{hik} = d)$$

donde $\hat{N}_d$ representa el tamaño estimado de la subpoblación perteneciente a la categoría d. Por su parte, $I(y_{hik} = d)$ es una variable binaria que toma el valor de 1 si la unidad k de la UPM i en el estrato h pertenece a la categoría d de la variable de interés y, y 0 en caso contrario. Nótese además que, si d fue utilizada en la calibración de los pesos, el valor de $\hat{N}_d$ coincidirá con el control externo aplicado y por ende su error estándar será nulo (Gutiérrez 2016).

En `R`, utilizando el diseño muestral definido previamente, es posible, generar resúmenes de la información de la encuesta para cada categoría de la variable `Zone`. Para ello, los datos se agrupan según la zona geográfica, lo que permite obtener resultados independientes para cada uno de estos dominios de estudio. A partir de este procedimiento se producen dos tipos principales de resultados. En primer lugar, se calcula el número de observaciones efectivamente recolectadas en la muestra, sin considerar los factores de expansión, usando la función `unweighted()`. En segundo lugar, se estima el tamaño poblacional asociado a cada dominio incorporando la información del diseño de muestreo. Junto con las estimaciones poblacionales también se obtienen medidas de precisión, como el error estándar y el intervalo de confianza. Los resultados se presentan en la tabla 4.1.

```r
survey_design %>%
  group_by(Zone) %>%
  summarise(
    n = unweighted(n()),
    size = survey_total(vartype = c("se", "ci"))
  )
```

Tabla 4.1.: Tamaño poblacional estimado por zona geográfica

Zone	n	size	size_se	size_low	size_upp
Rural	1297	72102	3062	66039	78165
Urban	1308	78164	2847	72526	83802

En efecto, el tamaño de muestra fue de 1297 personas en la zona rural y de 1308 en la zona urbana. A partir de estas muestras se estimó una población de 72,102 personas para la zona rural, con un error estándar de 3,062, y de 78,164 personas para la zona urbana, con un error estándar de 2,847. Considerando un nivel de confianza del 95 %, los intervalos de confianza para las estimaciones poblacionales fueron de 66,038.5 a 78,165.4 personas en la zona rural y de 72,526.2 a 83,801.7 personas en la zona urbana. De manera análoga, también es posible estimar el número de personas según los distintos niveles de pobreza.

```r
survey_design %>%
  group_by(Poverty) %>%
  summarise(size = survey_total(vartype = c("se", "ci")))
```

Tabla 4.2.: Tamaño poblacional estimado por condición de pobreza

Poverty	size	size_se	size_low	size_upp
NotPoor	91398	4395	82696	100101
Extreme	21519	4949	11719	31319
Relative	37349	3695	30032	44666

La tabla 4.2 refelja una población en la que la mayoría de las personas no vive en pobreza, pero en la que esa aparente normalidad convive con una fractura difícil de considerar marginal: cerca de 58.868 personas se encuentran en pobreza relativa o extrema. De ellas, 21.519 padecen las privaciones más severas y 37.349 permanecen por debajo de los niveles de vida socialmente aceptables.

Otra variable categórica relevante en las encuestas de hogares es el estado de ocupación. En este caso, la información se agrupa según las categorías de la variable `Employment`, lo que permite obtener estimaciones separadas para cada condición de actividad laboral de la población. Mediante el uso de la función `group_by()` es posible desagregar las estimaciones en niveles más detallados, facilitando el análisis comparativo entre los distintos grupos ocupacionales. En primer lugar, se excluyen los registros con valores faltantes en la variable de empleo, garantizando que los resultados se calculen únicamente con personas activas en la fuerza de trabajo. Posteriormente, para cada categoría de ocupación se estima el tamaño poblacional, junto con el error estándar y los intervalos de confianza.

```
survey_design %>%
  filter(!is.na(Employment)) %>%
  group_by(Employment) %>%
  summarise(size = survey_total(vartype = c("se", "ci")))
```

Tabla 4.3.: Tamaño poblacional estimado por estado de ocupación

Employment	size	size_se	size_low	size_upp
Unemployed	4635	761	3129	6141
Inactive	41465	2163	37183	45748
Employed	61877	2540	56847	66907

A partir de los resultados presentados en la tabla 4.3, se estima que la población ocupada asciende a 61.877 personas. Por su parte, la población inactiva se estima en 41.465 personas, mientras que cerca de 4.635 personas se encontrarían desempleadas. En conjunto, las estimaciones muestran que los desempleados representan el grupo de menor tamaño, aunque la amplitud relativa de su intervalo de confianza indica una mayor incertidumbre en comparación con las estimaciones de la población ocupada e inactiva.

Finalmente, las estimaciones también pueden desagregarse simultáneamente por más de una variable categórica. Por ejemplo, es posible analizar el estado ocupacional según el nivel de pobreza, cuyos resultados se presentan en la tabla 4.4, de donde se puede concluir que el empleo constituye una protección importante frente a la pobreza. Entre las 61.877 personas ocupadas, 17.277 continúan en situación de pobreza, de las cuales 5.128 se encuentran en pobreza extrema y 12.149 en pobreza relativa. La vulnerabilidad es aún más visible entre las 41.465 personas inactivas, pues 17.119 viven en pobreza, mientras que entre las 4.635 personas desempleadas apenas 1.768 son clasificadas como no pobres y 2.866 padecen pobreza extrema o relativa.

```
survey_design %>%
  filter(!is.na(Employment)) %>%
  group_by(Employment, Poverty) %>%
  cascade(
    size = survey_total(vartype = c("se", "ci")),
    .fill = "Total"
  )
```

Tabla 4.4.: Tamaño poblacional estimado por estado de ocupación y condición de pobreza

Employment	Poverty	size	size_se	size_low	size_upp
Unemployed	NotPoor	1768	405	966	2571
Unemployed	Extreme	1169	348	480	1859
Unemployed	Relative	1697	458	791	2604
Unemployed	Total	4635	761	3129	6141
Inactive	NotPoor	24346	1736	20908	27784
Inactive	Extreme	6422	1321	3807	9037

Tabla 4.4.: Tamaño poblacional estimado por estado de ocupación y condición de pobreza

Employment	Poverty	size	size_se	size_low	size_upp
Inactive	Relative	10697	1460	7806	13589
Inactive	Total	41465	2163	37183	45748
Employed	NotPoor	44600	2596	39460	49741
Employed	Extreme	5128	1122	2907	7349
Employed	Relative	12149	1347	9483	14816
Employed	Total	61877	2540	56847	66907
Total	Total	107977	3466	101115	114839

4.2. Estimación de proporciones

Las proporciones permiten expresar el peso relativo que tienen determinados grupos dentro de la población. Por ejemplo, conocer el porcentaje de hogares que se encuentran por debajo de la línea de pobreza es esencial para evaluar desigualdades y caracterizar condiciones de bienestar. Para obtener este tipo de estimaciones, se calcula el promedio ponderado de la variable dicotómica, lo que asegura que la estimación represente adecuadamente la distribución poblacional. Al convertir las categorías originales en variables indicadoras, la proporción estimada se calcula como:

$$\hat{p}_d = \frac{\hat{N}_d}{\hat{N}} = \frac{\sum_h \sum_i \sum_k w_{hik}\, I(y_{hik} = d)}{\sum_h \sum_i \sum_k w_{hik}}$$

Como explica Gutiérrez (2016), si bien este estimador es conceptualmente sencillo, corresponde a una función no lineal de los totales poblacionales. Por esta razón, para derivar sus propiedades estadísticas, particularmente su varianza y error estándar, es necesario recurrir a métodos de aproximación como la linealización de Taylor. En este contexto, se utiliza la función $z_{hik} = I(y_{hik} = d) - \hat{p}_d$. En la práctica, los programas estadísticos implementan automáticamente estos procedimientos y generan las estimaciones de proporciones junto con sus respectivos errores estándar e intervalos de confianza.

Cuando las proporciones estimadas toman valores cercanos a 0 o a 1, su precisión puede disminuir y las aproximaciones estadísticas tradicionales pueden resultar poco adecuadas. En estos casos, suele ser necesario disponer de tamaños de muestra mayores para obtener estimaciones suficientemente estables. Este problema adquiere especial importancia al analizar fenómenos poco frecuentes o muy concentrados, ya que pequeñas variaciones en la composición de la muestra pueden ocasionar cambios sustanciales en las proporciones estimadas (Edward L. Korn y Graubard 1999).

Esta dificultad también afecta la construcción de los intervalos de confianza. Como señala Agresti (2019), los intervalos basados directamente en la aproximación normal pueden producir límites inferiores menores que 0 o límites superiores mayores que 1, resultados incompatibles con el rango de una proporción. Una alternativa consiste en construir el intervalo en la escala logit y transformarlo posteriormente a la escala original. Para una proporción estimada $\hat{p}_d$, el intervalo de confianza con nivel $1 - \alpha$ puede expresarse como

$$CI(\hat{p}_d; 1-\alpha) = \left[\frac{\exp(L_d)}{1+\exp(L_d)}, \frac{\exp(U_d)}{1+\exp(U_d)} \right],$$

En donde

$$L_d = \log\left(\frac{\hat{p}_d}{1-\hat{p}_d}\right) t_{1-\alpha/2,df} \frac{\widehat{SE}(\hat{p}_d)}{\hat{p}_d(1-\hat{p}_d)}$$

y

$$U_d = \log\left(\frac{\hat{p}_d}{1-\hat{p}_d}\right) + t_{1-\alpha/2,df} \frac{\widehat{SE}(\hat{p}_d)}{\hat{p}_d(1-\hat{p}_d)}.$$

Al aplicar la transformación inversa del logit a los límites L_d y U_d, se garantiza que el intervalo resultante permanezca dentro del rango válido entre 0 y 1. Nótese que el término $t_{1-\alpha/2,df}$ representa el percentil asociado al nivel de confianza seleccionado, considerando df grados de libertad. Como se advirtió anteriormente, en el contexto de encuestas complejas, estos grados de libertad suelen definirse como la diferencia entre el número de unidades primarias de muestreo (UPM) y el número de estratos presentes en el subgrupo o dominio de análisis.

A diferencia de otros programas estadísticos que presentan los resultados en porcentaje, `R` reporta las proporciones en la escala [0,1]. Las funciones utilizadas permiten calcular directamente las estimaciones, sus errores estándar e intervalos de confianza, lo que facilita el análisis de variables categóricas en encuestas de hogares. A continuación se ilustra cómo obtener la proporción de personas por zona usando el diseño muestral previamente definido.

```
survey_design %>%
  group_by(Zone) %>%
  summarise(proportion = survey_mean(
    vartype = c("se", "ci"),
    proportion = TRUE
  ))
```

Tabla 4.5.: Proporción de personas por zona geográfica

Zone	proportion	proportion_se	proportion_low	proportion_upp
Rural	0.48	0.014	0.452	0.508
Urban	0.52	0.014	0.492	0.548

En este ejemplo, la función `survey_mean()` se utiliza para calcular la estimación de proporciones poblacionales y, mediante el argumento `proportion = TRUE`, se especifica que el interés se centra en estimar la distribución relativa de la población entre las distintas categorías de la variable analizada. Los resultados presentados en la tabla 4.5 indican que aproximadamente el 48 % de las personas reside en la zona rural, con un intervalo de confianza del 95 % entre 45.2 % y 50.8 %, mientras que el 52 % corresponde a población residente en la zona urbana, con un intervalo de confianza entre 49.2 % y 54.8 %.

La librería `srvyr` también dispone de la función `survey_prop()`, diseñada específicamente para la estimación de proporciones en encuestas complejas. Esta función produce resultados equivalentes a los obtenidos con `survey_mean()`, facilitando además una sintaxis más directa e intuitiva cuando el objetivo del análisis es exclusivamente el cálculo de proporciones. Además, si el interés se centra ahora en estimar proporciones para subpoblaciones específicas, por ejemplo, la proporción de hombres y mujeres que residen en la zona urbana, es necesario restringir el análisis a dicho dominio de estudio y posteriormente calcular las estimaciones correspondientes para cada categoría de sexo.

```
urban_subset %>%
  group_by(Sex) %>%
  summarise(proportion = survey_prop(vartype = c("se", "ci")))
```

Tabla 4.6.: Proporción de hombres y mujeres en zona urbana

Sex	proportion	proportion_se	proportion_low	proportion_upp
Female	0.537	0.013	0.511	0.563
Male	0.463	0.013	0.437	0.489

La tabla 4.6 muestra una leve mayoría femenina en la población urbana: las mujeres representan el 53,6 %, con un intervalo de confianza entre 51,0 % y 56,2 %, frente al 46,4 % de los hombres, cuyo intervalo se sitúa entre 43,7 % y 48,9 %.

Ahora bien, si el análisis se restringe únicamente a la población masculina incluida en la base de datos y el interés consiste en estimar la distribución porcentual de los hombres según la zona de residencia, es posible calcular la proporción de hombres que viven en áreas urbanas y rurales utilizando el diseño muestral previamente definido.

```
male_subset %>%
  group_by(Zone) %>%
  summarise(proportion = survey_prop(vartype = c("se", "ci")))
```

Tabla 4.7.: Distribución por zona en la subpoblación masculina

Zone	proportion	proportion_se	proportion_low	proportion_upp
Rural	0.491	0.018	0.455	0.526
Urban	0.509	0.018	0.474	0.545

Los resultados de la tabla 4.7 revelan una población masculina repartida casi por igual entre lo urbano y lo rural: el 50,9 % reside en áreas urbanas, con un intervalo de confianza entre 47,4 % y 54,5 %, mientras que el 49,1 % habita en zonas rurales, con un intervalo entre 45,5 % y 52,6 %.

La función `group_by()` permite generar distintos niveles de desagregación mediante la combinación de dos o más variables categóricas. Por ejemplo, es posible estimar la proporción de hombres según zona de residencia y condición de pobreza de manera simultánea. Este tipo de análisis permite caracterizar con mayor detalle las condiciones socioeconómicas de subpoblaciones específicas.

```r
male_subset %>%
  group_by(Zone, Poverty) %>%
  summarise(proportion = survey_prop(vartype = c("se", "ci")))
```

Tabla 4.8.: Proporción de hombres por zona y condición de pobreza

Zone	Poverty	proportion	proportion_se	proportion_low	proportion_upp
Rural	NotPoor	0.549	0.063	0.424	0.668
Rural	Extreme	0.198	0.067	0.096	0.364
Rural	Relative	0.254	0.037	0.187	0.334
Urban	NotPoor	0.660	0.037	0.584	0.728
Urban	Extreme	0.113	0.025	0.073	0.171
Urban	Relative	0.227	0.026	0.180	0.283

La tabla 4.8 revela una geografía social particular. Mientras en las zonas urbanas el 66,0 % de los hombres se encuentra fuera de la pobreza, en las rurales esta proporción desciende al 54,9 %; mientras que la pobreza extrema aumenta del 11,3 % al 19,8 % al pasar de las zonas urbanas a las rurales. La pobreza relativa, en cambio, permanece más cercana entre ambos territorios.

Las variables cuantitativas también pueden agruparse en categorías para facilitar su análisis conjunto con variables cualitativas. Por ejemplo, la edad puede clasificarse en rangos etarios y cruzarse posteriormente con la condición de actividad, lo que permite examinar cómo varía la participación laboral entre distintos grupos de población. En este caso, la población femenina se divide en dos categorías: mujeres de 18 a 35 años y mujeres pertenecientes a los demás grupos de edad. A partir de esta clasificación, se estima la distribución de la condición laboral en cada grupo etario.

```r
female_subset %>%
  filter(!is.na(Employment)) %>%
  mutate(age_range = case_when(
    Age >= 18 & Age <= 35 ~ "18 - 35",
    TRUE ~ "Other"
  )) %>%
  group_by(age_range, Employment) %>%
  summarise(proportion = survey_prop(vartype = c("se", "ci")))
```

Tabla 4.9.: Proporción de mujeres por rango de edad y estado de ocupación

age_range	Employment	proportion	proportion_se	proportion_low	proportion_upp
18 - 35	Unemployed	0.029	0.009	0.015	0.054
18 - 35	Inactive	0.517	0.038	0.442	0.591
18 - 35	Employed	0.455	0.036	0.385	0.526
Other	Unemployed	0.016	0.007	0.007	0.036
Other	Inactive	0.571	0.030	0.511	0.629
Other	Employed	0.413	0.029	0.356	0.472

La tabla 4.9 muestra que la inactividad constituye la condición predominante entre las mujeres de ambos grupos de edad: alcanza el 51,7 % entre aquellas de 18 a 35 años y el 57,1 % entre las demás. Las mujeres más jóvenes presentan una mayor proporción de ocupación: 45,5 %, frente a 41,3 %, pero también un desempleo algo más elevado, de 2,9 % frente a 1,6 %.

4.3. Contrastes y diferencias de proporciones

Cuando se analizan conjuntamente dos variables categóricas, es posible examinar cómo se distribuye una de ellas dentro de los grupos definidos por la otra. Por ejemplo, puede estudiarse la proporción de personas en situación de pobreza por sexo, zona de residencia o condición de actividad. Cada una de estas categorías define una subpoblación específica, para la cual pueden estimarse proporciones y sus correspondientes medidas de precisión (Edward L. Korn y Graubard 1999; Heeringa, West, y Berglund 2017).

Sin embargo, la estimación separada de estas proporciones no siempre es suficiente. Con frecuencia, el interés principal consiste en determinar cuánto difieren dos grupos y si la brecha observada es suficientemente grande en relación con la incertidumbre asociada al muestreo. Para ello, pueden construirse contrastes lineales que expresen la diferencia entre dos proporciones estimadas y permitan obtener su error estándar, intervalo de confianza y, cuando corresponda, una prueba de significancia.

Supóngase, por ejemplo, que se desea comparar la proporción de mujeres en situación de pobreza con la proporción correspondiente a los hombres. Este contraste puede expresarse como $\hat{p}_F - \hat{p}_M$, donde $\hat{p}_F$ representa la proporción estimada de mujeres en pobreza y $\hat{p}_M$ la proporción estimada de hombres en la misma condición.

```
sex_poverty_contrast <- svyby(
  formula = ~poor,
  by = ~Sex,
  design = survey_design,
  FUN = svymean,
  na.rm = TRUE,
  covmat = TRUE,
  vartype = c("se", "ci")
)
```

Tabla 4.10.: Proporción en condición de pobreza por sexo

	Sex	poor	se	ci_l	ci_u
Female	Female	0.389	0.032	0.327	0.451
Male	Male	0.395	0.037	0.323	0.466

Con base en los resultados presentados en la tabla 4.10, puede estimarse la diferencia entre las proporciones de mujeres y hombres en situación de pobreza, así como su correspondiente error estándar. Para ello, se resta la proporción estimada de hombres en pobreza de la proporción estimada de mujeres en la misma condición:

```
sex_poverty_estimate <- setNames(
  as.numeric(sex_poverty_contrast[["poor"]]),
  sex_poverty_contrast[["Sex"]]
)

sex_poverty_estimate["Female"] - sex_poverty_estimate["Male"]
```

```
  Female
-0.00532
```

Para estimar correctamente la variabilidad asociada a esta diferencia debe incorporarse la covarianza entre ambas estimaciones, ya que estas no provienen de dos muestras independientes. En efecto, hombres y mujeres pueden pertenecer a los mismos hogares y, además, los hogares de ambos grupos son seleccionados dentro de las mismas unidades primarias de muestreo (UPM). Por consiguiente, las dos proporciones comparten la estructura de dependencia inducida por el diseño de muestreo. Esto implica que las variaciones observadas en una estimación pueden estar relacionadas con las de la otra y, por tanto, que su covarianza no puede suponerse igual a cero. En este contexto, el error estándar de la diferencia (contraste) debe calcularse aplicando las propiedades de la varianza e incorporando explícitamente dicha covarianza, de la siguiente forma:

$$\widehat{se}\left(\hat{p}_F - \hat{p}_M\right) = \sqrt{\widehat{Var}\left(\hat{p}_F\right) + \widehat{Var}\left(\hat{p}_M\right) - 2\,\widehat{Cov}\left(\hat{p}_F, \hat{p}_M\right)}$$

En R, la matriz de varianzas y covarianzas puede obtenerse mediante la función `vcov`, cuyos resultados se presentan en la tabla 4.11:

```
sex_poverty_vcov <- vcov(sex_poverty_contrast)
```

Tabla 4.11.: Matriz de varianzas y covarianzas de la proporción de pobreza por sexo

	Female	Male
Female	0.000998	0.000918
Male	0.000918	0.001342

A partir de esta matriz, el error estándar de la diferencia de proporciones se estima, utilizando la expresión general para la varianza de una diferencia entre estimadores, la cual corresponde a la suma de las varianzas de cada estimación menos dos veces la covarianza entre ellas.

```
sqrt(
  sex_poverty_vcov["Female", "Female"] +
    sex_poverty_vcov["Male", "Male"] -
    2 * sex_poverty_vcov["Female", "Male"]
)
```

```
[1] 0.0224
```

No obstante, el paquete `survey` permite realizar este procedimiento de forma directa mediante la función `svycontrast`, la cual calcula tanto el contraste lineal como su error estándar asociado. Para obtener la diferencia entre las proporciones de mujeres y hombres en condición de pobreza, se utiliza el siguiente código:

```
svycontrast(
  stat = sex_poverty_contrast,
  contrasts = list(sex_difference = c(1, -1))
) %>%
  data.frame()
```

```
               contrast sex_difference
sex_difference -0.00532         0.0224
```

De lo que se concluye que la diferencia estimada entre las proporciones de mujeres y hombres en estado de pobreza es -0.005 (-0.5 %), con un error estándar de 0.022. Dada la magnitud del error estándar respecto de la diferencia, la estimación no debe interpretarse por sí sola como evidencia de una menor proporción de pobreza entre las mujeres.

Otro ejercicio que puede desarrollarse a partir de una encuesta de hogares consiste en estimar la proporción de personas desempleadas según la región de residencia. Para ello, se estiman las proporciones de desempleo para cada una de las regiones consideradas en el análisis. Los resultados obtenidos se presentan en la tabla 4.12:

```
region_unemployment_contrast <- svyby(
  formula = ~ unemployed,
  by = ~ Region,
  design = survey_design %>%
    filter(!is.na(unemployed)),
  FUN = svymean,
  na.rm = TRUE,
  covmat = TRUE,
  vartype = c("se", "ci")
)
```

Tabla 4.12.: Proporción en condición de desempleo por región

	Region	unemployed	se	ci_l	ci_u
Norte	Norte	0.049	0.020	0.010	0.088
Sur	Sur	0.066	0.024	0.019	0.112
Centro	Centro	0.039	0.012	0.014	0.063
Occidente	Occidente	0.040	0.012	0.016	0.064
Oriente	Oriente	0.030	0.013	0.005	0.054

Una vez estimadas las proporciones de desempleo por región, el siguiente paso consiste en evaluar si existen diferencias estadísticamente significativas entre algunas de estas proporciones mediante

contrastes lineales. En particular, el interés se centra en comparar las diferencias entre las proporciones estimadas de desempleo de distintas regiones. Los contrastes considerados corresponden a $\hat{p}_{Norte} - \hat{p}_{Centro} = 0.01004$, $\hat{p}_{Sur} - \hat{p}_{Centro} = 0.02691$ y $\hat{p}_{Occidente} - \hat{p}_{Oriente} = 0.01046$. De manera matricial, estos contrastes entre regiones pueden escribirse así:

$$\begin{bmatrix} 1 & 0 & -1 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1 \end{bmatrix}$$

Para calcular los errores estándar asociados a cada uno de los contrastes anteriores, es necesario considerar no solo las varianzas de las estimaciones regionales, sino también las covarianzas entre ellas.

```
region_unemployment_vcov <- vcov(region_unemployment_contrast)
```

Tabla 4.13.: Matriz de varianzas y covarianzas de la proporción de desempleo por región

	Norte	Sur	Centro	Occidente	Oriente
Norte	0.000401	0.000000	0.000000	0.000000	0.000000
Sur	0.000000	0.000564	0.000000	0.000000	0.000000
Centro	0.000000	0.000000	0.000154	0.000000	0.000000
Occidente	0.000000	0.000000	0.000000	0.000151	0.000000
Oriente	0.000000	0.000000	0.000000	0.000000	0.000158

La matriz de varianzas y covarianzas presentada en la tabla 4.13 muestra que las covarianzas entre las estimaciones regionales son iguales a cero, lo que refleja la independencia entre las muestras seleccionadas en cada región. A continuación se estiman los errores estándar asociados a acada contraste.

```
# Norte - Centro
sqrt(
  region_unemployment_vcov["Norte", "Norte"] +
    region_unemployment_vcov["Centro", "Centro"] -
    2 * region_unemployment_vcov["Norte", "Centro"]
)
```

```
[1] 0.0236
```

```
# Sur - Centro
sqrt(
  region_unemployment_vcov["Sur", "Sur"] +
    region_unemployment_vcov["Centro", "Centro"] -
    2 * region_unemployment_vcov["Sur", "Centro"]
)
```

```
[1] 0.0268
```

```
# Occidente - Oriente
sqrt(
  region_unemployment_vcov["Occidente", "Occidente"] +
    region_unemployment_vcov["Oriente", "Oriente"] -
    2 * region_unemployment_vcov["Occidente", "Oriente"]
)
```

```
[1] 0.0176
```

Alternativamente, el paquete `survey` permite calcular estos contrastes de manera directa mediante la función `svycontrast`, la cual obtiene tanto las diferencias estimadas entre proporciones como sus respectivos errores estándar. En este caso, la estimación de los contrastes definidos anteriormente se realiza de la siguiente manera:

```
svycontrast(
  stat = region_unemployment_contrast,
  contrasts = list(
    north_center = c(1, 0, -1, 0, 0),
    south_center = c(0, 1, -1, 0, 0),
    west_east    = c(0, 0, 0, 1, -1)
  )) %>%
  data.frame()
```

```
             contrast     SE
north_center   0.0100 0.0236
south_center   0.0269 0.0268
west_east      0.0105 0.0176
```

4.4. Pruebas de hipótesis sobre proporciones

Las pruebas de hipótesis también permiten evaluar si las diferencias observadas entre proporciones estimadas para distintos grupos reflejan diferencias en la población o si pueden atribuirse al error de muestreo (Edward L. Korn y Graubard 1995). Supóngase que p_H representa la proporción de hombres desocupados y p_M la proporción de mujeres desocupadas dentro de la población económicamente activa. El parámetro de interés corresponde a la diferencia entre ambas proporciones:

$$\Delta_p = p_H - p_M.$$

Para evaluar si las proporciones de desocupación de hombres y mujeres son diferentes, se plantea el siguiente sistema de hipótesis bilateral:

$$\begin{cases} H_0: & \Delta_p = 0 \\ H_1: & \Delta_p \neq 0 \end{cases}$$

El estimador muestral de esta diferencia se define como $\hat{\Delta}_p = \hat{p}_H - \hat{p}_M$. Su error estándar debe considerar tanto las varianzas de las proporciones estimadas como la covarianza entre ellas:

$$\widehat{se}(\hat{\Delta}_p) = \sqrt{\widehat{Var}(\hat{p}_H) + \widehat{Var}(\hat{p}_M) - 2\widehat{Cov}(\hat{p}_H, \hat{p}_M)}.$$

A partir de estas cantidades, el estadístico de prueba se expresa como:

$$t = \frac{\hat{\Delta}_p}{\widehat{se}(\hat{\Delta}_p)} \sim t_{df},$$

el cual sigue aproximadamente una distribución t de Student con df grados de libertad, determinados por el diseño de muestreo. Bajo la hipótesis nula, valores absolutos elevados del estadístico t proporcionan evidencia en contra de la igualdad de las proporciones.

Supóngase que el objetivo es evaluar, mediante una prueba de hipótesis, si la tasa de desempleo difiere significativamente entre hombres y mujeres. Para implementar esta prueba, se utiliza la variable indicadora `unemployed`, definida previamente, que toma el valor de uno para las personas desocupadas y de cero para el resto. Dado que la tasa de desempleo se calcula exclusivamente sobre la población económicamente activa, el análisis debe restringirse a las personas ocupadas y desocupadas, excluyendo tanto a la población inactiva como a las observaciones con valores perdidos en la variable `Employment`, que incluyen a los menores de edad. En consecuencia, la media de `unemployed` dentro de cada grupo de sexo corresponde a la proporción estimada de personas desocupadas, lo que permite contrastar formalmente ambas tasas.

```
lf_design <- survey_design %>%
  filter(
    !is.na(Employment),
    Employment %in% c("Employed", "Unemployed")
  )

lf_design %>%
  group_by(Sex) %>%
  summarise(
    proportion = survey_mean(
      unemployed,
      proportion = TRUE,
      vartype = c("se", "ci"),
      level = 0.95,
      na.rm = TRUE
    )
  )
```

Tabla 4.14.: Estimación de la tasa de desempleo por sexo

Sex	proportion	proportion_se	proportion_low	proportion_upp
Female	0.048	0.012	0.029	0.078
Male	0.084	0.015	0.059	0.118

Las tasas de desempleo estimadas por sexo se presentan en la tabla 4.14. En `R`, la función `svyttest()` del paquete `survey` permite contrastar la diferencia entre las medias de esta variable binaria y, por tanto, la diferencia entre las proporciones de desocupación de hombres y mujeres, incorporando los ajustes asociados al diseño muestral.

```
lf_design <- survey_design %>%
  filter(
    !is.na(Employment),
    Employment %in% c("Employed", "Unemployed")
  )

ttest_unemployment_sex <- svyttest(
  unemployed ~ Sex,
  design = lf_design,
  level = 0.95
)
```

Tabla 4.15.: Prueba de diferencia de proporciones de desocupación por sexo (población económicamente activa)

statistic	p_value	gl	ci_lower	ci_upper	estimated_difference
2.18	0.031	118	0.003	0.068	0.036

A partir de la tabla 4.15, se concluye que la diferencia estimada entre la proporción de hombres desocupados y la proporción de mujeres desocupadas es de 0.0359, es decir, la proporción correspondiente a los hombres es 3.59 puntos porcentuales mayor. El estadístico obtenido es $t = 2.182$, con 118 grados de libertad y un valor-p de 0.031. Dado que el valor-p es inferior al nivel de significancia del 5 %, se rechaza la hipótesis nula de igualdad de proporciones. Por tanto, existe evidencia estadística suficiente para afirmar que, dentro de la población económicamente activa, la proporción de desocupados es diferente entre hombres y mujeres.

4.5. Tablas cruzadas

De acuerdo con Heeringa, West, y Berglund (2017) y Agresti (2019), además de analizar cada variable de manera individual, resulta especialmente relevante estudiar si existe asociación entre dos variables categóricas. Este tipo de análisis permite identificar patrones y relaciones que aportan información valiosa para la toma de decisiones y la comprensión de fenómenos sociales. Por ejemplo,

en el ámbito de las políticas públicas es posible relacionar educación y empleo para diseñar estrategias orientadas al mercado laboral; en la evaluación de programas sociales, analizar diferencias en el acceso a servicios de salud según el nivel de ingresos; y en investigación social, estudiar cómo variables demográficas y condiciones de vida se relacionan entre sí para comprender dinámicas y tendencias de la población.

Formalmente, sean x y y dos variables categóricas con R categorías para las filas y C categorías para las columnas, respectivamente. En el contexto de encuestas por muestreo, el interés suele centrarse en estimar la distribución conjunta de ambas variables, es decir, la proporción de unidades poblacionales que pertenecen simultáneamente a cada combinación posible de categorías. Para ello, las entradas de una tabla cruzada, también conocida como tabla de contingencia, se estiman mediante frecuencias ponderadas obtenidas a partir de los factores de expansión y del diseño muestral. La estimación del total poblacional asociado a cada celda (r, c) se define como:

$$\hat{N}_{rc} = \sum_h \sum_i \sum_k w_{hik}\ I(x_{hik} = r,\ y_{hik} = c),$$

donde w_{hik} representa el factor de expansión asociado al elemento k de la UPM i perteneciente al estrato h, mientras que $I(x_{hik} = r,\ y_{hik} = c)$ corresponde a una función indicadora que toma el valor de uno cuando la unidad observada pertenece simultáneamente a la categoría r de la variable x y a la categoría c de la variable y, y toma el valor de cero en cualquier otro caso.

Además de las frecuencias conjuntas, también es posible estimar los tamaños marginales por filas y columnas, definidos respectivamente como $\hat{N}_{r+} = \sum_{c=1}^{C} \hat{N}_{rc}$ y $\hat{N}_{+c} = \sum_{r=1}^{R} \hat{N}_{rc}$. Mientras que el total general se expresa como $\hat{N}_{++} = \sum_{r=1}^{R} \sum_{c=1}^{C} \hat{N}_{rc}$. La estructura general de una tabla de contingencia con R filas y C columnas se presenta en la tabla 4.16:

Tabla 4.16.: Estructura general de una tabla de contingencia con R filas y C columnas

Variable 2	Variable 1		⋯		Marginal fila
	1	2	⋯	C	
1	$\hat{N}_{11}$	$\hat{N}_{12}$	⋯	$\hat{N}_{1C}$	$\hat{N}_{1+}$
2	$\hat{N}_{21}$	$\hat{N}_{22}$	⋯	$\hat{N}_{2C}$	$\hat{N}_{2+}$
⋮	⋮	⋮	⋱	⋮	⋮
R	$\hat{N}_{R1}$	$\hat{N}_{R2}$	⋯	$\hat{N}_{RC}$	$\hat{N}_{R+}$
Marginal columna	$\hat{N}_{+1}$	$\hat{N}_{+2}$	⋯	$\hat{N}_{+C}$	$\hat{N}_{++}$

De manera tradicional, las tablas de contingencia suelen representarse como arreglos bidimensionales de dimensión $R \times C$. Sin embargo, también pueden extenderse para incorporar una o más variables adicionales, generando subconjuntos o subtablas que permiten estudiar relaciones más complejas entre variables categóricas. A partir de los tamaños estimados, también pueden estimarse proporciones para cada celda de la tabla de contingencia, de la siguiente manera:

$$\hat{p}_{rc} = \frac{\hat{N}_{rc}}{\hat{N}_{++}}$$

A continuación, y siguiendo con la base de datos de ejemplo, se estima la distribución porcentual de hombres y mujeres según su condición de pobreza, incorporando además sus respectivos errores estándar e intervalos de confianza. Los resultados se presentan en la tabla 4.17. Este tipo de análisis permite describir la composición de la población pobre según sexo y evaluar la precisión de las estimaciones obtenidas a partir del diseño muestral.

```
poverty_design <- survey_design %>%
  mutate(poor = factor(
    poor,
    levels = 0:1,
    labels = c("Not poor", "Poor")
  ))
```

```
poverty_design %>%
  group_by(poor, Sex) %>%
  summarise(proportion = survey_prop(vartype = c("se", "ci")))
```

Tabla 4.17.: Proporción de hombres y mujeres en condición de pobreza y no pobreza

poor	Sex	proportion	proportion_se	proportion_low	proportion_upp
Not poor	Female	0.529	0.012	0.505	0.554
Not poor	Male	0.471	0.012	0.446	0.495
Poor	Female	0.524	0.016	0.492	0.555
Poor	Male	0.476	0.016	0.445	0.508

Los resultados indican que el 52.3 % de la población en condición de pobreza corresponde a mujeres, mientras que el 47.6 % corresponde a hombres. Para las mujeres, el intervalo de confianza al 95 % se encuentra entre 49.2 % y 55.5 %, mientras que para los hombres el intervalo correspondiente se ubica entre 44.5 % y 50.7 %. Estas estimaciones permiten observar la distribución relativa de la pobreza según sexo, considerando además la incertidumbre inherente al proceso de muestreo.

Con el mismo enfoque de `srvyr` también puede estimarse la tabla de contingencia mediante `group_by()` y `summarise()`. En el siguiente ejemplo se estima la distribución de la condición de pobreza según sexo, obteniendo además medidas de precisión como errores estándar e intervalos de confianza. El código correspondiente y sus resultados se presentan en la tabla 4.18.

```
sex_by_poverty_proportion <- poverty_design %>%
  group_by(poor, Sex) %>%
  summarise(proportion = survey_prop(vartype = c("se", "ci")))
```

Tabla 4.18.: Proporción de hombres y mujeres por condición de pobreza

poor	Sex	proportion	proportion_se	proportion_low	proportion_upp
Not poor	Female	0.529	0.012	0.505	0.554
Not poor	Male	0.471	0.012	0.446	0.495

Tabla 4.18.: Proporción de hombres y mujeres por condición de pobreza

poor	Sex	proportion	proportion_se	proportion_low	proportion_upp
Poor	Female	0.524	0.016	0.492	0.555
Poor	Male	0.476	0.016	0.445	0.508

Los intervalos de confianza ya fueron calculados por `survey_prop(vartype = c("se", "ci"))`. El siguiente código no vuelve a estimarlos mediante `confint()`; utiliza `select()` para extraer las columnas que contienen sus límites inferior y superior. Los resultados se presentan en la tabla 4.19 y coinciden, por construcción, con los generados anteriormente.

```
sex_by_poverty_proportion %>%
  dplyr::select(poor, Sex, proportion_low, proportion_upp)
```

Tabla 4.19.: Intervalos de confianza para la proporción por sexo y pobreza

poor	Sex	proportion_low	proportion_upp
Not poor	Female	0.505	0.554
Not poor	Male	0.446	0.495
Poor	Female	0.492	0.555
Poor	Male	0.445	0.508

Otro análisis de interés relacionado con tablas de doble entrada en encuestas de hogares consiste en estimar el porcentaje de personas desempleadas según sexo. El procedimiento correspondiente y sus resultados se presentan en la tabla 4.20.

```
sex_by_employment_proportion <- survey_design %>%
  filter(!is.na(Employment)) %>%
  group_by(Employment, Sex) %>%
  summarise(proportion = survey_prop(vartype = c("se", "ci")))
```

Tabla 4.20.: Proporción de hombres y mujeres por estado de ocupación

Employment	Sex	proportion	proportion_se	proportion_low	proportion_upp
Unemployed	Female	0.273	0.054	0.180	0.390
Unemployed	Male	0.727	0.054	0.610	0.820
Inactive	Female	0.770	0.023	0.721	0.813
Inactive	Male	0.230	0.023	0.187	0.279
Employed	Female	0.405	0.019	0.369	0.442
Employed	Male	0.595	0.019	0.558	0.631

De la anterior salida se puede observar que el 27.2 % de las personas desempleadas son mujeres y el 72.7 % son hombres, con errores estándar de 5.3 puntos porcentuales para ambas estimaciones. Los intervalos de confianza correspondientes se presentan en la tabla 4.21:

```
sex_by_employment_proportion %>%
  dplyr::select(Employment, Sex, proportion_low, proportion_upp)
```

Tabla 4.21.: Intervalos de confianza para la proporción por sexo y ocupación

Employment	Sex	proportion_low	proportion_upp
Unemployed	Female	0.180	0.390
Unemployed	Male	0.610	0.820
Inactive	Female	0.721	0.813
Inactive	Male	0.187	0.279
Employed	Female	0.369	0.442
Employed	Male	0.558	0.631

Si ahora el objetivo es estimar la pobreza por región, se utiliza la variable numérica `poor` dentro de un flujo `srvyr`, como se muestra en la tabla 4.22.

```
region_poverty_proportion <- survey_design %>%
  group_by(Region, poor) %>%
  summarise(proportion = survey_prop(vartype = c("se", "ci")))
```

Tabla 4.22.: Proporción en condición de pobreza por región

Region	poor	proportion	proportion_se	proportion_low	proportion_upp
Norte	0	0.641	0.055	0.526	0.742
Norte	1	0.359	0.055	0.258	0.474
Sur	0	0.656	0.043	0.566	0.737
Sur	1	0.344	0.043	0.263	0.434
Centro	0	0.635	0.079	0.470	0.773
Centro	1	0.365	0.079	0.227	0.530
Occidente	0	0.599	0.047	0.504	0.687
Occidente	1	0.401	0.047	0.313	0.496
Oriente	0	0.548	0.088	0.374	0.711
Oriente	1	0.452	0.088	0.289	0.626

De lo anterior se puede concluir que, en la región Norte, el 35 % de las personas están en estado de pobreza, mientras que en el Sur la proporción es del 34 %. La proporción estimada más alta se observa en la región Oriente, con un 45 %. Los errores estándar reportados permiten evaluar la precisión de estas estimaciones.

4.6. La razón de *odds*

De acuerdo con Díaz Monroy, Morales Rivera, y León Dávila (2018) y Agresti (2019), la razón de *odds* es una medida ampliamente utilizada para estudiar la asociación entre dos variables categóricas,

especialmente cuando el interés se centra en comparar la probabilidad de ocurrencia de un evento entre distintos grupos poblacionales. Su interpretación se basa en comparar las ventajas relativas (*odds*) de ocurrencia de un evento entre dos categorías de análisis.

Como explican Heeringa, West, y Berglund (2017), en este sentido, este parámetro permite evaluar cómo cambia dicha ventaja entre diferentes grupos de interés y constituye una herramienta fundamental en estudios sociales, económicos y de salud; además, esta medida también puede utilizarse para cuantificar la relación entre los niveles de una variable y un factor categórico, comparando las ventajas relativas de ocurrencia del evento en cada grupo.

Por ejemplo, suponga que se desea estudiar la asociación entre el sexo de las personas y la condición de pobreza. En particular, interesa evaluar si la ventaja relativa de pertenecer al grupo de personas no pobres frente al grupo de personas pobres es diferente entre mujeres y hombres. Para ello, puede expresarse la siguiente razón de *odds*:

$$OR = \frac{P(\text{pobreza} = 0 \mid \text{Sex} = \text{Female}) \big/ P(\text{pobreza} = 1 \mid \text{Sex} = \text{Female})}{P(\text{pobreza} = 0 \mid \text{Sex} = \text{Male}) \big/ P(\text{pobreza} = 1 \mid \text{Sex} = \text{Male})}$$

La expresión del numerador representa, entre las mujeres, la ventaja relativa de encontrarse en condición de no pobreza respecto de encontrarse en pobreza. De manera análoga, el denominador representa esa misma ventaja entre los hombres. Por tanto, la razón de odds compara ambas ventajas relativas y permite cuantificar si la asociación entre sexo y pobreza favorece más a un grupo que a otro.

Un valor igual a uno indicaría ausencia de asociación entre las variables, es decir, que la relación entre sexo y condición de pobreza es similar para hombres y mujeres. Valores superiores a uno indicarían una mayor ventaja relativa para las mujeres en comparación con los hombres, mientras que valores inferiores a uno sugerirían una mayor ventaja relativa para los hombres. En el contexto de encuestas de hogares, este tipo de medidas resulta especialmente útil para estudiar desigualdades sociodemográficas y brechas en condiciones de bienestar entre distintos grupos poblacionales.

El procedimiento para realizarlo en `R` sería primero estimar las proporciones de la tabla cruzada entre las variables sexo y pobreza, presentada en la tabla 4.23:

```
sex_poverty_interaction_proportion <- survey_design %>%
  group_by(Sex, poor) %>%
  summarise(proportion = survey_prop(vartype = c("se", "ci")))

sex_poverty_odds_contrast <- svymean(x = ~interaction(Sex, poor),
                                     design = survey_design,
                                     se = TRUE, na.rm = TRUE, ci = TRUE,
                                     keep.vars = TRUE)
```

Tabla 4.23.: Proporciones conjuntas de sexo y condición de pobreza

Sex	poor	proportion	proportion_se	proportion_low	proportion_upp
Female	0	0.611	0.032	0.547	0.671
Female	1	0.389	0.032	0.329	0.453

Tabla 4.23.: Proporciones conjuntas de sexo y condición de pobreza

Sex	poor	proportion	proportion_se	proportion_low	proportion_upp
Male	0	0.605	0.037	0.531	0.675
Male	1	0.395	0.037	0.325	0.469

Luego, la función `svycontrast()` realiza el contraste dividiendo cada uno de los elementos de la expresión mostrada anteriormente:

```
odds_ratio <- quote(
 (`interaction(Sex, poor)Female.0` /
  `interaction(Sex, poor)Female.1`) /
 (`interaction(Sex, poor)Male.0` /
  `interaction(Sex, poor)Male.1`)
)

svycontrast(
  stat = sex_poverty_odds_contrast,
  contrasts = odds_ratio
)
```

```
         nlcon  SE
contrast  1.02 0.1
```

Obteniendo que el *odds* estimado de que una mujer no se encuentre en situación de pobreza, en comparación con un hombre, es igual a 1.02. Esto significa que, sin considerar otras variables de la encuesta, las mujeres presentan una razón de probabilidades de no estar en pobreza aproximadamente 2 % mayor que la observada en los hombres. En otras palabras, la probabilidad relativa de no encontrarse en condición de pobreza es ligeramente superior para las mujeres respecto a los hombres.

4.7. Prueba de independencia χ^2

Como explica Agresti (2019), y como se introdujo en los capítulos anteriores, una prueba de hipótesis es un procedimiento estadístico utilizado para evaluar si la evidencia observada en una muestra es compatible con una afirmación acerca de la población. En el contexto de tablas de contingencia, las pruebas de independencia permiten analizar si existe asociación entre dos variables categóricas.

De acuerdo con Heeringa, West, y Berglund (2017), en este caso, la hipótesis nula (H_0) establece que ambas variables son independientes; es decir, que la distribución de una variable no depende de las categorías de la otra. Bajo este supuesto, las diferencias observadas en la muestra se atribuyen únicamente a la variabilidad muestral y no a una relación real entre las variables en la población. Matemáticamente, esta hipótesis puede expresarse como:

$$H_0 : p^0_{rc} = p_{r+} \times p_{+c}, \quad \text{para todo } r = 1, \dots, R \text{ y } c = 1, \dots, C$$

En donde p^0_{rc} representa la proporción esperada en la celda (r, c) bajo el supuesto de independencia, mientras que P_{r+} y P_{+c} corresponden a las proporciones marginales de las filas y columnas, respectivamente. Bajo esta formulación, la hipótesis de independencia implica que la probabilidad conjunta de cada celda puede expresarse como el producto de sus probabilidades marginales.

En consecuencia, la prueba de independencia consiste en comparar las proporciones observadas o estimadas $\hat{p}_{rc}$ con las proporciones esperadas p^0_{rc} bajo la hipótesis nula. Cuando las diferencias entre ambas son pequeñas, la evidencia empírica resulta consistente con el supuesto de independencia. Por el contrario, discrepancias suficientemente grandes sugieren la existencia de asociación entre las variables y conducen al rechazo de H_0.

Como explican J. N. K. Rao y Scott (1981), en las encuestas de hogares, características propias del diseño muestral complejo, como la estratificación, la selección por conglomerados y el uso de factores de expansión, afectan la variabilidad de los estimadores en comparación con la que se obtendría bajo un muestreo aleatorio simple. Como consecuencia, la prueba clásica de independencia χ^2 de Pearson no resulta adecuada para analizar tablas de contingencia provenientes de este tipo de diseños, ya que puede subestimar o sobreestimar la variabilidad real de las estimaciones.

Con el propósito de corregir este problema, Fay (1979), Fellegi (1980) y Thomas y Rao (1987) propusieron los primeros ajustes al estadístico chi-cuadrado de Pearson, basados en el efecto de diseño generalizado ($GDEFF$). Posteriormente, Jon N. K. Rao y Scott (1984) amplió y formalizó el marco teórico de estas correcciones, dando origen a lo que actualmente se conoce como la prueba de Rao-Scott, considerada el procedimiento de referencia para el análisis de datos categóricos obtenidos mediante encuestas complejas.

La idea central del ajuste de primer orden consiste en adaptar la prueba clásica de independencia mediante la incorporación del efecto de diseño generalizado, obteniendo así un estadístico robusto frente a las complejidades del diseño muestral:

$$\chi^2_{RS} = \frac{n_{++}}{GDEFF} \sum_r \sum_c \frac{(\hat{p}_{rc} - p^0_{rc})^2}{p^0_{rc}}$$

Como explica Lumley et al. (2026), esta expresión resume la corrección de primer orden. La opción `statistic = "F"` utilizada más adelante en `svychisq()` corresponde a la corrección de segundo orden de Rao-Scott expresada mediante una aproximación basada en la distribución F; por ello, no debe interpretarse como una aplicación numéricamente idéntica de la fórmula anterior.

donde n_{++} representa el tamaño total de la muestra, $\hat{p}_{rc}$ corresponde a la proporción estimada en la celda (r, c) de la tabla de contingencia y p^0_{rc} denota la proporción esperada bajo la hipótesis nula de independencia, calculada a partir del producto de las proporciones marginales de filas y columnas.

El término $GDEFF$ representa el efecto de diseño generalizado y mide cuánto se incrementa o disminuye la variabilidad de las estimaciones como consecuencia del diseño complejo de la encuesta respecto a la variabilidad que se observaría bajo un muestreo aleatorio simple. De esta manera, el ajuste de Rao–Scott permite que las inferencias derivadas de la prueba de independencia reflejen adecuadamente la estructura del diseño muestral.

Bajo la hipótesis nula H_0, el estadístico χ^2_{RS} sigue aproximadamente una distribución χ^2 con $(R-1)(C-1)$ grados de libertad, donde R y C corresponden al número de categorías de las filas y columnas, respectivamente. En situaciones de muestras pequeñas o con pocos grados de libertad, es frecuente utilizar ajustes basados en la distribución F, los cuales mejoran la precisión de la inferencia

estadística. Las pruebas de Rao-Scott constituyen el estándar para el análisis de asociación entre variables categóricas en encuestas complejas.

El paquete `survey` en `R` implementa este tipo de pruebas mediante la función `svychisq()`, la cual incorpora automáticamente los ajustes de Rao-Scott para tener en cuenta las características del diseño muestral complejo. Esta función permite evaluar la existencia de asociación entre dos variables categóricas utilizando tablas de contingencia ponderadas y corrigiendo adecuadamente la variabilidad de los estimadores. A modo de ejemplo, para evaluar si la condición de pobreza es independiente del sexo, se ejecuta el siguiente código:

```
svychisq(formula = ~Sex + poor, design = survey_design, statistic = "F")
```

```
    Pearson's X^2: Rao & Scott adjustment

data:  NextMethod()
F = 0.06, ndf = 1, ddf = 119, p-value = 0.8
```

En este caso, el argumento `formula` especifica las dos variables categóricas cuya independencia se desea evaluar; `design` corresponde al diseño muestral previamente definido; y `statistic = "F"` indica que la prueba se realizará utilizando la versión ajustada basada en la distribución F. Si el valor-p es superior al 5 %, no se rechaza la hipótesis nula H_0, concluyendo que no existe evidencia estadística suficiente para afirmar que pobreza y sexo están asociadas. Por el contrario, si el valor-p es inferior al 5 %, se rechaza H_0, lo que sugiere la existencia de dependencia o asociación entre ambas variables categóricas. De manera similar, se pueden evaluar otras relaciones, como desempleo y sexo o pobreza y región.

```
svychisq(formula = ~Sex + Employment, design = survey_design, statistic = "F")
```

```
    Pearson's X^2: Rao & Scott adjustment

data:  NextMethod()
F = 62, ndf = 2, ddf = 201, p-value <0.0000000000000002
```

En este primer caso, para el contraste de independencia entre las variables `Sex` y `Employment`, el estadístico obtenido fue $F = 62.251$, con un valor-p menor que 2.2×10^{-16}, lo que proporciona evidencia estadísticamente significativa para rechazar la hipótesis nula de independencia entre ambas variables. En consecuencia, los resultados sugieren que existe una asociación significativa entre el sexo y la condición de empleo en la población de estudio.

```
svychisq(formula = ~Region + poor, design = survey_design, statistic = "F")
```

```
    Pearson's X^2: Rao & Scott adjustment

data:  NextMethod()
F = 0.5, ndf = 3, ddf = 358, p-value = 0.7
```

Por otro lado, esta prueba de independencia entre las variables `Region` y `poor`, da como resultado $F = 0.48794$, con un valor-$p = 0.6914$. Dado que este valor-p es considerablemente mayor que niveles de significancia convencionales, no existe evidencia suficiente para rechazar la hipótesis nula de independencia. Por tanto, los resultados sugieren que, bajo el diseño muestral considerado, no se observa una asociación estadísticamente significativa entre la región y la condición de pobreza.

4.8. Visualización de variables categóricas

De acuerdo con Wickham (2016) y Heeringa, West, y Berglund (2017), la visualización de variables categóricas es fundamental en el análisis de encuestas de hogares, ya que permite comunicar de manera clara la distribución de grupos poblacionales y las comparaciones entre ellos. Al igual que en el análisis de variables continuas, los gráficos deben incorporar los pesos muestrales para que la representación visual corresponda a estimaciones válidas a nivel poblacional. Cuando la precisión estadística es relevante, es recomendable acompañar las barras con intervalos de confianza, que comunican tanto el valor central como la incertidumbre asociada al proceso de estimación.

En esta sección se utiliza el paquete `ggplot2`.

```
library(ggplot2)
```

4.8.1. Gráficos de barras

La construcción de gráficos de barras en `R` requiere, como paso previo, obtener las estimaciones que se desean representar visualmente. En el contexto de encuestas complejas, estas estimaciones pueden calcularse mediante las funciones `survey_total()` o `survey_prop()` del paquete `srvyr`, dependiendo de si el interés se centra en totales poblacionales o proporciones.

A continuación, se ilustra este procedimiento utilizando el tamaño poblacional estimado por zona de residencia. Los resultados gráficos correspondientes se presentan en la figura 4.1:

```
zone_size <- survey_design %>%
  group_by(Zone) %>%
  summarise(size = survey_total(vartype = c("se", "ci")))

zone_colors <- c(Urban = "#48C9B0", Rural = "#117864")

ggplot(
  data = zone_size,
  aes(
    x = Zone, y = size,
    ymax = size_upp, ymin = size_low,
    fill = Zone
  )
) +
  geom_bar(stat = "identity", position = "dodge") +
  geom_errorbar(position = position_dodge(width = 0.9), width = 0.3) +
```

```
  scale_fill_manual(values = zone_colors) +
  theme_bw() +
  theme(legend.position = "top")
```

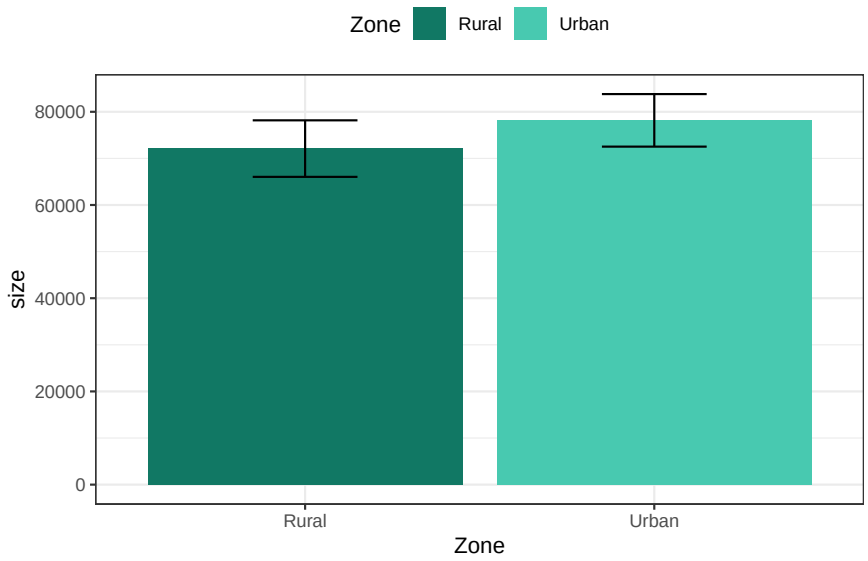

Figura 4.1.: Gráfico de barras del tamaño poblacional por zona geográfica

El procedimiento anterior también puede aplicarse a variables con más de dos categorías. En estos casos, el gráfico de barras permite comparar de manera sencilla las estimaciones asociadas a cada categoría de la variable de interés. Por ejemplo, a continuación se presenta la distribución estimada de la población según la condición de pobreza, cuyos resultados gráficos se muestran en la figura 4.2.

```
poverty_size <- survey_design %>%
  group_by(Poverty) %>%
  summarise(size = survey_total(vartype = c("se", "ci")))

ggplot(
  data = poverty_size,
  aes(
    x = Poverty, y = size,
    ymax = size_upp, ymin = size_low,
    fill = Poverty
  )
) +
  geom_bar(stat = "identity", position = "dodge") +
  geom_errorbar(position = position_dodge(width = 0.9), width = 0.3) +
  theme_bw() +
  theme(legend.position = "top")
```

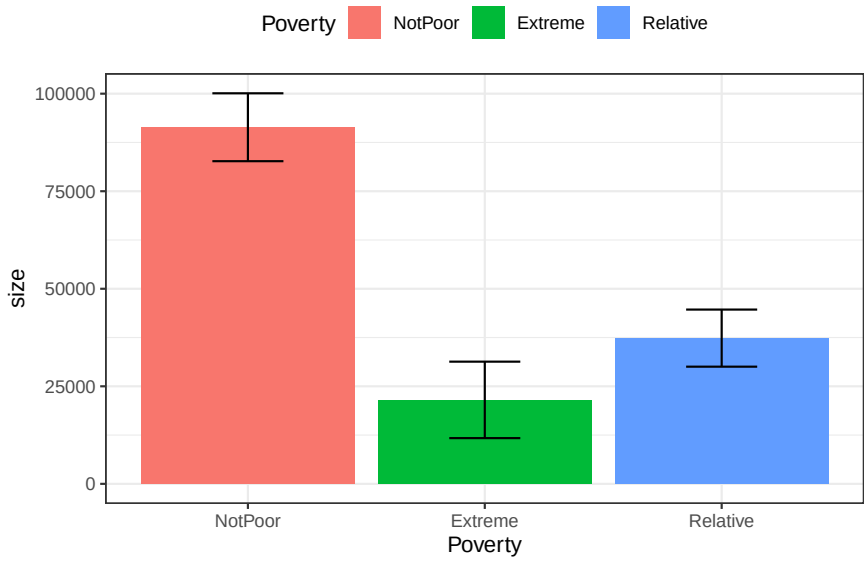

Figura 4.2.: Gráfico de barras del tamaño poblacional por condición de pobreza

Una de las ventajas de los gráficos de barras es que pueden extenderse fácilmente a comparaciones entre dos variables categóricas. Por ejemplo, se puede analizar el tamaño poblacional por nivel de pobreza cruzado con el estado de ocupación, presentada en la figura 4.3:

```
employment_poverty_size <- survey_design %>%
  group_by(unemployed, Poverty) %>%
  summarise(size = survey_total(vartype = c("se", "ci"))) %>%
  as.data.frame() %>%
  mutate(
    unemployed = case_when(
      unemployed == 0 ~ "Occupied",
      unemployed == 1 ~ "Unemployed",
      is.na(unemployed) ~ "Not in the Labour Force"
    )
  )

ggplot(
  data = employment_poverty_size,
  aes(
    x = Poverty, y = size,
    ymax = size_upp, ymin = size_low,
    fill = as.factor(unemployed)
  )
) +
  geom_bar(stat = "identity", position = "dodge") +
  geom_errorbar(position = position_dodge(width = 0.9), width = 0.3) +
  theme_bw() +
  theme(legend.position = "top")
```

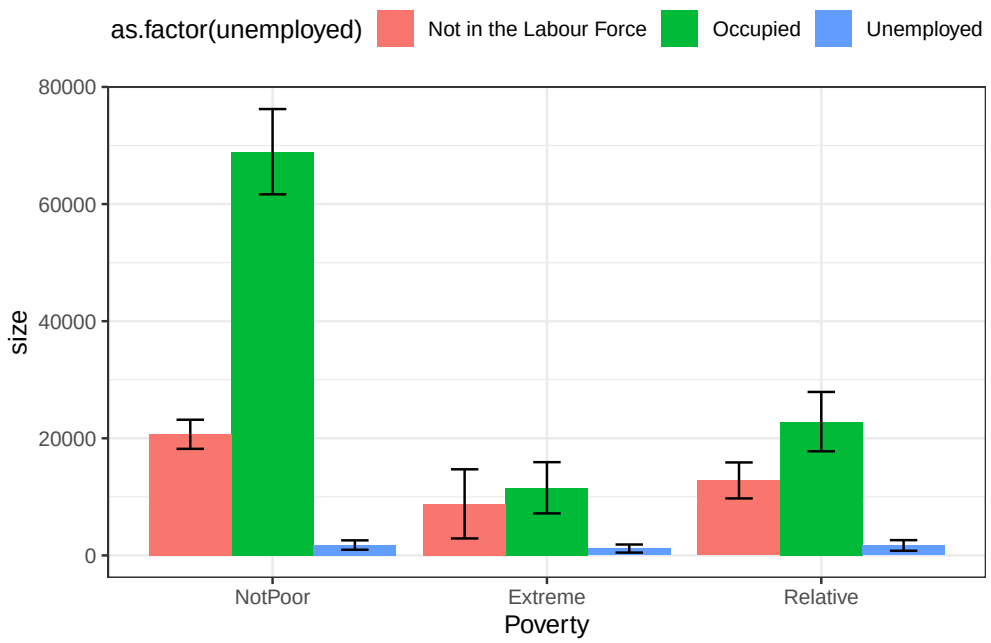

Figura 4.3.: Gráfico de barras del tamaño poblacional por estado de ocupación y pobreza

Este tipo de análisis gráfico resulta especialmente útil para identificar intersecciones de vulnerabilidad, al mostrar de manera simultánea dos o más características poblacionales. También es posible presentar proporciones en lugar de conteos. Por ejemplo, la proporción de hombres en condición de pobreza por zona, presentada en la figura 4.4:

```
male_zone_poverty_proportion <- male_subset %>%
  group_by(Zone, Poverty) %>%
  summarise(proportion = survey_prop(vartype = c("se", "ci")))

ggplot(
  data = male_zone_poverty_proportion,
  aes(
    x = Poverty, y = proportion,
    ymax = proportion_upp, ymin = proportion_low,
    fill = Zone
  )
) +
  geom_bar(stat = "identity", position = "dodge") +
  geom_errorbar(position = position_dodge(width = 0.9), width = 0.3) +
  scale_fill_manual(values = zone_colors) +
  theme_bw() +
  theme(legend.position = "top")
```

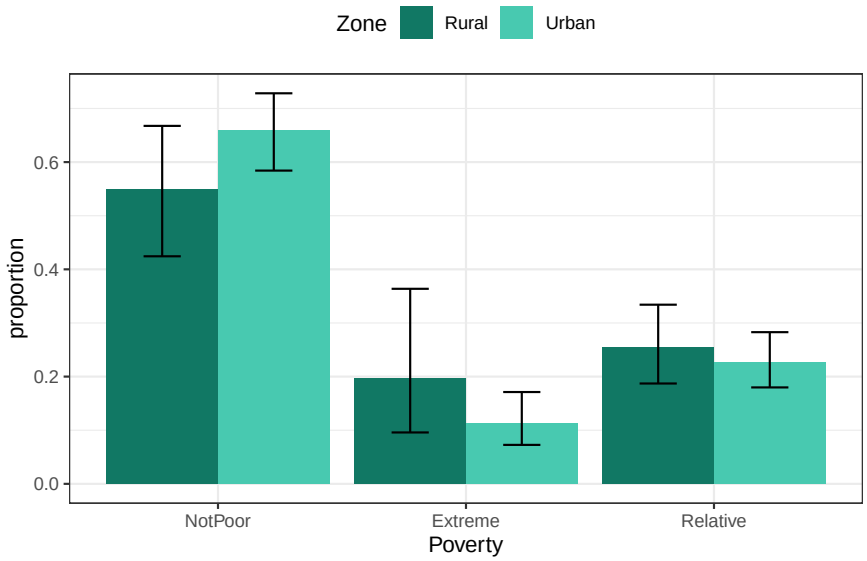

Figura 4.4.: Proporción de hombres en condición de pobreza por zona geográfica

De forma análoga, se puede graficar la proporción de hombres en condición de pobreza desagregada por región, presentada en la figura 4.5:

```
male_region_poverty_proportion <- male_subset %>%
  group_by(Region, poor) %>%
  summarise(proportion = survey_prop(vartype = c("se", "ci"))) %>%
  data.frame()

ggplot(
  data = male_region_poverty_proportion,
  aes(
    x = Region, y = proportion,
    ymax = proportion_upp, ymin = proportion_low,
    fill = as.factor(poor)
  )
) +
  geom_bar(stat = "identity", position = "dodge") +
  geom_errorbar(position = position_dodge(width = 0.9), width = 0.3) +
  theme_bw() +
  theme(legend.position = "top")
```

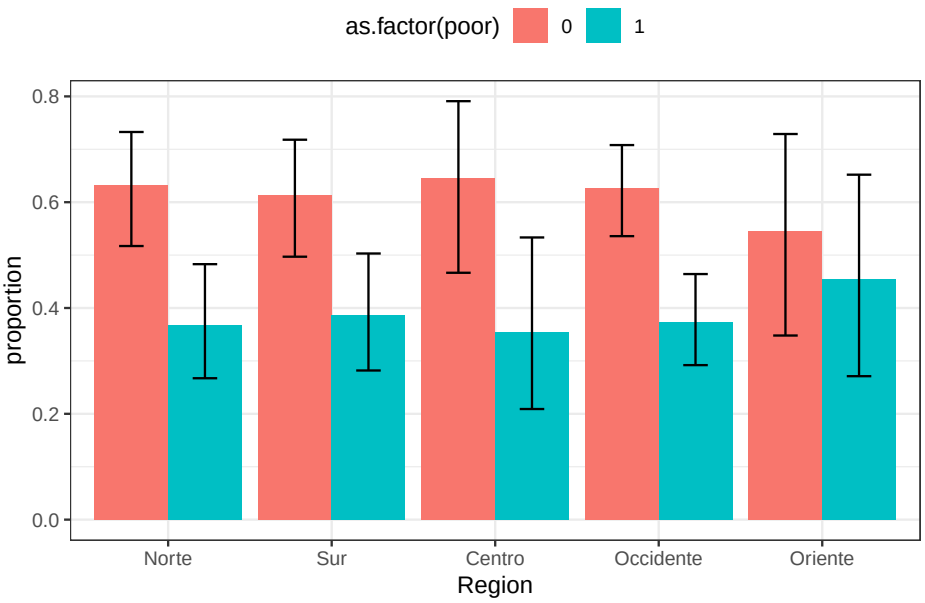

Figura 4.5.: Proporción de hombres en condición de pobreza por región

En síntesis, los gráficos de barras junto con sus correspondientes errores estándar o intervalos de confianza permiten una descripción clara de los patrones de pobreza, empleo u otras variables categóricas en la población y, cuando se aplican correctamente, constituyen una herramienta poderosa para comunicar hallazgos de encuestas de hogares manteniendo el rigor estadístico en la representación de los resultados.

4.8.2. Mapas

Como explica Tennekes (2018), los mapas temáticos constituyen una herramienta de visualización especialmente útil en el análisis de encuestas de hogares, ya que permiten representar la distribución territorial de indicadores sociales y demográficos. En este contexto, las estimaciones de proporciones calculadas en secciones anteriores —como el porcentaje de pobreza o desempleo por región— pueden proyectarse directamente sobre unidades geográficas, facilitando así la identificación de patrones espaciales y desigualdades regionales.

De acuerdo con Pebesma et al. (2026), Tennekes (2026) y Pebesma (2018), para construir este tipo de mapas en `R` se requiere información geoespacial en formato *shapefile*, la cual contiene los polígonos asociados a cada unidad geográfica de análisis. En este capítulo se utiliza el archivo `shapeBigCity/BigCity.shp`, cuyas regiones corresponden a las definidas en la base de datos `BigCity`. Los paquetes `sf` y `tmap` permiten cargar, manipular y visualizar esta información geográfica de manera integrada con los resultados obtenidos a partir de la encuesta.

El siguiente código carga los paquetes necesarios para trabajar con información geoespacial y visualización cartográfica en `R`. La librería `sf` permite leer y manipular objetos espaciales modernos, mientras que `tmap` proporciona herramientas para elaborar mapas temáticos. La instrucción `tmap_mode("plot")` establece el modo de visualización estática de los mapas, apropiado para documentos y reportes. Finalmente, la función `read_sf()` importa el archivo geográfico `BigCity.shp`,

almacenándolo en el objeto `bigcity_shape`, el cual contiene los polígonos correspondientes a las regiones de análisis.

```
library(sf)
library(tmap)
tmap_mode("plot")
bigcity_shape <- read_sf("../Data/shapeBigCity/BigCity.shp")
```

Con el shapefile cargado, se puede generar un mapa de las regiones usando `tm_shape()` y `tm_polygons()`, como se muestra en la figura 4.6:

```
tm_shape(bigcity_shape) +
  tm_polygons(col = "Region")
```

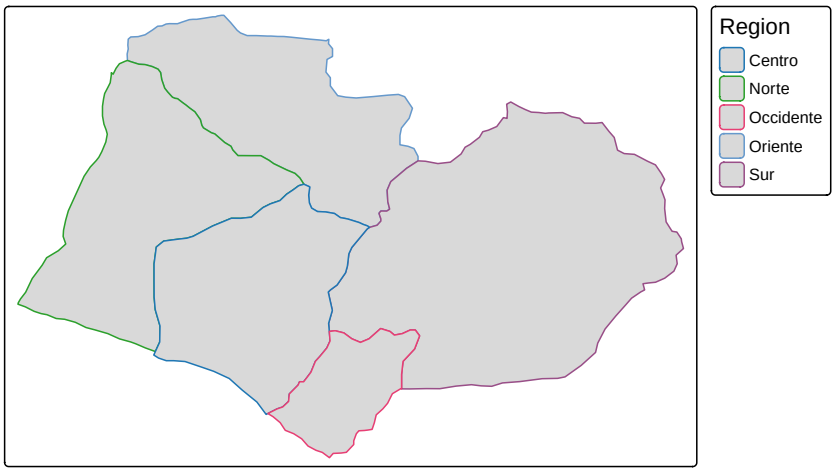

Figura 4.6.: Mapa de regiones geográficas de BigCity

Por ejemplo, si se desea representar geográficamente la proporción de hombres en condición de pobreza por región, estimada previamente en la figura 4.5, es necesario integrar las estimaciones obtenidas con la información espacial del shapefile. Para ello, el objeto `male_region_poverty_proportion`, que contiene las proporciones estimadas para cada región, se vincula con el archivo geográfico mediante la función `left_join()`. Posteriormente, se filtran únicamente las observaciones correspondientes a la categoría de pobreza (`poor == 1`), de manera que el mapa represente exclusivamente la distribución espacial de la población masculina en condición de pobreza.

```
shape_region_map <- tm_shape(
  bigcity_shape %>%
    left_join(
      male_region_poverty_proportion %>%
```

```
        filter(poor == 1),
      by = "Region"
    )
)
```

Una vez construido el objeto espacial con las estimaciones de interés, el argumento `breaks` ayuda a definir los puntos de corte que determinarán los intervalos de la escala de color utilizada en el mapa. Posteriormente, se genera el mapa temático, permitiendo visualizar la distribución regional de la proporción de hombres en condición de pobreza, como se presenta en la figura 4.7.

```
poverty_breaks <- c(0, 0.2, 0.4, 0.6, 0.8, 1)
shape_region_map +
  tm_polygons(
    col = "proportion",
    breaks = poverty_breaks,
    title = "Poverty proportion",
    palette = "YlOrRd"
  )
```

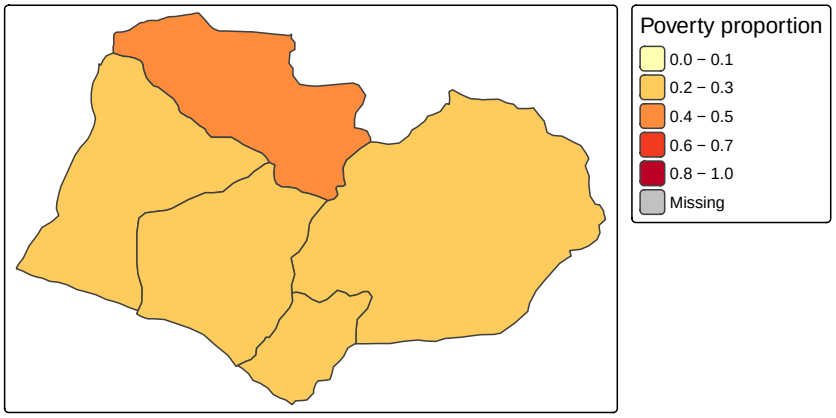

Figura 4.7.: Proporción de hombres en condición de pobreza por región

Otro uso de los mapas en encuestas, señalado por CEPAL (2023), es la visualización de la precisión de las estimaciones. Por ejemplo, usando el coeficiente de variación del ingreso medio por región es posible identificar las áreas donde las estimaciones son menos confiables, como se muestra en la figura 4.8.

El siguiente código calcula y representa geográficamente el coeficiente de variación del ingreso medio por región. En primer lugar, se estima el ingreso promedio para cada región. El argumento `vartype = "cv"` solicita el cálculo del coeficiente de variación de cada estimación. Posteriormente, se definen los intervalos de clasificación de la escala de color mediante el objeto `cv_breaks`, y las estimaciones

se integran con la información geográfica del shapefile usando `left_join()`. Finalmente, mediante las funciones `tm_shape()` y `tm_polygons()` del paquete `tmap`, se construye el mapa que representa espacialmente los coeficientes de variación utilizando una escala de colores en blanco y negro y ajustando la proporción del mapa con `tm_layout(asp = 0)`.

```
region_mean <- survey_design %>%
  group_by(Region) %>%
  summarise(mean = survey_mean(Income,
                               na.rm = TRUE,
                               vartype = c("cv")))

cv_breaks <- c(0, 0.1, 0.2, 1)
shape_cv_map <- tm_shape(
  bigcity_shape %>%
    left_join(region_mean, by = "Region")
)

shape_cv_map +
  tm_polygons(
    "mean_cv",
    breaks = cv_breaks,
    title = "Mean income CV",
    palette = c("#FFFFFF", "#000000")
  ) +
  tm_layout(asp = 0)
```

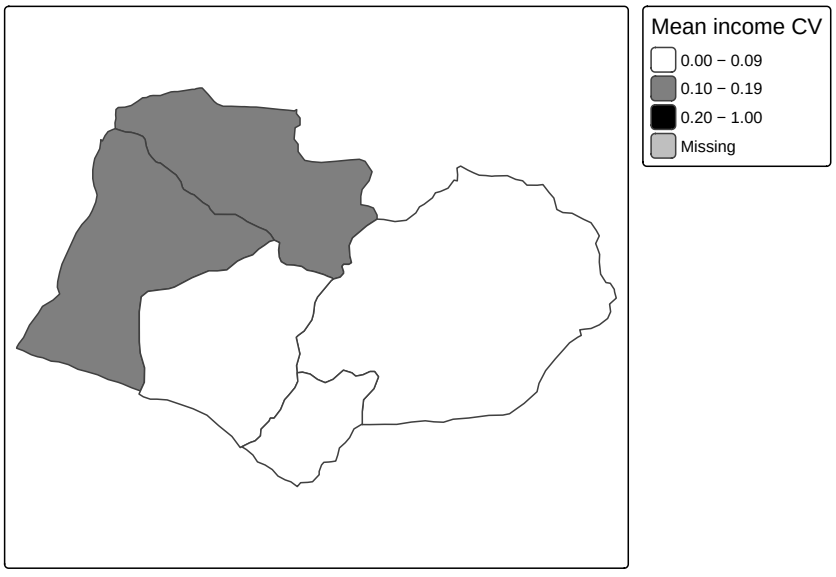

Figura 4.8.: Coeficiente de variación del ingreso medio por región

Los polígonos con tonos más oscuros corresponden a regiones donde la estimación del ingreso presenta mayor variabilidad relativa. Un coeficiente de variación elevado indica menor precisión relativa,

pero no permite atribuirla únicamente a un tamaño muestral insuficiente, pues también puede reflejar la ponderación y otras características del diseño.

De acuerdo con Prener, Grossenbacher, y Zehr (2025) y Wilke (2025), cuando se desea representar simultáneamente dos variables —por ejemplo, la proporción de pobreza y su coeficiente de variación— es posible construir un mapa bivariante utilizando `ggplot2` junto con los paquetes `biscale` y `cowplot`. Este tipo de visualización permite identificar regiones con alta pobreza y alta incertidumbre en la estimación de forma conjunta, como se presenta en la figura 4.9. El siguiente código estima la proporción de mujeres en condición de pobreza en zonas rurales para cada región, incorporando además el cálculo del coeficiente de variación de las estimaciones, generando así una base de datos lista para la construcción de un mapa bivariado.

```
region_sex_poverty_proportion <- survey_design %>%
  group_by(Region, Zone, Sex, poor) %>%
  summarise(proportion = survey_mean(vartype = "cv")) %>%
  filter(poor == 1, Zone == "Rural", Sex == "Female")
```

El siguiente código construye un mapa bivariado que permite visualizar simultáneamente la proporción de pobreza femenina rural y el coeficiente de variación asociado a dicha estimación por región. En primer lugar, se integran las estimaciones contenidas en `region_sex_poverty_proportion` con la información geográfica del shapefile mediante `left_join()`. Posteriormente, la función `bi_class()` del paquete `biscale` clasifica conjuntamente las variables `proportion` y `proportion_cv` en una cuadrícula de (3 ×3) categorías (`dim = 3`), empleando el método de clasificación de Fisher (`style = "fisher"`). Luego, con `ggplot2` y `geom_sf()`, se genera el mapa temático utilizando colores bivariados definidos por `bi_scale_fill()`, mientras que `bi_theme()` aplica un estilo minimalista y se elimina la leyenda predeterminada.

A continuación, la función `bi_legend()` crea una leyenda especializada que permite interpretar simultáneamente ambas dimensiones del mapa: el coeficiente de variación y la pobreza femenina rural. Finalmente, mediante las funciones `ggdraw()` y `draw_plot()` del paquete `cowplot`, se combinan el mapa y la leyenda en una única visualización.

```
library(biscale)
library(cowplot)

bivariate_shape <- bigcity_shape %>%
  left_join(region_sex_poverty_proportion, by = "Region")

k <- 3
bivariate_data <- bi_class(
  bivariate_shape,
  y = proportion,
  x = proportion_cv,
  dim = k,
  style = "fisher"
)

bivariate_map <- ggplot() +
```

```
  geom_sf(
    data = bivariate_data,
    aes(fill = bi_class, geometry = geometry),
    colour = "white",
    size = 0.1
  ) +
  bi_scale_fill(pal = "GrPink", dim = k) +
  bi_theme() +
  theme(legend.position = "none")

bivariate_legend <- bi_legend(
  pal = "GrPink",
  dim = k,
  xlab = "Coefficient of variation",
  ylab = "Rural female poverty",
  size = 8
)

ggdraw() +
  draw_plot(bivariate_map, 0, 0, 1, scale = 0.7) +
  draw_plot(bivariate_legend, 0.75, 0.4, 0.2, 0.2)
```

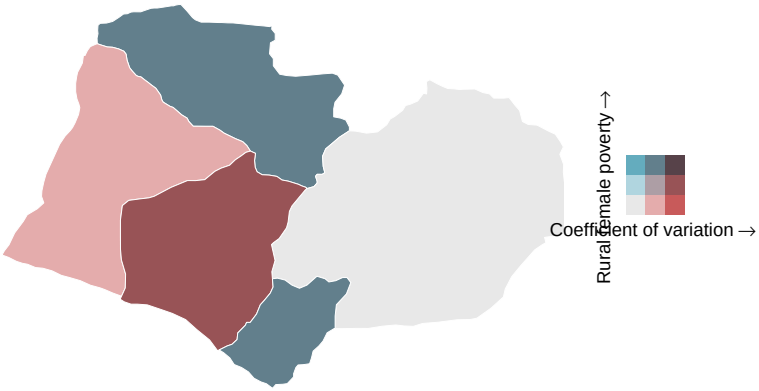

Figura 4.9.: Mapa bivariante: proporción de pobreza femenina rural y su coeficiente de variación por región

En este mapa, cada celda de la leyenda combina un tono de la proporción de pobreza (eje vertical) con un tono del coeficiente de variación (eje horizontal), permitiendo identificar simultáneamente

regiones con alta pobreza y alta incertidumbre en la estimación. Este tipo de representación puede ayudar a priorizar una evaluación del diseño y de la precisión en zonas geográficas específicas; por sí solo, sin embargo, no determina automáticamente que deba ampliarse el tamaño muestral.

5. Modelos de regresión

Los modelos de regresión constituyen una de las herramientas más utilizadas en el análisis de datos provenientes de encuestas, ya que permiten estudiar la relación entre una variable de interés y un conjunto de variables explicativas (Chambers y Skinner 2003; Heeringa, West, y Berglund 2017). A través de estos modelos, es posible evaluar cómo cambian determinadas características de la población según factores demográficos, sociales o económicos observados en la muestra. No obstante, la validez de los resultados obtenidos depende de una adecuada especificación del modelo y de la correcta consideración de las características del diseño muestral de la encuesta.

En términos generales, los modelos de regresión buscan describir y cuantificar la asociación entre una variable respuesta (dependiente) y una o más variables explicativas (independientes), proporcionando elementos para la interpretación estadística y la formulación de inferencias sobre la población de estudio. No obstante la validez de los resultados obtenidos depende, en gran medida, de una adecuada especificación del modelo y de la correcta consideración de las características del diseño muestral de la encuesta.

Por ejemplo, es posible analizar el comportamiento del ingreso de los hogares como variable de respuesta, en función de características como la edad del jefe de hogar, el nivel educativo alcanzado y la situación laboral de sus integrantes, consideradas variables explicativas. Utilizando información proveniente de encuestas de hogares, este tipo de modelos permite identificar patrones de asociación, cuantificar el efecto de distintos factores socioeconómicos sobre el ingreso y generar evidencia empírica útil para el diseño, seguimiento y evaluación de políticas públicas.

Como explica Pfeffermann (2011), dado que las encuestas de hogares están sustentadas en diseños de muestreo complejos, los métodos clásicos de regresión, desarrollados bajo supuestos de muestreo aleatorio simple, pueden resultar inapropiados. Ignorar el diseño puede sesgar los coeficientes cuando la selección es informativa o el modelo está mal especificado y, aun cuando los coeficientes no cambien de manera apreciable, puede producir errores estándar y pruebas de hipótesis incorrectos.

En consecuencia, el análisis de datos provenientes de encuestas exige una atención detallada al diseño de muestreo. La incorporación de los pesos de la encuesta y los ajustes correspondientes a la estratificación y la conglomeración permite obtener inferencias válidas y precisas. Adicionalmente, en algunos casos se han propuesto alternativas simplificadas, como el uso de pesos normalizados o enfoques de ponderación aproximada, que buscan equilibrar la complejidad metodológica con la factibilidad práctica del análisis.

El estudio de la regresión bajo diseños de muestreo complejos cuenta con una trayectoria bien documentada. Como señaló Kish y Frankel (1974) en sus primeras discusiones, estos diseños afectan las inferencias derivadas de modelos de regresión. Posteriormente, Fuller (1975) desarrolló estimadores de varianza apoyados en técnicas de linealización para modelos de regresión lineal múltiple con ponderación desigual y métodos específicos para diseños estratificados y de dos etapas.

Más adelante, Shah, Holt, y Folsom (1977) y David A. Binder (1983) abordaron el problema de las violaciones a los supuestos clásicos al trabajar con datos de encuestas mediante alternativas de

inferencia robusta para los parámetros. En paralelo, las distribuciones muestrales de los estimadores de regresión en poblaciones finitas se incorporaron a procedimientos para estimar varianzas bajo esquemas complejos.

En los años siguientes, Skinner, Holt, y Smith (1989) y Fuller (2002) ampliaron estos aportes mediante estimadores de varianza para los coeficientes de regresión que incorporan la estratificación y la conglomeración, así como mediante compendios de métodos de estimación aplicables a modelos de regresión en encuestas complejas. La discusión reciente incluye, además, métodos de ponderación *q-weighted* y evidencia empírica sobre su utilidad.

5.1. Formulación del modelo

De acuerdo con Chambers y Skinner (2003) y Heeringa, West, y Berglund (2017), los modelos de regresión bajo diseños de muestreo complejos permiten trascender las estadísticas descriptivas y estudiar asociaciones o realizar predicciones; una interpretación causal requiere, además, un diseño de identificación y supuestos causales que no se obtienen únicamente por incorporar el diseño muestral. Su correcta aplicación abre la posibilidad de examinar cómo las características sociodemográficas y económicas se asocian con distintos resultados de interés, aportando evidencia clave para la formulación de políticas públicas.

Como explican Fox y Weisberg (2019), la formulación más básica corresponde al modelo de regresión lineal simple, el cual describe la relación entre una variable respuesta y una única variable explicativa mediante la siguiente expresión:

$$y = \beta_0 + \beta_1 x + \varepsilon$$

En donde y representa la variable de respuesta (dependiente), x la variable explicativa (independiente), β_0 el intercepto del modelo, β_1 el coeficiente asociado a la variable independiente y ε el término de error, que recoge la variabilidad no explicada por el modelo y puede interpretarse como la diferencia entre el valor observado y el valor estimado por el modelo, denotado por $\hat{y}_k$.

En muchas aplicaciones empíricas, y especialmente en el análisis de encuestas de hogares, el fenómeno de interés depende simultáneamente de múltiples factores. En estos casos se emplean modelos de regresión lineal múltiple, los cuales incorporan varias variables explicativas:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon,$$

donde cada coeficiente β_j cuantifica la asociación entre la variable respuesta y la correspondiente covariable x_j, manteniendo constantes las demás variables incluidas en el modelo. De manera más compacta, el modelo puede expresarse utilizando notación matricial como:

$$y_k = \mathbf{x}_k \beta + \varepsilon_k, \quad k = 1, \dots, n,$$

donde $\mathbf{x}_k = [1, x_{1k}, \dots, x_{pk}]$ representa el vector de covariables asociado a la unidad k, mientras que $\beta = [\beta_0, \beta_1, \dots, \beta_p]'$ corresponde al vector de parámetros desconocidos del modelo. En este contexto, el valor esperado de la variable respuesta condicionado al conjunto de covariables puede expresarse como:

$$E(y \mid \mathbf{x}) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

Los modelos de regresión lineal se sustentan en una serie de supuestos teóricos que garantizan la validez de las estimaciones y de las inferencias derivadas del modelo. En primer lugar, se asume que el valor esperado de los residuos condicionado a las covariables es igual a cero, es decir, $E(\varepsilon_k \mid \mathbf{x}_k) = 0$, lo que implica la ausencia de sesgos sistemáticos en la estimación de la variable respuesta. Asimismo, se supone homogeneidad de la varianza de los errores, de modo que la variabilidad residual permanece constante para todos los valores de las covariables, esto es, $Var(\varepsilon_k \mid \mathbf{x}_k) = \sigma^2$. Adicionalmente, se considera que los errores siguen una distribución normal con media cero y varianza constante, expresada como $\varepsilon_k \mid \mathbf{x}_k \sim N(0, \sigma^2)$, y que los residuos asociados a distintas observaciones son independientes entre sí, de manera que $Cov(\varepsilon_k, \varepsilon_j \mid \mathbf{x}_k, \mathbf{x}_j) = 0$.

La validez de los modelos de regresión lineal depende del cumplimiento de algunos supuestos clásicos ampliamente discutidos en la literatura especializada. En conjunto, estos supuestos permiten que los estimadores obtenidos mediante el modelo posean propiedades estadísticas deseables, tales como insesgamiento, eficiencia y consistencia.

5.2. Uso de los pesos de muestreo

Cuando se trabaja con datos provenientes de encuestas basadas en diseños muestrales complejos, los supuestos clásicos de los modelos de regresión rara vez se cumplen de manera estricta, puesto que las observaciones no provienen de muestras aleatorias simples e independientes, sino de esquemas de selección que incorporan elementos como estratificación, conglomeración y probabilidades desiguales de selección. Por ejemplo, la presencia de conglomeración puede violar el supuesto de independencia de los errores, ya que individuos pertenecientes a un mismo hogar, segmento o área geográfica tienden a compartir características similares. Del mismo modo, las probabilidades desiguales de selección y los ajustes de ponderación pueden generar heterogeneidad en la varianza de las observaciones, incumpliendo el supuesto de homocedasticidad.

En este contexto, surge una pregunta fundamental: ¿de qué manera deben incorporarse los pesos muestrales en los modelos de regresión? La respuesta no es sencilla, ya que los pesos de la encuesta reflejan no solo las probabilidades de selección, sino también diversos ajustes asociados a la no respuesta, calibración y correcciones de cobertura. Aunque su utilización permite producir estimaciones representativas de la población, la incorporación directa de ponderaciones en los modelos también puede reducir la eficiencia estadística e incrementar la variabilidad de los estimadores.

En términos generales, Pfeffermann (2011) distingue dos enfoques principales para abordar este problema. El primero corresponde al enfoque basado en el diseño muestral, cuyo objetivo es obtener inferencias válidas para la población objetivo respetando las características del proceso de selección de la muestra. De acuerdo con Heeringa, West, y Berglund (2017), desde esta perspectiva, los pesos muestrales son esenciales para corregir las probabilidades desiguales de inclusión y producir estimaciones insesgadas de los coeficientes de regresión. Sin embargo, este enfoque no protege frente a posibles errores de especificación del modelo; es decir, aun cuando las estimaciones sean válidas desde el punto de vista del diseño, el modelo puede no representar adecuadamente las relaciones existentes en la población.

El segundo corresponde al enfoque orientado al modelo, según el cual los pesos no son necesariamente requeridos siempre que el modelo esté correctamente especificado y el mecanismo de muestreo

sea no informativo. Bajo este supuesto, las relaciones entre las variables observadas en la muestra coinciden con las de la población, de manera que el diseño muestral no introduce sesgos relevantes en la estimación de los parámetros. Desde esta perspectiva, incorporar ponderaciones podría incluso resultar contraproducente, al aumentar innecesariamente la varianza de los estimadores y, en consecuencia, los errores estándar.

Según Skinner, Holt, y Smith (1989) y Naciones Unidas (2026), la discusión sobre la conveniencia de utilizar ponderaciones en modelos de regresión ha sido ampliamente desarrollada en la literatura especializada. En la práctica, una recomendación metodológica frecuente consiste en estimar los modelos tanto con ponderaciones como sin ellas y comparar posteriormente los resultados obtenidos. Si la inclusión de los pesos produce cambios importantes en los coeficientes estimados o modifica las conclusiones sustantivas del análisis, ello sugiere que el diseño de muestreo es informativo o que el modelo presenta problemas de especificación, por lo que el uso de ponderaciones resulta aconsejable. Por el contrario, si los coeficientes permanecen relativamente estables y las ponderaciones únicamente incrementan los errores estándar, puede considerarse que el modelo captura adecuadamente la estructura de los datos y que el uso de pesos no es estrictamente necesario.

En términos aplicados, esta decisión suele depender del propósito analítico del estudio. Cuando el objetivo es realizar inferencia descriptiva y producir estimaciones representativas de la población, la utilización de ponderaciones es indispensable. En cambio, en contextos de inferencia analítica orientados al estudio de asociaciones, relaciones causales o contrastes de hipótesis, pueden emplearse tanto modelos ponderados como no ponderados, especialmente cuando el modelo incorpora variables relacionadas con el diseño muestral, como estratos o conglomerados. No obstante, el uso de modelos no ponderados debe justificarse cuidadosamente, dado que implica asumir condiciones más restrictivas sobre el mecanismo de selección de la muestra y la correcta especificación del modelo estadístico.

En resumen, cuando se opta por ponderar, las ponderaciones corrigen posibles sesgos de sobre o subrepresentación de determinados grupos y contribuyen a obtener estimaciones de varianza más exactas. Dentro del enfoque basado en el diseño, esto permite que los resultados se aproximen a valores insesgados comparables a los que se obtendrían en un censo completo, incluso cuando el modelo no esté formulado de manera óptima. Sin embargo, cuando los pesos presentan gran dispersión, pueden aumentar la varianza de los parámetros estimados y volver inestables las estimaciones, razón por la que en contextos explicativos o analíticos los modelos sin ponderar pueden, en ocasiones, arrojar resultados más consistentes y eficientes.

En cualquier caso, si el modelo se encuentra mal especificado, ignorar las ponderaciones muestrales puede conducir a estimaciones sesgadas, poco informativas o incluso carentes de validez inferencial. Por esta razón, la selección adecuada de las variables incluidas en el modelo constituye un aspecto fundamental del análisis. Con el fin de abordar estas problemáticas, se han propuesto diversas estrategias de ajuste orientadas a lograr un equilibrio entre ambos objetivos. Entre los procedimientos más utilizados se destacan los siguientes:

1. Pesos tipo Senado: este procedimiento ajusta los pesos de manera que su suma coincida con el tamaño de la muestra, en lugar del tamaño de la población. Esta transformación multiplica todos los pesos por una constante y conserva su dispersión relativa; por tanto, no constituye un recorte de pesos extremos. Las nuevas ponderaciones son $w_k^{Senate} = w_k \times \frac{n}{\sum w_k}$.
2. Pesos normalizados: en este enfoque los pesos originales se reescalan para que su suma sea igual a uno. Al ser también una multiplicación por una constante común, el reescalamiento

no reduce por sí mismo la variabilidad relativa de los pesos ni sustituye la especificación del diseño; su efecto sobre las varianzas depende del método y del software utilizados. Las nuevas ponderaciones son $w_k^{Normalized} = \frac{w_k}{\sum w_k}$.

Es importante destacar que, en ambos métodos, los pesos ajustados se obtienen mediante transformaciones multiplicativas directas de los pesos muestrales originales. Por ello, no deben emplearse para calcular tamaños o totales poblacionales. Asimismo, estos procedimientos no alteran las estimaciones de razones, como medias o proporciones, ya que en tales casos los pesos se cancelan en el cociente. En la práctica, el uso de estos ajustes puede considerarse una solución pragmática en contextos donde no se dispone de software especializado para encuestas. No obstante, cuando se cuenta con herramientas como los paquetes `survey` y `srvyr` en `R`, descritos en los capítulos anteriores, el reescalamiento deja de ser necesario, pues dichos paquetes realizan el tratamiento adecuado para preservar tanto la representatividad como las propiedades inferenciales del modelo.

5.3. Estimación de los parámetros del modelo

De acuerdo con Molina y Skinner (1992) y David A. Binder (2011), en términos del ajuste de modelos de regresión con encuestas de hogares, se han desarrollado metodologías inferenciales avanzadas que integran las fuentes de incertidumbre en un mismo marco analítico, buscando reflejar tanto la estructura del diseño como los supuestos y limitaciones del modelo. Entre las aproximaciones más relevantes se encuentran la pseudo-verosimilitud y la inferencia combinada.

El método de pseudo-verosimilitud extiende las técnicas tradicionales de máxima verosimilitud para ajustarlas a las particularidades de los diseños muestrales complejos. En este enfoque, la distribución de muestreo definida por el diseño ocupa un rol central, mientras que la distribución del modelo pasa a un segundo plano. Si bien en contextos de modelos bien especificados los estimadores basados en pseudo-verosimilitud tienden a ser insesgados o consistentes, su principal virtud radica en evitar los sesgos que se originarían al ignorar el diseño muestral. En términos prácticos, este método traduce el modelo tradicional en uno que respeta la forma en que los datos fueron obtenidos, garantizando inferencias más sólidas.

En contraste, la inferencia combinada propone un marco unificado en el que se integran simultáneamente la variabilidad del muestreo y la incertidumbre del modelo. Considerar ambas fuentes permite caracterizar de manera más completa la variabilidad total, pero no garantiza automáticamente estimaciones más precisas ni elimina todo sesgo; esas propiedades dependen del diseño y de la especificación del modelo. De este modo, la inferencia combinada resulta especialmente útil en aplicaciones donde se requiere equilibrar representatividad poblacional y solidez estadística de los modelos ajustados.

Como explica Wolter (2007), cuando se desea estimar los parámetros de un modelo de regresión lineal a partir de una muestra con un diseño complejo, el enfoque estándar cambia. En consecuencia, los estimadores tradicionales, como los obtenidos por máxima verosimilitud o mínimos cuadrados, pueden resultar sesgados y producir errores estándar poco confiables. Frente a esta situación, los métodos de estimación de varianza, como la linealización de Taylor y el bootstrap, son válidos en la medida en que su implementación reproduzca las características pertinentes del diseño.

Según Fuller (2002) y Chambers y Skinner (2003), en el caso de un modelo de regresión lineal simple, la estimación de β_1 bajo un esquema de muestreo complejo se realiza, mediante un estimador ponderado, que puede expresarse de la siguiente forma:

$$\hat{\beta}_1 = \frac{\sum_h \sum_i \sum_k w_{hik}\,(y_{hik} - \hat{\bar{y}})(x_{hik} - \hat{\bar{x}})}{\sum_h \sum_i \sum_k w_{hik}\,(x_{hik} - \hat{\bar{x}})^2} = \frac{\hat{t}_{xy}}{\hat{t}_{xx}}$$

En donde $\hat{\bar{y}} = \hat{t}_y/\hat{N}$ y $\hat{\bar{x}} = \hat{t}_x/\hat{N}$ corresponden a las medias estimadas de las variables de estudio y auxiliar, respectivamente, mientras que w_{hik} representa los pesos de muestreo asociados a cada unidad de observación. Esta formulación incorpora explícitamente dichos pesos, lo que permite considerar las probabilidades desiguales de selección y las características propias del diseño muestral. Asimismo, la varianza estimada de $\hat{\beta}_1$ puede aproximarse como:

$$\widehat{Var}\left(\hat{\beta}_1\right) \approx \frac{\widehat{Var}(\hat{t}_{xy}) + \hat{\beta}_1^2 \widehat{Var}(\hat{t}_{xx}) - 2\hat{\beta}_1 \widehat{Cov}(\hat{t}_{xy}, \hat{t}_{xx})}{(\hat{t}_{xx})^2}$$

Si $\hat{\beta}_1$ corresponde al estimador de la pendiente en un modelo de regresión lineal simple ponderado, entonces el estimador del intercepto $\hat{\beta}_0$ está dado por

$$\hat{\beta}_0 = \hat{\bar{y}} - \hat{\beta}_1 \hat{\bar{x}}$$

Nótese que la varianza de $\hat{\beta}_0$ depende simultáneamente de los estimadores $\hat{\bar{y}}$, $\hat{\bar{x}}$ y $\hat{\beta}_1$. Por esta razón, su aproximación puede obtenerse mediante las propiedades de la varianza de combinaciones lineales de estimadores, en conjunto con técnicas de linearización de Taylor. A partir de este procedimiento se deriva la matriz de varianzas y covarianzas de los coeficientes del modelo, de la cual se obtienen tanto los errores estándar como las covarianzas asociadas a las estimaciones de los parámetros.

De acuerdo con Kish y Frankel (1974), en el caso de la regresión múltiple, si $\hat{\beta}$ representa el estimador de β, entonces la varianza de cada coeficiente se estima considerando su interdependencia con los demás parámetros, lo que se refleja en la construcción de una matriz de varianzas-covarianzas que recoge tanto la variabilidad individual como las covarianzas entre todos los estimadores. Este cálculo requiere utilizar totales ponderados de cuadrados y productos cruzados de todas las combinaciones entre la variable dependiente y y el conjunto de predictores $\mathbf{x} = (1, x_1, \ldots, x_p)$. En términos generales:

$$\widehat{Var}(\hat{\beta}) = \hat{\Sigma} = \begin{bmatrix} \widehat{Var}(\hat{\beta}_0) & \widehat{Cov}(\hat{\beta}_0, \hat{\beta}_1) & \cdots & \widehat{Cov}(\hat{\beta}_0, \hat{\beta}_p) \\ \widehat{Cov}(\hat{\beta}_0, \hat{\beta}_1) & \widehat{Var}(\hat{\beta}_1) & \cdots & \widehat{Cov}(\hat{\beta}_1, \hat{\beta}_p) \\ \vdots & \vdots & \ddots & \vdots \\ \widehat{Cov}(\hat{\beta}_0, \hat{\beta}_p) & \widehat{Cov}(\hat{\beta}_1, \hat{\beta}_p) & \cdots & \widehat{Var}(\hat{\beta}_p) \end{bmatrix}$$

En este capítulo se mostrará cómo implementar estos modelos en `R` mediante las librerías `survey` y `srvyr`, junto con `tidyverse` para organizar el flujo de trabajo. Se abordará la especificación del diseño muestral, la estimación de modelos lineales y logísticos, así como la obtención de errores estándar y pruebas de hipótesis ajustadas al diseño, integrando los fundamentos teóricos con ejemplos prácticos reproducibles. Para ejemplificar los conceptos desarrollados hasta este punto, se empleará la misma base de datos utilizada a lo largo del libro. El proceso inicia con la carga de las librerías, los datos y la definición del diseño de muestreo:

```
library(survey)
library(srvyr)
library(tidyverse)

data(BigCity, package = "TeachingSampling")
survey_data <- readRDS("../Data/encuesta.rds")

survey_design <- survey_data %>%
  as_survey_design(
    strata = Stratum,
    ids = PSU,
    weights = wk,
    nest = TRUE
  )
```

En los ejemplos desarrollados a lo largo de este capítulo se emplearán los mismos subgrupos considerados en los capítulos anteriores.

```
urban_subset <- survey_design %>%
  filter(Zone == "Urban")

rural_subset <- survey_design %>%
  filter(Zone == "Rural")

female_subset <- survey_design %>%
  filter(Sex == "Female")

male_subset <- survey_design %>%
  filter(Sex == "Male")
```

Con el objetivo de ajustar un modelo de regresión entre las variables ingreso y gasto, resulta conveniente explorar previamente la relación existente entre ambas variables. Para ello, se construye un diagrama de dispersión mediante `ggplot2`, el cual permite evaluar visualmente la forma y la intensidad de la asociación. El resultado se presenta en la figura 5.1.

```
unweighted_plot <- ggplot(
  data = survey_data,
  aes(x = Expenditure, y = Income)
) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE)

unweighted_plot
```

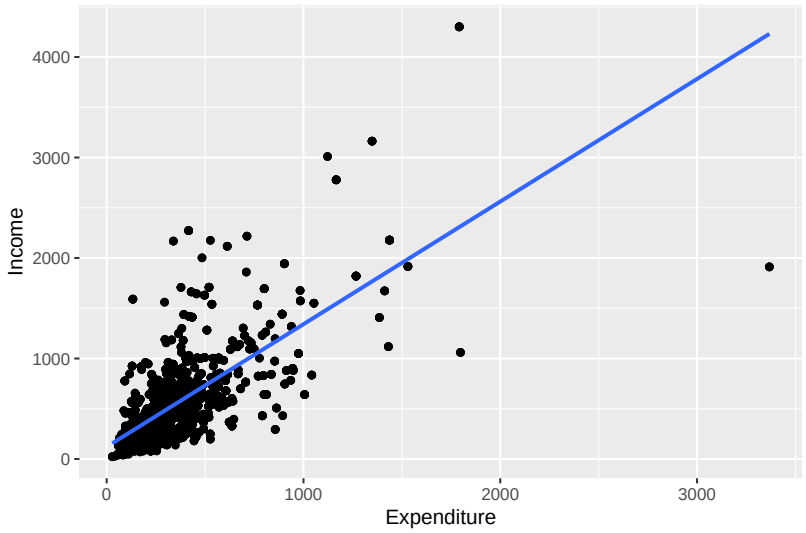

Figura 5.1.: Diagrama de dispersión entre gasto e ingreso en la muestra (sin ponderar)

Los datos muestrales conservan una relación aproximadamente lineal entre ingreso y gasto, aunque se observa una mayor dispersión en los valores correspondientes a hogares con niveles de gasto más elevados. Una vez completada la exploración gráfica de los datos, se procede al ajuste de los modelos de regresión lineal.

Para ajustar modelos incorporando los factores de expansión se usa la función `svyglm` de la librería `survey`, que permite estimar modelos lineales incorporando explícitamente las características del diseño muestral. En este caso, la expresión `Income ~ Expenditure` especifica que la variable `Income` se modela en función de la variable `Expenditure`; es decir, se ajusta una regresión lineal simple en la que el gasto actúa como variable explicativa del ingreso. El argumento `design = survey_design` indica que el modelo debe utilizar el objeto de diseño muestral previamente definido, mientras que `family = stats::gaussian()` especifica que se ajusta un modelo lineal con distribución normal, equivalente a una regresión lineal clásica.

```
fit_svy <- svyglm(
  Income ~ Expenditure + Zone + Sex,
  design = survey_design,
  family = stats::gaussian()
)

fit_svy
```

```
Stratified 1 - level Cluster Sampling design (with replacement)
With (238) clusters.
Called via srvyr
Sampling variables:
  - ids: PSU
  - strata: Stratum
  - weights: wk
```

```
Call:  svyglm(formula = Income ~ Expenditure + Zone + Sex, design = survey_design,
    family = stats::gaussian())

Coefficients:
(Intercept)  Expenditure    ZoneUrban      SexMale
      73.58         1.22        66.65        20.64

Degrees of Freedom: 2604 Total (i.e. Null);  116 Residual
Null Deviance:      635000000
Residual Deviance: 309000000    AIC: 38300
```

La salida anterior corresponde al ajuste de un modelo de regresión lineal en el que la variable `Income` se explica a partir de las variables `Expenditure`, `Zone` y `Sex`. Los resultados indican la existencia de una relación positiva entre ingreso y gasto, de modo que, manteniendo constantes las demás variables, el ingreso esperado aumenta en aproximadamente 1.22 unidades por cada unidad adicional de gasto.

Asimismo, se evidencian diferencias sistemáticas asociadas a las características sociodemográficas incluidas en el modelo: residir en zona urbana se asocia con un incremento promedio de 66.65 unidades en el ingreso respecto de la categoría de referencia, mientras que ser hombre se asocia con un aumento de 20.64 unidades en comparación con la categoría base del sexo. El intercepto estimado, igual a 73.58, representa el nivel esperado de ingreso para la categoría de referencia del modelo (zona rural y sexo femenino) cuando el gasto es igual a cero.

5.4. Diagnóstico del modelo

De acuerdo con Li y Valliant (2015) y Fox y Weisberg (2019), en el análisis de encuestas de hogares, la evaluación de un modelo estadístico constituye un aspecto importante que contribuye a valorar la validez de las inferencias. Un modelo adecuadamente especificado no depende únicamente de la selección de las covariables, sino también del examen de los supuestos que sustentan la consistencia y confiabilidad de los resultados.

En el caso de los modelos de regresión lineal aplicados a encuestas complejas, resulta necesario examinar distintos aspectos relacionados con la calidad del ajuste y el comportamiento de los errores. Como se indicó anteriormente, entre ellos se encuentran la capacidad explicativa del modelo, la normalidad y homogeneidad de la varianza de los residuos, la independencia entre errores y la posible presencia de observaciones influyentes o valores atípicos que puedan afectar las estimaciones.

Estas verificaciones adquieren particular importancia en el contexto de diseños muestrales complejos, donde características como la estratificación, la conglomeración y el uso de factores de expansión pueden intensificar problemas asociados con la heterocedasticidad, la dependencia entre observaciones o la sensibilidad a valores extremos. Por esta razón, el diagnóstico del modelo no debe limitarse a los supuestos clásicos de la regresión, sino incorporar también las particularidades derivadas del diseño de la encuesta. La aplicación sistemática de estos procedimientos permite evaluar la solidez del modelo, fortalecer la confianza en las inferencias y asegurar que los resultados obtenidos sean representativos y útiles tanto para la investigación social aplicada como para el análisis de políticas públicas.

5.4.1. Coeficiente de determinación

Como explican Fox y Weisberg (2019), una de las medidas más utilizadas para evaluar el ajuste de un modelo de regresión es el coeficiente de determinación, denotado por R^2; no debe confundirse con el coeficiente de correlación múltiple R, aunque en el modelo lineal con intercepto se relacionan mediante R^2. Este indicador cuantifica la proporción de la variabilidad de la variable dependiente que es explicada por el modelo. Sus valores oscilan entre cero y uno: cuanto más cercano se encuentre a la unidad, mayor será la capacidad explicativa del modelo; en contraste, valores próximos a cero indican un bajo nivel de explicación de la variabilidad observada. La magnitud del R^2 no debe evaluarse en términos absolutos, sino considerando el fenómeno estudiado y el tipo de información disponible.

El coeficiente se calcula a partir de las sumas de cuadrados totales y de error, como $R^2 = 1 - \frac{SSE}{SST}$; en donde SST representa la suma de cuadrados totales y SSE la suma de cuadrados del error.

Según Heeringa, West, y Berglund (2017), en encuestas con diseños de muestreo complejos, esta medida debe ajustarse para incorporar la estructura del diseño y los pesos muestrales. El estimador ponderado del coeficiente de determinación se define como:

$$\hat{R}^2 = 1 - \frac{\widehat{SSE}}{\widehat{SST}}$$

donde $\widehat{SSE}$ es la suma ponderada de errores al cuadrado, calculada como $\widehat{SSE} = \sum_h \sum_i \sum_k w_{hik}\,(y_{hik} - x_{hik}\hat{\beta})^2$; mientras que $\widehat{SST}$ representa la suma total ponderada de cuadrados, definida por $\widehat{SST} = \sum_h \sum_i \sum_k w_{hik}\,(y_{hik} - \hat{\bar{y}})^2$.

De acuerdo con Pfeffermann (2011), dado que R^2 tiende a incrementarse conforme se incorporan más variables al modelo, conviene complementar su interpretación con el coeficiente de determinación ajustado (R^2_{adj}), el cual introduce una penalización en función del número de covariables incluidas y del tamaño muestral. Este coeficiente se define como $\hat{R}^2_{adj} = 1 - \frac{(n-1)}{(n-p)}(1 - \hat{R}^2)$, donde n corresponde al número de observaciones, p representa el número de parámetros estimados y $\hat{R}^2$ denota el coeficiente de determinación ponderado definido previamente.

De esta manera, el ajuste permite evaluar la capacidad explicativa del modelo controlando el efecto derivado de la incorporación de nuevas variables. Este ajuste facilita una comparación más equitativa entre modelos con distinto número de predictores y resulta particularmente útil en el análisis de encuestas, donde la complejidad del diseño y el uso de ponderaciones pueden influir notablemente en la magnitud del R^2.

En `R`, el coeficiente de determinación R^2 puede estimarse a partir de los modelos ajustados previamente. Para ello, se ajusta inicialmente un modelo nulo (incluyendo únicamente el intercepto), el cual permite calcular la suma total ponderada de cuadrados ($\widehat{SST}$).

```
null_model <- svyglm(Income ~ 1, design = survey_design)

s1 <- summary(fit_svy)
s0 <- summary(null_model)
```

```
wsst <- s0$dispersion[1]
wsse <- s1$dispersion[1]
```

La estimación del coeficiente de determinación y su contraparte ajustada se obtiene a partir del siguiente cálculo:

```
n <- nrow(survey_design)
p <- length(coef(fit_svy))
r2 <- 1 - wsse / wsst
r2_adj <- 1 - ((n - 1) / (n - p)) * (1 - r2)

r2
```

```
[1] 0.5138
```

```
r2_adj
```

```
[1] 0.5133
```

5.4.2. Residuales

De acuerdo con Li y Valliant (2015) y Fox y Weisberg (2019), en el diagnóstico de modelos, el análisis de los residuales es una de las herramientas más relevantes. Si el modelo está correctamente especificado, los residuales actúan como una aproximación de los errores no observados y permiten evaluar, de forma indirecta, si los supuestos del modelo se cumplen razonablemente en los datos. Su revisión sistemática permite identificar posibles desviaciones en la especificación, lo que ayuda a determinar si el ajuste es adecuado o si, por el contrario, es necesario reconsiderar la forma funcional del modelo o el método de estimación utilizado.

Según Heeringa, West, y Berglund (2017), en encuestas con diseños de muestreo complejos, los residuales de Pearson constituyen una herramienta habitual para evaluar las discrepancias entre los valores observados y los valores esperados bajo el modelo ajustado. Estos se definen como:

$$r_k = \frac{y_k - \mu_k}{\sqrt{\widehat{Var}(\mu_k)/w_k}},$$

donde $\hat{\mu}_k = \mathbf{x}_k^\top \hat{\beta}$ representa el valor esperado de y_k bajo el modelo ajustado, w_k corresponde al peso muestral asociado a la unidad k, y $\widehat{Var}(\hat{\mu}_k)$ denota la varianza estimada bajo la familia del modelo.

Los residuales de Pearson también se utilizan para evaluar aspectos clave del ajuste del modelo, en particular la normalidad y la homogeneidad de la varianza de los errores. De forma complementaria, el gráfico de residuos frente a valores ajustados constituye una herramienta diagnóstica especialmente informativa, ya que su inspección visual permite detectar posibles patrones sistemáticos que sugieran incumplimientos de los supuestos de independencia o heterocedasticidad.

La normalidad de los errores puede examinarse como diagnóstico, aunque no es un requisito estricto para la estimación de los coeficientes y su importancia inferencial depende del método de varianza y del tamaño de la muestra. Para ello, una herramienta estándar es el gráfico cuantil-cuantil (QQ-plot), el cual compara los cuantiles de los residuos observados con los cuantiles teóricos de una distribución normal con la misma media y varianza. Una alineación aproximada de los puntos sobre la recta de 45° indica que el supuesto de normalidad resulta razonable dentro del contexto del modelo.

De acuerdo con Valliant (2024), a continuación, estos diagnósticos se aplican a los modelos previamente ajustados. Para ello, se utiliza el paquete `svydiags`, una extensión del paquete `survey` para el diagnóstico de modelos estimados bajo diseños muestrales complejos. Este paquete permite obtener los residuales estandarizados de manera directa, facilitando la evaluación de los supuestos del modelo en este tipo de contextos:

```
library(svydiags)
stdresids <- as.numeric(svystdres(fit_svy)$stdresids)
survey_design$variables <- survey_design$variables %>%
  mutate(stdresids = stdresids)
```

El análisis de normalidad puede complementarse con el histograma de los residuales estandarizados, presentado en la figura 5.2. El histograma sugiere una desviación respecto a la curva normal teórica (en rojo), pero la inspección gráfica por sí sola no constituye una prueba formal ni demuestra que el modelo sea inválido. La distribución de los residuos (representada en azul) muestra un pico mucho más agudo y estrecho que la curva normal, además de una asimetría positiva marcada con una cola extendida hacia la derecha. Esta discrepancia motiva examinar transformaciones de la variable dependiente o familias más flexibles, además de evaluar la especificación de la media.

```
ggplot(
  data = survey_design$variables,
  aes(x = stdresids)
) +
  geom_histogram(
    aes(y = ..density..),
    colour = "black",
    fill = "blue",
    alpha = 0.3
  ) +
  geom_density(size = 1, colour = "blue") +
  geom_function(fun = dnorm, colour = "red", size = 1) +
  labs(y = "")
```

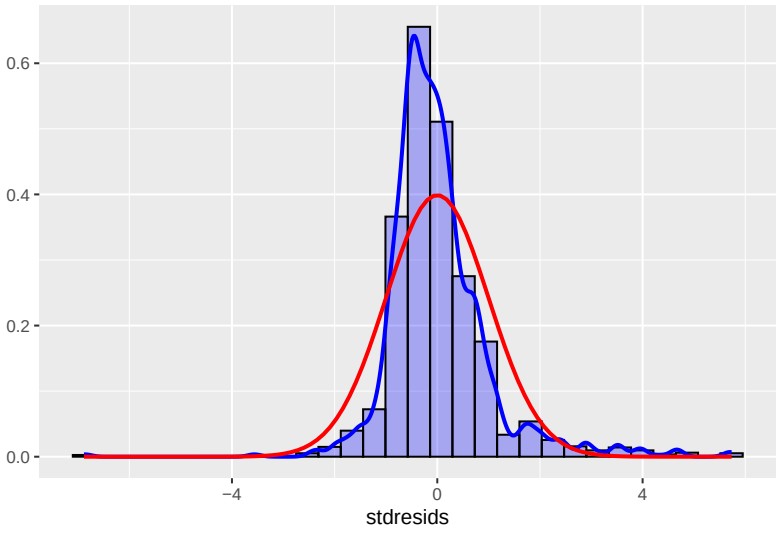

Figura 5.2.: Histograma de residuales estandarizados con curva de densidad estimada y distribución normal teórica

Según Chambers y Skinner (2003), la homocedasticidad de los errores (varianza constante) constituye uno de los supuestos centrales en los modelos de regresión. Su incumplimiento afecta la eficiencia de los estimadores y las varianzas basadas exclusivamente en el modelo; las varianzas robustas ajustadas al diseño ofrecen protección adicional, aunque no corrigen una especificación incorrecta de la media. Dado que los residuos representan las discrepancias entre los valores observados y los valores ajustados por el modelo, su análisis en función de los valores predichos o de las covariables constituye una herramienta diagnóstica fundamental. Bajo un modelo adecuadamente especificado, se espera que los residuos se distribuyan de forma aleatoria alrededor de cero, sin patrones estructurados; en cambio, la aparición de formas sistemáticas (como estructuras en embudo o curvaturas) sugiere la presencia de heterocedasticidad (varianza no constante) o posibles relaciones funcionales no captadas por la especificación del modelo.

Para evaluar este supuesto, se analizan gráficamente los residuos en función de los valores estimados $\hat{\mu}_k$ o de covariables específicas del modelo. La aparición de patrones sistemáticos en estos gráficos constituye evidencia de heterocedasticidad o de posibles problemas de especificación funcional. En particular, para evaluar la homocedasticidad respecto de las covariables incluidas en el modelo, se construyen gráficos de residuos frente a cada una de ellas y se combinan con `patchwork`.

```
library(patchwork)

g1 <- ggplot(
  data = survey_design$variables,
  aes(x = Expenditure, y = stdresids)
) +
  geom_point() +
  geom_hline(yintercept = 0)

g2 <- ggplot(
  data = survey_design$variables,
```

```
  aes(x = Zone, y = stdresids)
) +
  geom_point() +
  geom_hline(yintercept = 0)

g3 <- ggplot(
  data = survey_design$variables,
  aes(x = Sex, y = stdresids)
) +
  geom_point() +
  geom_hline(yintercept = 0)

(g1 | g2 | g3)
```

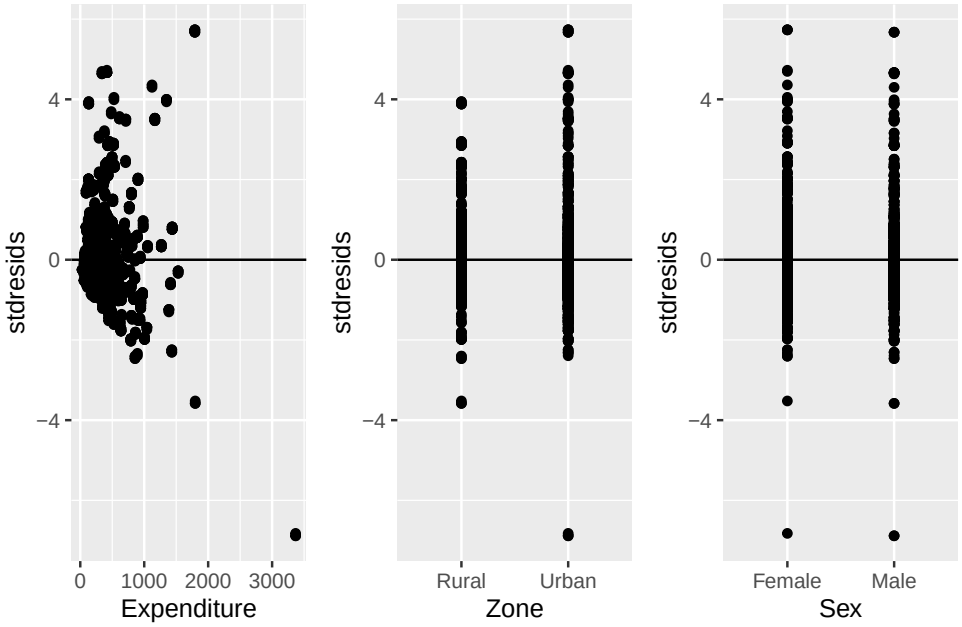

Figura 5.3.: Gráficos de residuales estandarizados frente a cada covariable del modelo

La figura 5.3 presenta los gráficos de dispersión de los residuos estandarizados frente a las covariables del modelo. En conjunto, estos resultados sugieren posibles limitaciones en la especificación funcional. En particular, para `Expenditure` se observa una ligera curvatura en la tendencia de los residuos, lo que podría indicar la presencia de una relación no lineal no completamente capturada por el modelo. Por su parte, en los gráficos asociados a `Zone` y `Sex` se aprecia una dispersión desigual de los residuos entre categorías, lo que sugiere diferencias en la variabilidad condicional o en el ajuste de la media entre grupos.

5.4.3. Observaciones influyentes

En el análisis diagnóstico, la identificación de observaciones influyentes es una técnica de especial importancia. Estas son unidades muestrales cuyo impacto sobre el ajuste del modelo resulta desproporcionado respecto al resto de la muestra. Conviene subrayar que una observación influyente no necesariamente corresponde a un valor atípico: mientras un atípico puede estar alejado del patrón general de los datos sin afectar sustancialmente el ajuste, una observación influyente puede alterar de manera significativa las estimaciones de los parámetros, incluso si no luce inusual.

Según Heeringa, West, y Berglund (2017, 213), la distancia de Cook mide el impacto que tendría la eliminación de la observación k sobre el ajuste global del modelo, al combinar simultáneamente la magnitud del residuo, la varianza estimada del modelo y su apalancamiento (*leverage*). En el contexto de encuestas con diseños muestrales complejos, la distancia de Cook suele denotarse como c_k y su expresión incorpora explícitamente los pesos muestrales. Para evaluar la magnitud de este estadístico, se utiliza una aproximación basada en su comparación con una distribución F, mediante la expresión:

$$\frac{(df - p + 1)\, c_k}{df} \sim F_{(p,\, df-p)},$$

Donde p denota los parámetros del modelo y df representa los grados de libertad asociados al diseño muestral. En la práctica, una observación puede considerarse influyente si su distancia de Cook supera valores de referencia aproximados entre 2 y 3.

En `R`, se evalúa la presencia de observaciones influyentes mediante la distancia de Cook, estimada con la función `svyCooksD` del paquete `svydiags`. Los resultados obtenidos se presentan en la figura 5.4, en donde se observa que ninguna de las distancias de Cook supera el umbral de referencia de 3, por lo que, bajo este criterio diagnóstico, no se identifican observaciones con influencia práctica elevada sobre el ajuste; el término «significativa» no corresponde aquí a una prueba de hipótesis.

```
cook_values <- svyCooksD(fit_svy, doplot = TRUE)
```

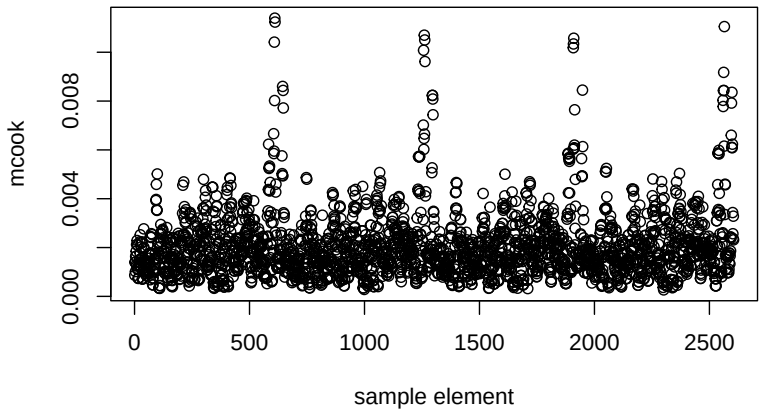

Figura 5.4.: Distancia de Cook para cada observación de la muestra

Otro estadístico que mide el grado de influencia de una observación es el DFBETA, denotado como $D_f\text{Beta}_{(k)j}$, que mide el cambio que experimenta el coeficiente de regresión $\widehat{\beta}_j$ cuando la observación k es eliminada del proceso de estimación. En consecuencia, esta medida evalúa la influencia individual de cada observación sobre los parámetros del modelo, identificando aquellas unidades cuya presencia altera de manera importante las estimaciones obtenidas. Valores elevados de $D_f\text{Beta}_{(k)j}$ indican que la observación ejerce una influencia considerable sobre el coeficiente correspondiente, lo que puede afectar la estabilidad e interpretación del modelo.

Como criterio práctico de diagnóstico, Li y Valliant (2015) señalan que una observación suele considerarse influyente sobre el coeficiente $\widehat{\beta}_j$ cuando

$$\left|D_f\text{Beta}_{(k)j}\right| \geq \frac{z}{\sqrt{n_I \times DEFF}},$$

Donde z toma comúnmente valores de 2 o 3, n_I representa el número de UPMs seleccionadas en la muestra de la primera etapa y $DEFF$ corresponde al efecto de diseño. A continuación, se evalúa la influencia de las observaciones sobre los coeficientes del modelo de ejemplo mediante la función `svydfbetas`. Los resultados correspondientes a las primeras diez observaciones se presentan en la tabla 5.1.

```
d_dfbetas <- data.frame(t(svydfbetas(fit_svy)$Dfbetas))
colnames(d_dfbetas) <- paste0("Beta_", 1:4)
d_dfbetas %>%
  slice(1:10L)
```

```
d_dfbetas <- data.frame(t(svydfbetas(fit_svy)$Dfbetas))
colnames(d_dfbetas) <- paste0("Beta_", 1:4)
d_dfbetas %>%
  slice(1:10L) %>%
  tabla_fmt()
```

Tabla 5.1.: Valores DfBetas para las primeras diez observaciones de la muestra

Beta_1	Beta_2	Beta_3	Beta_4
-0.0003	-0.0001	0.0013	-0.0033
-0.0009	-0.0001	0.0010	0.0024
-0.0009	-0.0001	0.0010	0.0024
-0.0003	-0.0001	0.0013	-0.0033
-0.0009	-0.0001	0.0010	0.0024
0.0025	0.0005	-0.0034	-0.0076
0.0025	0.0005	-0.0034	-0.0076
0.0005	0.0005	-0.0031	0.0075
0.0025	0.0005	-0.0034	-0.0076
-0.0007	0.0004	0.0007	-0.0029

Una vez obtenidos los valores de la medida DFBETA para cada observación y coeficiente del modelo, se calcula el correspondiente umbral de influencia a partir de los resultados generados por la función

`svydfbetas`. Con base en este criterio, se construye una variable dicotómica que permite clasificar las observaciones según su nivel de influencia sobre las estimaciones del modelo. Los resultados de esta clasificación se presentan en la tabla 5.2.

```
d_dfbetas$id <- 1:nrow(d_dfbetas)
d_dfbetas <- reshape2::melt(d_dfbetas, id.vars = "id")
cutoff <- svydfbetas(fit_svy)$cutoff
d_dfbetas <- d_dfbetas %>%
  mutate(criterion = ifelse(abs(value) > cutoff, "Yes", "No"))

label_text <- d_dfbetas %>%
  filter(criterion == "Yes") %>%
  arrange(desc(abs(value))) %>%
  slice(1:10L)

label_text
```

```
d_dfbetas$id <- 1:nrow(d_dfbetas)
d_dfbetas <- reshape2::melt(d_dfbetas, id.vars = "id")
cutoff <- svydfbetas(fit_svy)$cutoff
d_dfbetas <- d_dfbetas %>%
  mutate(criterion = ifelse(abs(value) > cutoff, "Yes", "No"))

label_text <- d_dfbetas %>%
  filter(criterion == "Yes") %>%
  arrange(desc(abs(value))) %>%
  slice(1:10L)

label_text %>%
  tabla_fmt()
```

Tabla 5.2.: Las diez observaciones más influyentes según el criterio DfBetas

id	variable	value	criterion
889	Beta_1	0.2837	Yes
891	Beta_1	0.2837	Yes
890	Beta_2	-0.2802	Yes
889	Beta_2	-0.2748	Yes
891	Beta_2	-0.2748	Yes
890	Beta_1	0.2497	Yes
890	Beta_3	0.1612	Yes
890	Beta_4	0.1580	Yes
889	Beta_3	0.1558	Yes
891	Beta_3	0.1558	Yes

Los resultados muestran que varias observaciones superan el umbral establecido por el criterio $D_f\text{Beta}_{(k)j}$, lo que indica que dichas unidades ejercen una influencia importante sobre la estimación

de algunos coeficientes del modelo, lo que sugiere la presencia de unidades con alto impacto sobre la estabilidad del ajuste. La figura 5.5 presenta la representación gráfica de los valores de D_fBeta junto con el umbral de referencia utilizado para identificar observaciones influyentes. En la figura, los puntos resaltados en rojo corresponden a las observaciones que exceden dicho umbral y que, por tanto, requieren una revisión más detallada dentro del proceso de diagnóstico del modelo.

```
ggplot(d_dfbetas, aes(y = abs(value), x = id)) +
  geom_point(aes(col = criterion)) +
  geom_text(
    data = label_text,
    angle = 45,
    vjust = -1,
    aes(label = id)
  ) +
  geom_hline(aes(yintercept = cutoff)) +
  facet_wrap(. ~ variable, nrow = 2) +
  scale_color_manual(
    values = c("Yes" = "red", "No" = "black")
  )
```

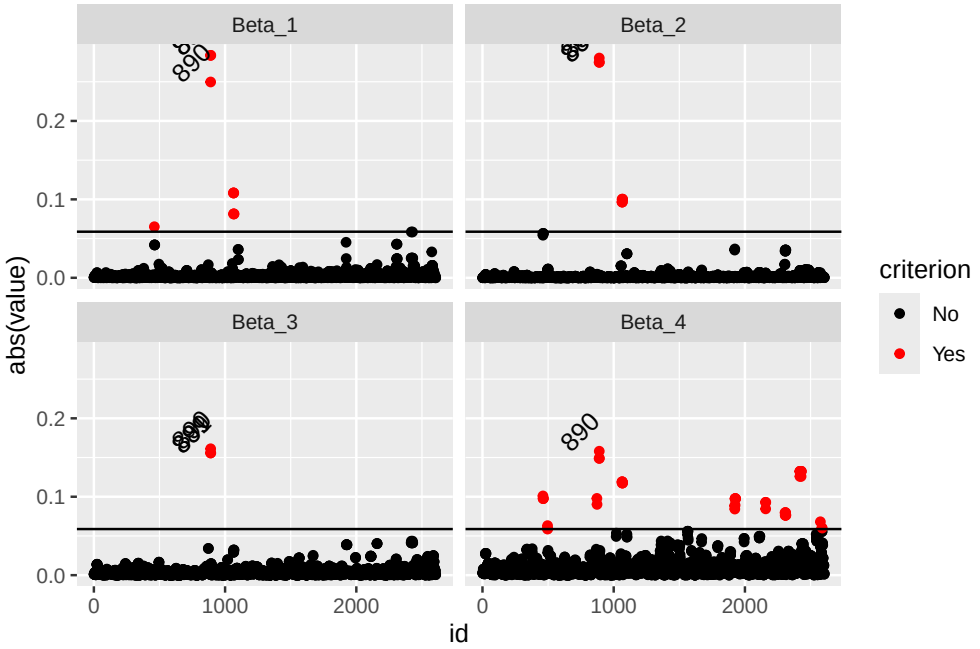

Figura 5.5.: Valores DfBetas por parámetro: observaciones influyentes señaladas en rojo

Por último, el estadístico DFFITS, denotado por $D_f\text{Fits}_{(k)}$, mide la influencia que ejerce la observación k sobre los valores ajustados del modelo. En particular, este estadístico evalúa cuánto cambian las predicciones del modelo cuando una observación es eliminada del proceso de estimación. En consecuencia, valores elevados de $D_f\text{Fits}_{(k)}$ indican observaciones con capacidad de alterar de manera importante el ajuste global del modelo y, por tanto, potencialmente influyentes en el proceso de inferencia. Una observación se considera influyente si:

$$|D_f\text{Fits}_{(k)}| \geq z\sqrt{\frac{p}{n \times DEFF}}$$

Donde z toma comúnmente valores de 2 o 3, mientras que n representa el tamaño muestral total de elementos incluidos en la encuesta. En el contexto de encuestas complejas, el uso de umbrales ajustados por efecto de diseño resulta especialmente relevante, ya que incorpora la pérdida de precisión derivada de la estratificación, la conglomeración y la utilización de pesos desiguales.

En `R`, el cálculo de este estadístico puede realizarse mediante la función `svydffits` del paquete `svydiags`, la cual obtiene las medidas de influencia considerando explícitamente la estructura del diseño muestral. Los resultados gráficos correspondientes se presentan en la figura 5.6, en donde se confirma la existencia de algunas observaciones influyentes (señaladas en rojo).

```
d_dffits <- data.frame(
  dffits = svydffits(fit_svy)$Dffits,
  id = 1:length(svydffits(fit_svy)$Dffits)
)
cutoff <- svydffits(fit_svy)$cutoff
d_dffits <- d_dffits %>%
  mutate(cutoff_criterion = ifelse(abs(dffits) > cutoff, "Yes", "No"))

ggplot(d_dffits, aes(y = abs(dffits), x = id)) +
  geom_point(aes(col = cutoff_criterion)) +
  geom_hline(yintercept = cutoff) +
  scale_color_manual(
    values = c("Yes" = "red", "No" = "black")
  )
```

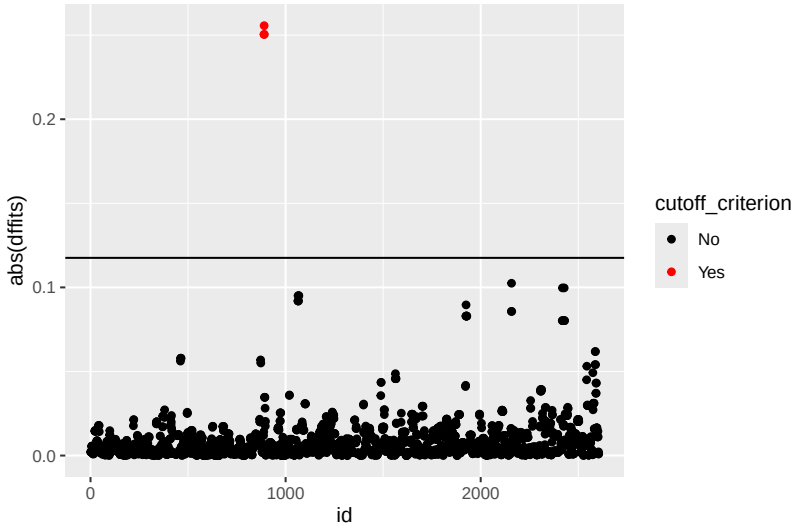

Figura 5.6.: Valores DfFits: observaciones influyentes sobre el ajuste global señaladas en rojo

5.5. Inferencia sobre los parámetros del modelo

Una vez evaluado el ajuste del modelo y examinados —sin asumir que se cumplen automáticamente— los supuestos asociados a los errores, el siguiente paso consiste en determinar la significación estadística de los parámetros estimados. Este análisis permite establecer si las covariables incluidas en el modelo aportan información relevante para explicar la variabilidad de la variable respuesta.

De acuerdo con Chambers y Skinner (2003) y Heeringa, West, y Berglund (2017), en modelos ajustados con datos provenientes de encuestas complejas, las pruebas de hipótesis sobre los coeficientes de regresión se construyen utilizando estimadores de varianza consistentes con el diseño muestral. Para evaluar el efecto de una covariable asociada al parámetro β_j, se considera usualmente el contraste

$$H_0 : \beta_j = 0 \qquad \text{frente a} \qquad H_1 : \beta_j \neq 0.$$

El estadístico de prueba correspondiente se define como

$$t_j = \frac{\hat{\beta}_j - \beta_j}{\sqrt{\widehat{Var}(\hat{\beta}_j)}}$$

el cual, bajo la hipótesis nula sigue aproximadamente una distribución t de Student con df grados de libertad asociados al diseño muestral (usualmente definidos como el número de unidades primarias de muestreo menos el número de estratos). La decisión inferencial se basa en comparar el valor absoluto del estadístico t_j con el valor crítico de la distribución de referencia. Cuando el valor observado del estadístico resulta suficientemente grande en magnitud, se rechaza la hipótesis nula y se concluye que la covariable correspondiente presenta una asociación estadísticamente significativa con la variable respuesta.

Las mismas propiedades distribucionales permiten construir intervalos de confianza para los coeficientes de regresión. Así, un intervalo de confianza al nivel $(1-\alpha) \times 100\,\%$ para el parámetro β_j está dado por

$$\hat{\beta}_j \pm t_{1-\frac{\alpha}{2}, df} \sqrt{\widehat{Var}(\hat{\beta}_j)},$$

donde $t_{1-\frac{\alpha}{2}, df}$ representa el cuantil de orden $1-\alpha/2$ de la distribución t de Student con df grados de libertad.

Para implementar estos procedimientos en `R`, se utilizan las funciones `summary.svyglm` y `confint.svyglm`, las cuales permiten obtener, respectivamente, las pruebas de significación basadas en el estadístico t y los intervalos de confianza de los coeficientes estimados bajo el diseño muestral complejo. Los resultados correspondientes se presentan en la tabla 5.3.

```
summary_table <- summary(fit_svy)$coefficients %>%
  as.data.frame() %>%
  tibble::rownames_to_column("Parameter")

confidence_table <- confint(fit_svy) %>%
  as.data.frame() %>%
```

```
  tibble::rownames_to_column("Parameter")

summary_table <- dplyr::left_join(
  summary_table,
  confidence_table,
  by = "Parameter"
)
```

Se incorporan explícitamente los límites producidos por `confint()` para que la tabla contenga tanto las pruebas t como los intervalos de confianza anunciados.

```
summary_table %>%
  tabla_fmt()
```

Tabla 5.3.: Pruebas t e intervalos de confianza al 95 % para los parámetros del modelo

Parameter	Estimate	Std. Error	t value	Pr(>|t|)	2.5 %	97.5 %
(Intercept)	73.580	59.4700	1.237	0.2185	-44.2077	191.368
Expenditure	1.222	0.1968	6.212	0.0000	0.8326	1.612
ZoneUrban	66.652	39.6655	1.680	0.0956	-11.9106	145.214
SexMale	20.644	15.6115	1.322	0.1886	-10.2762	51.565

Los resultados del modelo indican que la variable `Expenditure` presenta una asociación estadísticamente significativa con la variable de interés (`Income`). Este efecto resulta altamente significativo desde el punto de vista estadístico ($p < 0.001$), lo que proporciona evidencia sólida de una relación lineal positiva entre ambas variables. Por su parte, las variables categóricas `ZoneUrban` y `SexMale` presentan valores-p de 0.096 y 0.189, respectivamente, por lo que la evidencia no resulta concluyente al nivel convencional del 5 %.

5.6. Estimación y predicción

Según Heeringa, West, y Berglund (2017), los modelos de regresión cumplen dos objetivos fundamentales. El primero consiste en explicar la variabilidad de la variable respuesta a partir de las covariables disponibles, propósito que ha sido desarrollado a lo largo de este capítulo. El segundo corresponde a la predicción del valor esperado de la variable de interés para nuevas observaciones, tanto dentro como fuera del dominio muestral. La siguiente expresión representa el valor esperado de la variable respuesta y_k condicionada al vector de covariables observadas $\mathbf{x}_k$. En otras palabras, el modelo utiliza la información disponible de las variables explicativas para estimar el valor promedio esperado de la variable de interés para la unidad k.

$$\hat{\mu}_k = \hat{E}(y_k \mid \mathbf{x}_k) = \mathbf{x}_k \hat{\beta}$$

Asimismo, la varianza asociada a la estimación del valor esperado, definida como $Var(\hat{\mu}_k) = Var(\mathbf{x}_k \hat{\beta})$, puede estimarse mediante la siguiente expresión:

$$\widehat{Var}(\hat{\mu}_k) = \mathbf{x}_k' \, \hat{\Sigma} \, \mathbf{x}_k$$

Como explican Fox y Weisberg (2019), en donde $\hat{\Sigma} = \widehat{Var}(\hat{\beta})$ corresponde a la matriz de varianzas y covarianzas estimada de los parámetros del modelo de regresión. Esta expresión cuantifica únicamente la incertidumbre del valor medio estimado debida a los coeficientes; un intervalo de predicción para una nueva respuesta debe incorporar además la variabilidad residual. En consecuencia, aquellas unidades cuyos valores de las covariables estén asociados a regiones con mayor incertidumbre en la estimación de los coeficientes presentarán varianzas más elevadas. Por tanto, la precisión de las estimaciones depende directamente de la precisión con la que se estimen los coeficientes del modelo de regresión. Continuando con el modelo del ejemplo, recordemos que los coeficientes estimados del modelo se presentan en la tabla 5.4:

De acuerdo con Robinson et al. (2026), los coeficientes del modelo se convierten, mediante `broom::tidy()`, a una tabla ordenada, lo que facilita presentar estimaciones, errores estándar y pruebas en una estructura rectangular.

```
fit_svy %>%
  broom::tidy() %>%
  tabla_fmt()
```

Tabla 5.4.: Coeficientes estimados del modelo de regresión con diseño muestral complejo

term	estimate	std.error	statistic	p.value
(Intercept)	73.580	59.4700	1.237	0.2185
Expenditure	1.222	0.1968	6.212	0.0000
ZoneUrban	66.652	39.6655	1.680	0.0956
SexMale	20.644	15.6115	1.322	0.1886

A partir de los coeficientes estimados del modelo, la estimación del valor esperado de la variable respuesta para la unidad k puede expresarse como:

$$\hat{E}(y_k \mid \mathbf{x}_k) = 73.58 + 1.22 \times x_{1k} + 66.65 \times x_{2k} + 20.64 \times x_{3k}$$

En la expresión anterior se unifica el subíndice k para que coincida con la unidad definida en la formulación del modelo.

El siguiente código calcula las predicciones del modelo de regresión para las observaciones incluidas en la muestra, utilizando la matriz de diseño y el vector de coeficientes estimados:

```
x_obs <- model.matrix(fit_svy) %>%
  data.frame() %>%
  as.matrix()
hat_beta <- coef(fit_svy)
hat_mu <- as.numeric(x_obs %*% hat_beta)
hat_mu %>%
  head(10)
```

```
 [1] 517.6 496.9 496.9 517.6 496.9 553.1 553.1 573.7 553.1 184.8
```

En primer lugar, la instrucción `model.matrix(fit_svy)` construye la matriz de diseño asociada al modelo ajustado, incorporando automáticamente las variables explicativas. A continuación, `coef(fit_svy)` extrae explícitamente el vector de coeficientes estimados, evitando depender de la coincidencia parcial de nombres mediante `fit_svy$coe`. Posteriormente, se obtienen las estimaciones mediante el producto matricial `x_obs %*% hat_beta`. Finalmente, la función `head(10)` permite visualizar las primeras diez estimaciones obtenidas.

Para calcular la varianza de las estimaciones se requiere la matriz de varianzas y covarianzas de los parámetros del modelo. Los elementos de la diagonal principal corresponden a las varianzas de los parámetros estimados, mientras que los elementos fuera de la diagonal representan las covarianzas entre pares de coeficientes. La matriz estimada de varianzas y covarianzas se presenta en la tabla 5.5:

```
vcov(fit_svy) %>%
  as.data.frame() %>%
  tibble::rownames_to_column("Parameter")
```

```
vcov(fit_svy) %>%
  as.data.frame() %>%
  tibble::rownames_to_column("Parameter") %>%
  tabla_fmt()
```

Tabla 5.5.: Matriz de varianzas y covarianzas de los coeficientes del modelo

Parameter	(Intercept)	Expenditure	ZoneUrban	SexMale
(Intercept)	3536.68	-10.5505	123.689	326.1346
Expenditure	-10.55	0.0387	-3.295	-0.5704
ZoneUrban	123.69	-3.2949	1573.353	-121.4459
SexMale	326.13	-0.5704	-121.446	243.7178

En términos computacionales, esto implica combinar el vector de covariables de cada observación con la matriz $\hat{\Sigma}$ para obtener la precisión de las predicciones individuales. El siguiente código calcula la varianza asociada a las predicciones obtenidas a partir del modelo de regresión:

```
cov_beta <- vcov(fit_svy) %>%
  as.matrix()
mu_variance <- diag(x_obs %*% cov_beta %*% t(x_obs))
mu_variance %>%
  head(10)
```

```
   1    2    3    4    5    6    7    8    9   10
1375  874  874 1375  874 1218 1218 1667 1218 2998
```

Se utiliza `diag()` porque cada varianza corresponde a un elemento de la diagonal de $\mathbf{X}\hat{\Sigma}\mathbf{X}^\top$; convertir toda la matriz con `as.numeric()` produciría también las covarianzas entre observaciones y no el vector de varianzas individuales.

El intervalo de confianza, que proporciona un rango de valores plausibles para el valor esperado de la variable respuesta condicionado a las covariables observadas, se obtiene mediante:

$$\mathbf{x}_k\hat{\beta} \pm t_{(1-\frac{\alpha}{2},df)}\sqrt{\mathbf{x}_k'\ \hat{\Sigma}\ \mathbf{x}_k}$$

De acuerdo con Lumley et al. (2026), en `R`, la función `predict(fit_svy, type = "link")` genera estimaciones del valor medio en la escala del predictor lineal. El argumento `type = "link"` indica que las estimaciones se obtienen en la escala del predictor lineal, es decir, utilizando directamente la combinación lineal de las covariables y los coeficientes estimados. Posteriormente, la función `confint()` calcula los intervalos de confianza asociados a dichas estimaciones.

```
hat_mu <- data.frame(predict(fit_svy, type = "link"))
mu_ci <- data.frame(confint(predict(fit_svy, type = "link")))
colnames(mu_ci) <- c("lower_limit", "upper_limit")
```

Las estimaciones obtenidas a partir del modelo, junto con sus respectivos intervalos de confianza, pueden visualizarse gráficamente en la figura 5.7.

```
pred <- cbind(hat_mu, mu_ci)
pred$Expenditure <- survey_data$Expenditure
pd <- position_dodge(width = 0.2)
ggplot(
  pred %>%
    slice(1:100L),
  aes(x = Expenditure, y = link)
) +
  geom_errorbar(
    aes(ymin = lower_limit, ymax = upper_limit),
    width = .1,
    linetype = 1
  ) +
  geom_point(size = 2, position = pd) +
  theme_bw()
```

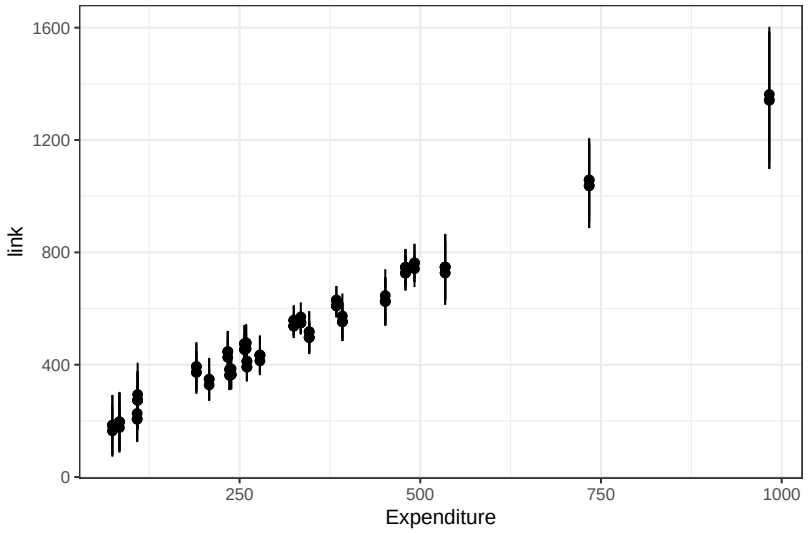

Figura 5.7.: Predicciones puntuales e intervalos de confianza para las primeras 100 observaciones de la muestra

Cuando el interés está en realizar predicciones para valores de las covariables que se encuentran fuera del rango observado en la muestra, la extrapolación puede aumentar la incertidumbre y depender fuertemente de la forma funcional supuesta. Además, para predecir una respuesta individual —dentro o fuera del rango observado— debe incorporarse la variabilidad inherente al término aleatorio de la regresión. Por esta razón, la varianza de predicción incluye un componente adicional asociado al error residual del modelo:

$$\widehat{Var}\left(y_{k,\text{nuevo}} - \hat{\mu}_{k,\text{nuevo}}\right) = \mathbf{x}_k' \, \hat{\Sigma} \, \mathbf{x}_k + \hat{\sigma}^2_{yx}$$

En esta expresión, el término $\hat{\sigma}^2_{yx}$ corresponde a la varianza residual estimada, es decir, la parte de la variabilidad de la variable respuesta que no es explicada por las covariables incluidas en el modelo. La incorporación de este segundo componente es fundamental cuando se desea predecir sobre unidades no observadas, ya que reconoce explícitamente la existencia de variabilidad individual alrededor de la media estimada.

En consecuencia, los intervalos de predicción suelen ser más amplios que los intervalos de confianza para la media, reflejando una mayor incertidumbre asociada a la predicción de valores individuales. De esta manera, el intervalo de predicción para una nueva observación puede expresarse como:

$$\mathbf{x}_k\hat{\beta} \pm t_{(1-\frac{\alpha}{2},df)} \sqrt{\mathbf{x}_k^\top \hat{\Sigma}_{\hat{\beta}} \mathbf{x}_k + \hat{\sigma}^2_{yx}}$$

Supóngase, por ejemplo, que se desea obtener la predicción del modelo para una nueva observación correspondiente a un individuo de sexo masculino (`Male`), residente en zona urbana (`Urban`) y con un nivel de gasto (`Expenditure`) igual a 1600. Este perfil puede definirse en `R` mediante el siguiente conjunto de datos:

```r
new_data <- data.frame(
  Expenditure = 1600,
  Sex = "Male",
  Zone = "Urban"
)

predict(fit_svy, newdata = new_data, type = "link")
```

```
  link  SE
1 2117 243
```

```r
confint(predict(fit_svy, newdata = new_data))
```

```
  2.5 % 97.5 %
1  1641   2593
```

En primer lugar, el objeto `new_data` especifica los valores de las covariables incluidas en el modelo. Posteriormente, la función `predict()` calcula la predicción puntual asociada a dicha observación. El argumento `newdata = new_data` indica que la predicción debe realizarse utilizando la nueva información suministrada, mientras que `type = "link"` solicita el resultado en la escala del predictor lineal. Finalmente, `confint()` aplicado al resultado de `predict.svyglm()` produce un intervalo de confianza para el valor medio estimado; no incorpora automáticamente el término residual requerido para un intervalo de predicción individual.

6. Modelos lineales generalizados

Nelder y Wedderburn (1972) introdujeron los modelos lineales generalizados como un marco capaz de integrar diversos métodos estadísticos ampliamente utilizados, como la regresión logística, los modelos log-lineales para datos de conteo y ciertos modelos avanzados de regresión para variables continuas. Posteriormente, McCullagh y Nelder (1989) sistematizaron esta formulación, que permitió extender las herramientas de regresión más allá de los escenarios en los que la variable respuesta sigue una distribución normal. Cuando la variable de interés no es continua o los supuestos del modelo lineal clásico no se satisfacen, los modelos lineales generalizados constituyen una alternativa adecuada para modelar las relaciones de interés.

La necesidad de esta generalización surge porque los modelos lineales clásicos asumen, entre otras cosas, normalidad y homogeneidad de varianza en la variable respuesta, condiciones que rara vez se cumplen cuando se analizan variables categóricas, proporciones o conteos. En este tipo de situaciones, los modelos lineales generalizados ofrecen un marco flexible para modelar adecuadamente la relación entre las variables mediante la selección de una distribución de probabilidad apropiada para la respuesta y una función de enlace que conecte la media de dicha distribución con el predictor lineal.

Para ilustrar los modelos presentados en este capítulo se utilizan la base de datos `survey_data` y la población `BigCity`, siguiendo la misma estructura utilizada en capítulos anteriores. Después de cargar `tidyverse`, `survey`, `srvyr`, `broom` y `jtools`, el siguiente bloque de código crea dos nuevas variables categóricas (`poor` y `unemployed`) a partir de las variables originales `Poverty` y `Employment`, respectivamente.

```
options(digits = 4)
library(tidyverse)
library(survey)
library(srvyr)
library(broom)
library(jtools)

survey_data <- readRDS("../Data/encuesta.rds")
data("BigCity", package = "TeachingSampling")

survey_design <- survey_data %>%
  as_survey_design(
    strata = Stratum,
    ids = PSU,
    weights = wk,
    nest = TRUE
  )
```

```r
survey_design <- survey_design %>%
  mutate(
    poor = factor(
      ifelse(Poverty != "NotPoor", 1, 0),
      levels = c(0, 1),
      labels = c("Not poor", "Poor")
    ),
    unemployed = factor(ifelse(Employment == "Unemployed", 1, 0),
      levels = c(0, 1),
      labels = c("Not unemployed", "Unemployed")
    )
  )
```

6.1. Modelo de regresión logística

De acuerdo con McCullagh y Nelder (1989) y Agresti (2019), cuando la variable de interés es dicotómica, tomando valores de cero o uno, el modelo lineal clásico no resulta adecuado debido a que puede generar predicciones fuera del intervalo $(0, 1)$ y porque la varianza de la respuesta depende de su media. Para abordar estas limitaciones, el modelo de regresión logística constituye uno de los modelos lineales generalizados más utilizados para analizar variables binarias. En el contexto de encuestas de hogares, la regresión logística puede emplearse para analizar la asociación entre una condición de interés, como la pobreza, y características sociodemográficas de los individuos o los hogares.

Considere, una variable dicotómica $Y_k \in \{0, 1\}$, cuya esperanza es $\theta_k = Pr(Y_k = 1)$. El modelo de regresión logística relaciona dicha probabilidad con un conjunto de variables explicativas mediante la función de enlace logit, definida como

$$\log\left(\frac{\theta_k}{1-\theta_k}\right) = \beta_0 + \beta_1 x_{k1} + \cdots + \beta_p x_{kp} = \mathbf{x}_k\beta$$

En donde β es el vector de coeficientes de regresión asociado a las covariables. Esta transformación logit convierte el intervalo restringido de probabilidades $(0, 1)$ en toda la recta real, permitiendo modelar la relación entre la respuesta y las variables auxiliares mediante un predictor lineal. De manera equivalente, la probabilidad de éxito puede expresarse mediante la siguiente relación:

$$\theta_k = Pr(Y_k = 1 \mid \mathbf{x}_k) = \frac{\exp(\mathbf{x}_k\beta)}{1 + \exp(\mathbf{x}_k\beta)}$$

De manera equivalente, la probabilidad de que la unidad no presente la característica de interés es $Pr(Y_k = 0 \mid \mathbf{x}_k) = 1 - \theta_k$. Bajo esta parametrización, cada coeficiente de β representa el efecto de una covariable sobre el logaritmo de la razón de probabilidades de que ocurra el evento de interés, manteniendo constantes las demás variables del modelo. Según Heeringa, West, y Berglund (2017), la estimación de los parámetros puede realizarse mediante el enfoque de pseudo-máxima verosimilitud. Este método incorpora los factores de expansión en la función de verosimilitud convencional con

el fin de obtener estimadores consistentes respecto de la población objetivo. La función de pseudo-verosimilitud para el modelo logístico se define como

$$PL(\beta) = \prod_k \left[\theta_k^{y_k} \{1-\theta_k\}^{1-y_k}\right]^{w_k}$$

En donde el subíndice k representa una unidad observada de la muestra. Asimismo, θ_k denota la probabilidad de que la unidad k de la UPM i en el estrato h presente la característica de interés, y_k corresponde al valor observado de la variable dicotómica para dicha unidad y w_k representa el factor de expansión asociado a la observación.

La contribución de cada observación al proceso de estimación está determinada por su respectivo factor de expansión, permitiendo que la función refleje adecuadamente las características del diseño complejo de la encuesta. Dado que la maximización de la pseudo-verosimilitud resulta equivalente a maximizar su logaritmo, entonces el problema se reduce a optimizar la siguiente función:

$$\ell(\beta) = \sum_k w_k \left[y_k \log(\theta_k) + (1-y_k)\log(1-\theta_k)\right]$$

Esta expresión constituye la función objetivo utilizada por los algoritmos de optimización numérica para obtener el estimador de los parámetros del modelo. Es importante señalar que la función anterior no corresponde a una verosimilitud en sentido estricto bajo el diseño complejo de la encuesta. Por esta razón, se utiliza el término *pseudo-verosimilitud* y los estimadores obtenidos mediante su maximización son conocidos como estimadores de pseudo-máxima verosimilitud (*Pseudo Maximum Likelihood Estimators*, PMLE).

De acuerdo con Molina y Skinner (1992) y Chambers y Skinner (2003), bajo condiciones de regularidad, estos estimadores son consistentes y asintóticamente normales, constituyendo la base de procedimientos de inferencia para el análisis de datos de encuestas complejas.

Como explica David A. Binder (1983), una vez obtenidos los estimadores de pseudo-máxima verosimilitud, sus varianzas se estiman, mediante técnicas de linealización de Taylor, las cuales incorporan explícitamente las características del diseño complejo de muestreo. La matriz de varianzas y covarianzas de los parámetros estimados, denotada por $\text{Var}(\widehat{\beta}) = \Sigma$, se obtiene a partir de una aproximación lineal de las ecuaciones de estimación alrededor del verdadero valor de los parámetros.

Esta matriz describe la precisión de los estimadores, ya que sus elementos diagonales corresponden a las varianzas de cada coeficiente, mientras que los elementos fuera de la diagonal representan las covarianzas entre pares de coeficientes. Su estimación constituye la base para el cálculo de errores estándar, la construcción de intervalos de confianza y la realización de pruebas de hipótesis sobre los parámetros del modelo.

Como los datos provienen de un diseño muestral complejo, es necesario incorporar adecuadamente las ponderaciones, la estratificación y la conglomeración para obtener estimadores consistentes de los parámetros y de sus errores estándar. En `R`, esta tarea puede realizarse mediante la función `svyglm()` del paquete `survey`, la cual extiende los modelos lineales generalizados al contexto de encuestas complejas.

Siguiendo con la encuesta de ejemplo, y con el fin de analizar los factores asociados a la condición de pobreza, se ajusta un modelo de regresión logística utilizando como variables explicativas el gasto, la

situación en el empleo y el sexo. Dado que los datos provienen de una encuesta con diseño muestral complejo, la estimación se realiza mediante la función `svyglm()` del paquete `survey`, especificando `family = binomial` para indicar que la variable respuesta es dicotómica y que se utilizará la función de enlace logit. Los resultados del ajuste se presentan en la tabla 6.1.

```
logit_model <- svyglm(
  formula = poor ~ Expenditure + Employment + Sex,
  family = binomial,
  design = survey_design
)
```

En la especificación del modelo, el argumento `formula = poor ~ Expenditure + Employment + Sex` define la relación entre la variable respuesta `poor` y las covariables incluidas en el predictor lineal. Para las variables categóricas, la función genera automáticamente las variables indicadoras correspondientes y estima sus efectos con respecto a una categoría de referencia. Por su parte, el argumento `design = survey_design` incorpora la información del diseño muestral. El objeto resultante, `logit_model`, contiene los parámetros estimados, sus medidas de precisión y las estadísticas necesarias para realizar inferencia sobre la relación entre la condición de pobreza y las variables explicativas consideradas en el modelo.

Los resultados de la tabla 6.1 muestran las estimaciones de los coeficientes de regresión, junto con sus errores estándar, intervalos de confianza y valores-p. Del modelo se deduce que un mayor nivel de gasto se asocia con una menor probabilidad de encontrarse en condición de pobreza. Asimismo, en comparación con la categoría de referencia (desocupados), tanto las personas inactivas como las ocupadas presentan una menor propensión a ser pobres, siendo este efecto más marcado entre estas últimas. En contraste, el sexo tiene una influencia muy reducida sobre la probabilidad de pobreza una vez controlado el efecto de las demás covariables incluidas en el modelo.

```
logit_results %>%
  as.data.frame() %>%
  tibble::rownames_to_column("Parameter") %>%
  tabla_fmt()
```

Tabla 6.1.: Intervalos de confianza al 95 % para los parámetros del modelo logístico

Parameter	term	estimate	std.error	statistic	p.value	conf.low	conf.high
1	(Intercept)	2.1877	0.3731	5.863	0.0000	1.449	2.9269
2	Expenditure	-0.0054	0.0008	-6.497	0.0000	-0.007	-0.0038
3	EmploymentInactive	-0.8634	0.3811	-2.266	0.0253	-1.618	-0.1085
4	EmploymentEmployed	-1.4069	0.3381	-4.161	0.0001	-2.077	-0.7371
5	SexMale	0.0025	0.1573	0.016	0.9873	-0.309	0.3141

La figura 6.1 presenta la distribución de los estimadores de los coeficientes de regresión. Se observa que, para todas las covariables excepto el sexo, los intervalos de confianza no incluyen el valor cero, lo que proporciona evidencia de una asociación estadísticamente significativa. En contraste, el intervalo de confianza asociado a la variable sexo contiene el valor cero, por lo que no se encuentra

evidencia suficiente para concluir que esta variable tenga un efecto significativo una vez controladas las demás covariables incluidas en el modelo.

De acuerdo con Henry, Wickham, y Chang (2024) y Long (2026), para visualizar la distribución de los coeficientes se emplea `ggstance` para la geometría horizontal de los intervalos, junto con `plot_summs()` del paquete `jtools`.

```
library(ggstance)
plot_summs(logit_model,
  scale = TRUE,
  plot.distributions = TRUE
)
```

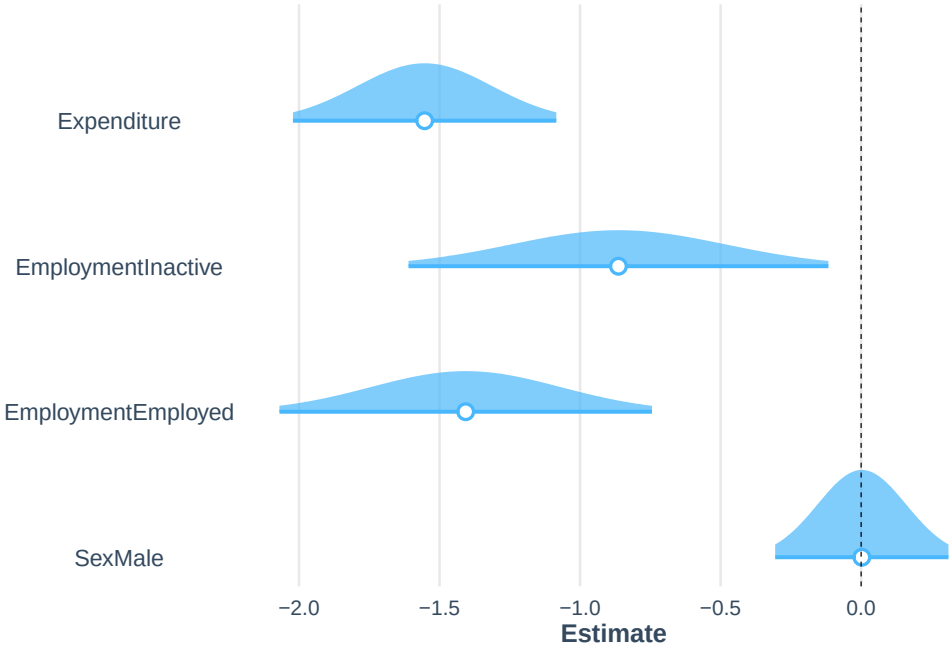

Figura 6.1.: Distribución de los parámetros del modelo logístico base

El modelo puede ampliarse incorporando términos de interacción para evaluar si el efecto de una variable explicativa depende de los valores que toma otra. En este caso, se incluye una interacción entre el sexo y la situación laboral mediante el término `Sex:Employment`, con el objetivo de analizar si la relación entre el estado ocupacional y la probabilidad de pobreza es la misma para hombres y mujeres. Sin este término, el modelo asume que el efecto de la situación laboral sobre la pobreza es idéntico para ambos sexos. En cambio, al incorporar la interacción, se permite que dicho efecto varíe entre hombres y mujeres, capturando posibles diferencias en la forma en que la condición laboral se asocia con la pobreza en cada grupo. Los resultados del modelo ajustado se presentan en la tabla 6.2.

```
interaction_logit_model <- svyglm(
  formula = poor ~ Expenditure + Employment + Sex + Sex:Employment,
  family = binomial,
```

```
  design = survey_design
)
```

```
interaction_logit_results %>%
  tabla_fmt()
```

Tabla 6.2.: Coeficientes del modelo logístico con interacciones

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	1.7698	0.6103	2.8996	0.0045	0.5606	2.9790
Expenditure	-0.0054	0.0008	-6.4726	0.0000	-0.0070	-0.0037
EmploymentInactive	-0.3952	0.5941	-0.6652	0.5072	-1.5721	0.7817
EmploymentEmployed	-1.0593	0.5696	-1.8599	0.0655	-2.1877	0.0691
SexMale	0.5872	0.7479	0.7851	0.4340	-0.8945	2.0689
EmploymentInactive:SexMale	-0.8465	0.8600	-0.9843	0.3271	-2.5503	0.8573
EmploymentEmployed:SexMale	-0.4812	0.7604	-0.6328	0.5281	-1.9878	1.0254

Al incorporar la interacción entre sexo y situación laboral, el gasto continúa mostrando una asociación negativa y estadísticamente significativa con la probabilidad de pobreza, manteniéndose como la covariable con mayor evidencia explicativa dentro del modelo. En contraste, los efectos principales de la situación laboral pierden importancia estadística una vez que se incluyen los términos de interacción, lo que sugiere una mayor incertidumbre en la estimación de sus efectos. Por su parte, ni el efecto principal del sexo ni los coeficientes asociados a las interacciones entre sexo y situación laboral resultan estadísticamente significativos al nivel del 5 %. Estos resultados individuales no bastan para afirmar que la interacción no aporta información; esa conclusión requiere una prueba conjunta de los términos de interacción ajustada al diseño.

La figura 6.2 presenta una comparación de las estimaciones de los coeficientes y sus respectivos intervalos de confianza para los modelos con y sin interacción. Se observa que la incorporación de los términos de interacción incrementa la incertidumbre asociada a varios de los parámetros, lo que se refleja en intervalos de confianza más amplios. Asimismo, los coeficientes de interacción presentan estimaciones cercanas a cero y amplios intervalos de confianza que incluyen dicho valor, lo que coincide con la falta de evidencia estadística para concluir que el efecto de la situación laboral sobre la probabilidad de pobreza difiere entre hombres y mujeres.

```
plot_summs(interaction_logit_model, logit_model,
  scale = TRUE,
  plot.distributions = TRUE
)
```

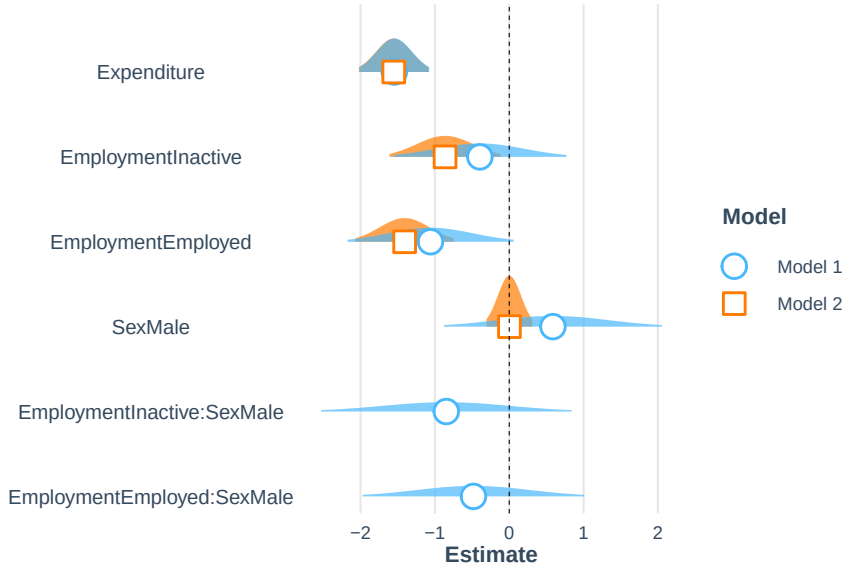

Figura 6.2.: Comparación de los parámetros del modelo logístico con y sin interacciones

6.2. Modelo de regresión multinomial

De acuerdo con McCullagh y Nelder (1989) y Agresti (2019), en las encuestas de hogares es frecuente encontrar variables de respuesta con más de dos categorías, como la situación laboral, cuyos posibles estados incluyen empleado, desempleado e inactivo. Cuando estas categorías son mutuamente excluyentes y no poseen un orden natural, la regresión logística multinomial ofrece una herramienta adecuada para modelar su relación con un conjunto de covariables y puede entenderse como una extensión de la regresión logística binaria.

Este modelo permite estimar la probabilidad de pertenecer a cada categoría de la variable respuesta en función de variables explicativas de naturaleza continua o categórica. Su aplicación requiere que las categorías de respuesta sean exhaustivas y mutuamente excluyentes y que no exista multicolinealidad severa entre las covariables. Además, para las covariables continuas, se asume una relación lineal entre estas y el logaritmo de las razones de probabilidades (*log-odds*) de cada categoría respecto a una categoría de referencia. El logit multinomial de categoría base también implica la independencia de alternativas irrelevantes, por lo que conviene valorar si ese supuesto es razonable para las categorías analizadas.

El modelo logístico multinomial especifica, la probabilidad $\theta_{k(j)}$ de que la unidad k pertenezca a la categoría j, con $j = 1, \dots, J$, dado el vector de covariables $\mathbf{x}_k$, mediante la siguiente expresión:

$$Pr(Y_k = j \mid \mathbf{x}_k) = \theta_{k(j)} = \frac{\exp(\mathbf{x}_k \beta_j)}{\sum_{l=1}^{J} \exp(\mathbf{x}_k \beta_l)}$$

En donde β_j es el vector de coeficientes de regresión para la j-ésima categoría. Para identificar el modelo debe fijarse el vector de una categoría de referencia en cero; de ese modo, la expresión produce $J-1$ conjuntos de coeficientes estimables. Esta formulación expresa directamente las probabilidades de pertenencia a cada categoría, y en la práctica resulta más conveniente reescribir el

modelo utilizando una categoría de referencia. Hay parametrizaciones alternativas que facilitan la interpretación de los parámetros, ya que cada conjunto de coeficientes describe el efecto de las covariables sobre el logaritmo de las razones de probabilidades (*log-odds*) de pertenecer a una determinada categoría en comparación con la categoría de referencia.

Según Heeringa, West, y Berglund (2017), la estimación de los parámetros del modelo se realiza mediante el enfoque de pseudo-máxima verosimilitud. La función de pseudo-verosimilitud multinomial se define como

$$PL(\beta \mid \mathbf{X}) = \prod_{k} \left\{ \prod_{j=1}^{J} \theta_{k(j)}^{y_{k(j)}} \right\}^{w_k}$$

En donde k representa las unidades de observación pertenecientes a la muestra compleja. Asimismo, $y_{k(j)}$ es una variable indicadora que toma el valor de uno cuando dicha unidad pertenece a la categoría j y cero en caso contrario, y w_k corresponde al factor de expansión asociado a la unidad observada. Por conveniencia matemática y computacional, la estimación de los parámetros se realiza a partir de la pseudo-logverosimilitud, cuya maximización conduce a los mismos estimadores que la pseudo-verosimilitud original.

$$\ell(\beta) = \sum_{k} w_k \sum_{j=1}^{J} y_{k(j)} \log(\theta_{k(j)})$$

De acuerdo con Molina y Skinner (1992) y David A. Binder (1983), la inferencia estadística para los parámetros de interés se basa en estimaciones de varianza obtenidas mediante técnicas de linealización de Taylor que incorporan explícitamente las características del diseño muestral complejo, permitiendo calcular errores estándar, intervalos de confianza y pruebas de hipótesis apropiados para datos de encuestas. Bajo condiciones regulares, los estimadores obtenidos mediante este procedimiento son consistentes y asintóticamente normales.

Para el ajuste del modelo usando `R`, como referencia inicial, se estiman las proporciones de personas por estado de empleo, restringiendo las unidades de interés a mayores de 15 años, como se presenta en la tabla 6.3.

```
survey_design %>%
  filter(Age >= 15) %>%
  group_by(Employment) %>%
  summarise(proportion = survey_mean(vartype = c("se", "ci")))
```

```
employment_proportion %>%
  tabla_fmt()
```

Tabla 6.3.: Proporción estimada de personas por estado de empleo (mayores de 15 años)

Employment	proportion	proportion_se	proportion_low	proportion_upp
Unemployed	0.0429	0.0071	0.0288	0.0571
Inactive	0.3840	0.0152	0.3539	0.4142

Tabla 6.3.: Proporción estimada de personas por estado de empleo (mayores de 15 años)

Employment	proportion	proportion_se	proportion_low	proportion_upp
Employed	0.5731	0.0140	0.5453	0.6008

De acuerdo con Lumley (2026), para modelar una variable de respuesta multinomial se emplea la función `svy_vglm()` del paquete `svyVGAM`, que permite ajustar modelos de regresión multinomial incorporando explícitamente las características del diseño complejo de la encuesta. En este ejemplo, se especifica un modelo donde la condición de actividad se explica en función de la edad, el sexo y la zona de residencia. A diferencia de los modelos multinomiales convencionales, `svy_vglm()` calcula las varianzas de los estimadores considerando la complejidad del diseño muestral, incluyendo la estratificación, la conglomeración y los pesos de expansión. De esta forma, tanto las estimaciones puntuales como sus errores estándar e inferencias asociadas son consistentes con el diseño de la encuesta.

```
library(svyVGAM)
survey_design_15 <- survey_design %>%
  filter(Age >= 15)
multinomial_model <- svy_vglm(
  formula = Employment ~ Age + Sex + Zone,
  design = survey_design_15,
  crit = "coef",
  family = multinomial(refLevel = "Unemployed")
)
```

El argumento `family = multinomial(refLevel = "Unemployed")` define un modelo logístico multinomial tomando la categoría `Unemployed` como referencia. Bajo esta especificación, se estiman de manera simultánea los logaritmos de las razones de probabilidades (*log-odds*) entre cada una de las categorías de `Employment` y la categoría de referencia. Los parámetros se obtienen mediante pseudo-máxima verosimilitud, incorporando los factores de expansión de la encuesta en la función objetivo.

Dado que `broom::tidy()` no puede aplicarse directamente a objetos de clase `svyVGAM`, se define la siguiente función auxiliar para estructurar los resultados del modelo en un formato tabular estandarizado[1]:

```
tidy.svyVGAM <- function(x, conf.int = FALSE, conf.level = 0.95,
                         exponentiate = FALSE, ...) {
  ret <- as_tibble(summary(x)$coeftable, rownames = "term")
  colnames(ret) <- c("term", "estimate", "std.error", "statistic", "p.value")
  coefs <- tibble::enframe(stats::coef(x), name = "term", value = "estimate")
  ret <- left_join(coefs, ret, by = c("term", "estimate"))
  if (conf.int) {
    ci <- broom:::broom_confint_terms(x, level = conf.level, ...)
    ret <- dplyr::left_join(ret, ci, by = "term")
```

[1] Adaptada de https://tech.popdata.org/pma-data-hub/posts/2021-08-15-covid-analysis/

```
  }
  if (exponentiate) ret <- broom:::exponentiate(ret)
  ret %>%
    tidyr::separate(term, into = c("term", "y.level"), sep = ":") %>%
    arrange(y.level) %>%
    relocate(y.level, .before = term)
}
```

La tabla 6.4 presenta los coeficientes estimados del modelo logístico multinomial, tomando la categoría `Unemployed` como referencia. Los resultados muestran que la edad tiene una asociación positiva y estadísticamente significativa con las probabilidades relativas de pertenecer a las categorías `Inactive` y `Employed` frente a `Unemployed`. Asimismo, la variable `SexMale` presenta coeficientes negativos y significativos, indicando que los hombres tienen menores probabilidades relativas que las mujeres de ubicarse en dichas categorías. Por el contrario, la variable de zona de residencia no muestra efectos estadísticamente significativos al nivel convencional del 5 %.

```
tidy.svyVGAM(multinomial_model, exponentiate = FALSE, conf.int = FALSE)
```

```
multinomial_results %>%
  tabla_fmt()
```

Tabla 6.4.: Coeficientes estimados del modelo de regresión multinomial

y.level	term	estimate	std.error	statistic	p.value
1	(Intercept)	2.5565	0.5500	4.6481	0.0000
1	Age	0.0221	0.0104	2.1168	0.0343
1	SexMale	-2.1851	0.3166	-6.9015	0.0000
1	ZoneUrban	-0.2536	0.4045	-0.6270	0.5307
2	(Intercept)	2.3063	0.4614	4.9983	0.0000
2	Age	0.0181	0.0088	2.0556	0.0398
2	SexMale	-0.5783	0.2766	-2.0908	0.0365
2	ZoneUrban	0.0392	0.3605	0.1087	0.9135

De acuerdo con Warnes et al. (2023) y Solt y Hu (2025), la figura 6.3 muestra los intervalos de confianza de los coeficientes estimados para cada ecuación del modelo multinomial. Se usa `gtools::stars.pval()` para codificar la significación estadística, y la visualización se construye con `dotwhisker::dwplot()`, lo que permite comparar coeficientes e intervalos en una escala común. Las covariables cuyos intervalos no incluyen el valor cero presentan evidencia de asociación estadística significativa con la variable respuesta, mientras que aquellos cuyos intervalos contienen el cero no muestran efectos significativos al nivel de confianza considerado.

```
multinomial_results %>%
  mutate(
    model = if_else(y.level == 1, "Inactive", "Employed"),
    sig = gtools::stars.pval(p.value)
```

```
) %>%
dotwhisker::dwplot(
  dodge_size = 0.3,
  vline = geom_vline(xintercept = 0, colour = "grey60", linetype = 2)
) +
guides(color = guide_legend(reverse = TRUE)) +
theme_bw() +
theme(legend.position = "top")
```

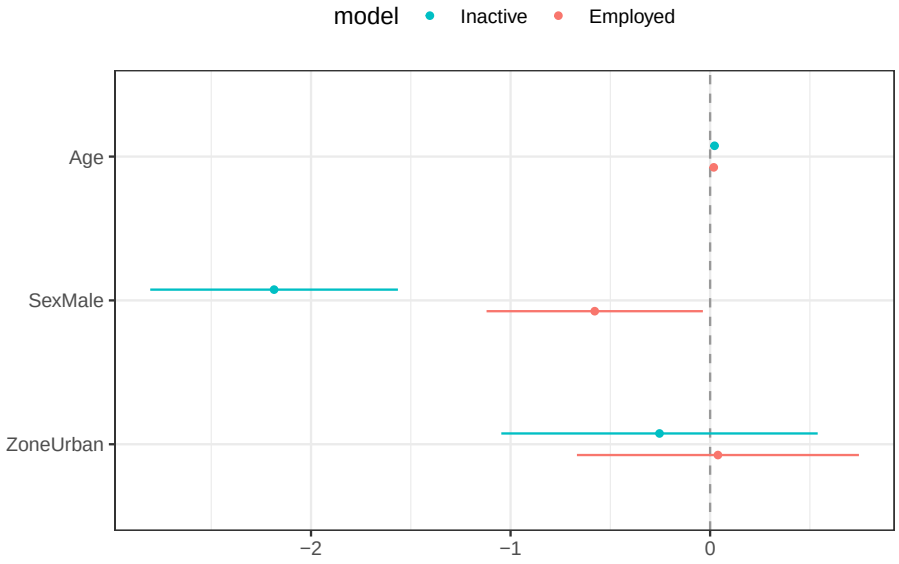

Figura 6.3.: Intervalos de confianza para los coeficientes del modelo multinomial

6.3. Modelo de regresión Gamma

Según McCullagh y Nelder (1989), la regresión Gamma constituye una extensión de los modelos lineales generalizados adecuada para variables respuesta continuas, estrictamente positivas y cuya variabilidad aumenta con la media. Este comportamiento es frecuente en variables económicas y sociales, como los ingresos de los hogares o los gastos en consumo.

Si Y_k, la variable de interés observada para la unidad k, sigue una distribución Gamma con media μ_k y parámetro de dispersión ϕ, el modelo relaciona el valor esperado de la variable respuesta con un conjunto de covariables mediante una función de enlace. El enlace logarítmico es una opción frecuente y garantiza predicciones positivas, aunque el enlace canónico de la familia Gamma es el inverso. En este caso, el modelo se expresa como

$$\log(\mu_k) = \mathbf{x}_k\beta$$

En donde $\mathbf{x}_k$ corresponde al vector de variables explicativas asociado a la unidad observada y β es el vector de parámetros desconocidos del modelo. De forma equivalente, se tiene que:

$$\mu_k = E(Y_k \mid \mathbf{x}_k) = \exp(\mathbf{x}_k\beta),$$

Con el enlace logarítmico, cada parámetro puede interpretarse como el efecto de una covariable sobre el logaritmo de la media esperada, de manera que $\exp(\beta_j)$ representa el cambio multiplicativo asociado a un incremento unitario en la covariable correspondiente, manteniendo constantes las demás variables del modelo. Bajo la distribución Gamma, la función de densidad condicional de Y_k puede escribirse como

$$f(y_k \mid \mu_k, \phi) = \frac{1}{\Gamma(1/\phi)(\phi\mu_k)^{1/\phi}} y_k^{\frac{1}{\phi}-1} \exp\left(-\frac{y_k}{\phi\mu_k}\right), \qquad y_k > 0,$$

donde $\Gamma(\cdot)$ denota la función Gamma y ϕ representa el parámetro de dispersión. Bajo el enfoque de inferencia basado en pseudo-máxima verosimilitud, la función objetivo que se optimiza para obtener los estimadores de los parámetros del modelo está dada por

$$\ell(\beta, \phi) = \sum_k w_k \log\left[f(y_k \mid \mu_k, \phi)\right]$$

En donde k representa las unidades de observación de la muestra y w_k corresponde al factor de expansión asociado a cada una de ellas. Esta función define el criterio de optimización utilizado para estimar los parámetros del modelo bajo el enfoque de pseudo-máxima verosimilitud, descrito para encuestas complejas por Chambers y Skinner (2003). En `R`, se hace necesario primero definir el diseño de muestreo.

```
gamma_design <- survey_data %>%
  as_survey_design(
    strata = Stratum,
    ids = PSU,
    weights = wk,
    nest = TRUE
  )
```

El siguiente código ajusta, mediante `svyglm()`, un modelo de regresión Gamma para la variable de ingresos del hogar (`Income`), incorporando el diseño complejo de la encuesta definido en el objeto `gamma_design`. El modelo utiliza como variables explicativas el gasto del hogar (`Expenditure`), la edad (`Age`), el sexo (`Sex`) y la zona de residencia (`Zone`). Asimismo, se especifica una función de enlace logarítmica, garantizando predicciones positivas y permitiendo interpretar los efectos de las variables explicativas en términos multiplicativos sobre el ingreso esperado. Los coeficientes estimados se presentan en la tabla 6.5:

```
gamma_model <- svyglm(
  formula = Income ~ Expenditure + Age + Sex + Zone,
  design = gamma_design,
  family = Gamma(link = "log")
)
```

```
gamma_results %>%
  tabla_fmt()
```

Tabla 6.5.: Coeficientes del modelo Gamma para el ingreso de los hogares

term	estimate	std.error	statistic	p.value
(Intercept)	5.4094	0.0980	55.184	0.0000
Expenditure	0.0018	0.0001	16.891	0.0000
Age	0.0015	0.0006	2.430	0.0166
SexMale	0.0484	0.0384	1.260	0.2104
ZoneUrban	0.1514	0.0890	1.701	0.0916

Dado que el modelo utiliza un enlace logarítmico, los coeficientes pueden interpretarse como efectos multiplicativos sobre el ingreso esperado. El coeficiente asociado al gasto del hogar es positivo y estadísticamente significativo ($p < 0.001$), indicando que, manteniendo constantes las demás variables del modelo, un incremento de una unidad en el gasto se asocia con un aumento aproximado del 0.2 % en el ingreso esperado, ya que $\exp(0.002) \approx 1.002$. La edad también presenta un efecto positivo y significativo ($p = 0.017$). En promedio, cada año adicional de edad se asocia con un incremento cercano al 0.2 % en el ingreso esperado, manteniendo constantes las demás covariables. Por el contrario, las variables de sexo y zona de residencia no muestran evidencia estadística suficiente para concluir que influyen sobre el ingreso.

7. Modelos multinivel

De acuerdo con Goldstein (2011) y Rabe-Hesketh y Skrondal (2012), los modelos multinivel, también conocidos como modelos jerárquicos, son una técnica estadística diseñada para el análisis de datos con estructura jerárquica, como los que provienen de encuestas de hogares, en donde las unidades de observación no son independientes entre sí. Los individuos pertenecen a hogares, los hogares se ubican en áreas geográficas específicas (UPMs) y estas, a su vez, forman parte de unidades territoriales más amplias (estratos o dominios). Como consecuencia, las observaciones que comparten un mismo contexto suelen presentar características más similares entre sí que aquellas pertenecientes a contextos diferentes. Esta estructura de dependencia viola uno de los supuestos fundamentales de los modelos de regresión convencionales y puede conducir a estimaciones ineficientes e inferencias incorrectas.

A diferencia de los modelos de regresión convencionales, los modelos multinivel reconocen que las observaciones pueden estar agrupadas en distintos niveles y que cada uno de ellos puede contribuir a explicar la variabilidad observada. En consecuencia, permiten representar simultáneamente los efectos asociados a las características de las unidades de análisis y aquellos derivados del entorno o contexto en el que estas se encuentran. Esta formulación resulta especialmente útil cuando se busca estudiar fenómenos determinados tanto por factores individuales como por características de los hogares, comunidades o áreas geográficas.

Como explica Snijders y Bosker (2011), otra característica relevante de los modelos multinivel es que permiten cuantificar la proporción de la variabilidad asociada a cada nivel de agrupación presente en los datos. Mientras los efectos fijos resumen la relación promedio entre la variable respuesta y las covariables incluidas en el modelo, los efectos aleatorios representan las diferencias sistemáticas entre los grupos que no son explicadas por dichas covariables. Gracias a esta estructura, es posible modelar explícitamente la dependencia entre observaciones pertenecientes a un mismo grupo y obtener inferencias más adecuadas cuando los datos presentan organización jerárquica.

De acuerdo con Gelman y Hill (2006), el desarrollo teórico y metodológico de los modelos multinivel ha sido ampliamente documentado en la literatura de muestreo. Sus fundamentos conceptuales, estrategias de estimación y aplicaciones en contextos sociales y demográficos se encuentran consolidados en distintas referencias especializadas.

Como muestra Browne y Draper (2006), por otra parte, la comparación entre enfoques de estimación para modelos jerárquicos permite evaluar diferencias entre los métodos basados en máxima verosimilitud y los enfoques bayesianos.

Según Merlo et al. (2006), en el ámbito de la epidemiología social, los modelos multinivel resultan útiles para estudiar fenómenos contextuales, mostrando cómo factores asociados al entorno pueden contribuir a explicar desigualdades en salud y otros resultados de interés poblacional.

Como advierten Pfeffermann et al. (1998) y Carle (2009), es necesario distinguir entre la jerarquía sustantiva que se desea modelar y la estructura del diseño muestral: una UPM, un estrato de diseño o un dominio no constituye automáticamente un contexto intercambiable susceptible de

representarse mediante un efecto aleatorio. Esa decisión requiere una justificación conceptual y suficiente información en cada grupo.

7.1. Motivación

De acuerdo con Bates et al. (2015), para iniciar este capítulo, se cargan `tidyverse` para la manipulación de datos y la generación de gráficos, y `lme4` para la estimación de modelos multinivel. Asimismo, se importa la base de datos que será utilizada a lo largo de los ejemplos. Los títulos matemáticos de algunas figuras se formatean con `latex2exp`, llamado de forma explícita en el código.

```
library(tidyverse)
library(lme4)
```

```
survey_data <- readRDS("../Data/encuesta.rds") %>%
  mutate(poor = ifelse(Poverty != "NotPoor", 1, 0))
```

Con fines ilustrativos, los ejemplos de esta sección se desarrollan a partir de una submuestra de la encuesta. El siguiente bloque de código construye la base de datos utilizada en los ejemplos del capítulo. Para ello, identifica los tres estratos con menor número de hogares y conserva únicamente las variables requeridas para el análisis: ingreso, gasto, estrato, sexo, región, zona de residencia y condición de pobreza.

```
plot_data <- survey_data %>%
  select(HHID, Stratum) %>%
  distinct() %>%
  group_by(Stratum) %>%
  tally() %>%
  arrange(n) %>%
  select(-n) %>%
  slice(1:3L) %>%
  inner_join(survey_data, by = "Stratum") %>%
  select(Income, Expenditure, Stratum, Sex, Region, Zone, poor)
```

Como punto de partida, se ajusta un modelo de regresión lineal convencional que relaciona el ingreso de los hogares con su nivel de gasto, ignorando temporalmente la estructura jerárquica de los datos. Bajo esta especificación, se asume que todas las observaciones son independientes y que la relación entre ingreso y gasto es homogénea para todos los hogares, independientemente del estrato al que pertenezcan. La figura 7.1 presenta la recta de regresión estimada junto con los datos observados.

```
ggplot(data = plot_data,
       aes(y = Income, x = Expenditure)) +
  geom_jitter() +
  geom_smooth(formula = y ~ x, method = "lm", se = FALSE) +
  ggtitle(latex2exp::TeX(
```

```
    "$Income_{i} \\sim \\hat{\\beta}_{0} + \\hat{\\beta}_{1}Expenditure_{i} + \\epsilon_{i}$"
  )) +
  theme(
    legend.position = "none",
    plot.title = element_text(hjust = 0.5)
  )
```

Figura 7.1.: Modelo de regresión lineal simple para el ingreso en función del gasto (sin considerar estratos)

Como explica Finch, Bolin, y Kelley (2019), si bien este modelo resulta útil como referencia inicial, sus supuestos suelen ser poco realistas en el contexto de encuestas de hogares. En particular, la independencia entre observaciones puede verse comprometida cuando los hogares pertenecen a estratos que comparten características socioeconómicas, demográficas o geográficas similares. Asimismo, es posible que los niveles promedio de ingreso difieran sistemáticamente entre estratos, aun cuando la relación entre ingreso y gasto sea semejante. Ignorar esta estructura jerárquica puede conducir a estimaciones incorrectas de los errores estándar y, en consecuencia, a inferencias estadísticas poco confiables.

Con fines exclusivamente ilustrativos, se incorpora gradualmente la estructura jerárquica de los datos al proceso de modelación. Para ello, se ajusta inicialmente un modelo que permite que cada estrato tenga su propio intercepto, mientras mantiene una pendiente común para todas las observaciones. Esta especificación resulta útil para visualizar cómo los niveles promedio de ingreso pueden diferir entre estratos, aun cuando la relación entre ingreso y gasto se considere constante.

```
beta_1 <- coef(lm(Income ~ Expenditure, data = plot_data))[2]
model_coef <- plot_data %>%
  group_by(Stratum) %>%
  summarise(beta_0 = coef(lm(Income ~ Expenditure))[1]) %>%
  mutate(beta_1 = beta_1)
```

De esta manera, el modelo introduce una primera fuente de heterogeneidad entre grupos y permite apreciar las limitaciones del modelo de regresión simple presentado anteriormente, el cual asumía una única recta de regresión para toda la población. La figura 7.2 presenta las rectas de regresión con interceptos diferenciados por estrato:

```
ggplot(data = plot_data,
       aes(y = Income, x = Expenditure, colour = Stratum)) +
  geom_jitter() +
  geom_abline(data = model_coef,
              mapping = aes(slope = beta_1, intercept = beta_0, colour = Stratum)) +
  ggtitle(latex2exp::TeX(
    "$Income_{kj} \\sim \\hat{\\beta}_{0j} + \\hat{\\beta}_{1}Expenditure_{kj} + \\epsilon_{kj}
  )) +
  theme(
    legend.position = "none",
    plot.title = element_text(hjust = 0.5)
  )
```

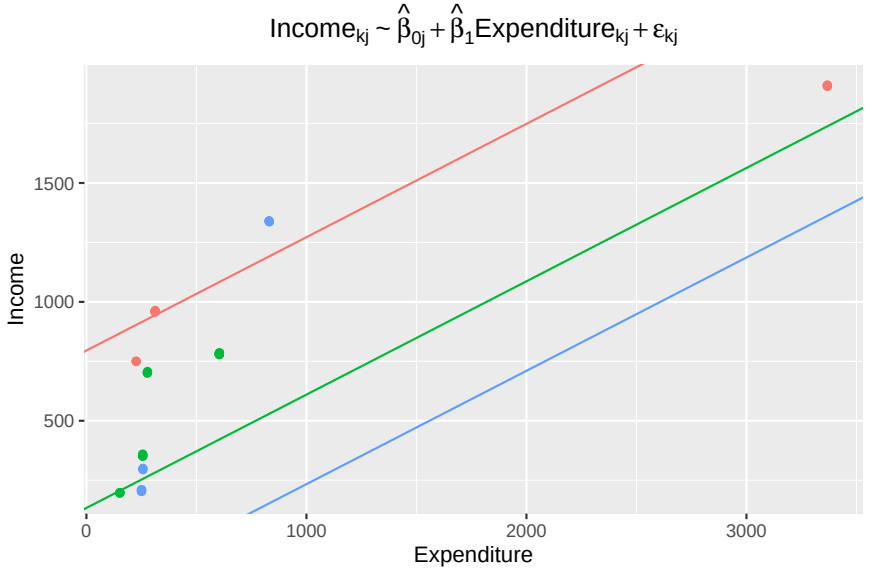

Figura 7.2.: Modelo con intercepto específico por estrato y pendiente común

Como siguiente paso, se considera una especificación alternativa en la que todos los estratos comparten un mismo nivel promedio de ingreso, representado por un intercepto común, pero se permite que la relación entre ingreso y gasto varíe entre estratos mediante pendientes específicas para cada uno de ellos.

```
beta_0 <- coef(lm(Income ~ Expenditure, data = plot_data))[1]
model_coef <- plot_data %>%
  group_by(Stratum) %>%
  summarise(beta_1 = coef(lm(Income ~ Expenditure))[2]) %>%
  mutate(beta_0 = beta_0)
```

Bajo esta especificación, las diferencias entre estratos se manifiestan en la intensidad de la asociación entre ingreso y gasto, más que en sus niveles promedio. La figura 7.3 permite visualizar estas diferencias en las trayectorias de regresión estimadas para cada estrato.

```
ggplot(data = plot_data,
       aes(y = Income, x = Expenditure, colour = Stratum)) +
  geom_jitter() +
  geom_abline(data = model_coef,
              mapping = aes(slope = beta_1, intercept = beta_0, colour = Stratum)) +
  ggtitle(latex2exp::TeX(
    "$Income_{kj} \\sim \\hat{\\beta}_{0} + \\hat{\\beta}_{1j}Expenditure_{kj} + \\epsilon_{kj}
  )) +
  theme(
    legend.position = "none",
    plot.title = element_text(hjust = 0.5)
  )
```

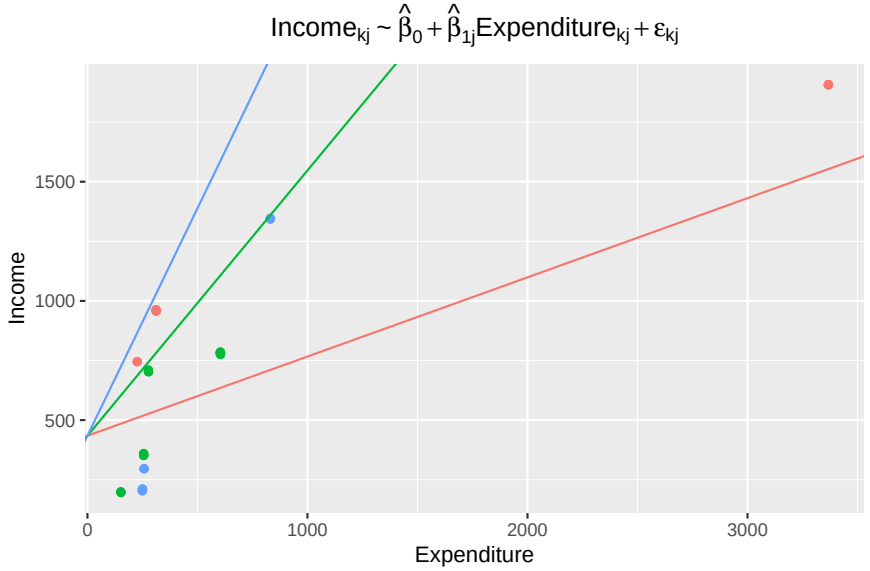

Figura 7.3.: Modelo con pendiente específica por estrato e intercepto común

Finalmente, se considera la especificación más flexible de las presentadas hasta el momento, permitiendo que tanto el nivel promedio de ingreso como la relación entre ingreso y gasto varíen entre estratos. Bajo este enfoque, cada estrato posee su propia recta de regresión, con interceptos y pendientes estimados de manera independiente.

```
model_coef <- plot_data %>%
  group_by(Stratum) %>%
  summarise(
    beta_0 = coef(lm(Income ~ Expenditure))[1],
    beta_1 = coef(lm(Income ~ Expenditure))[2]
  )
```

Esta formulación permite capturar simultáneamente diferencias en los niveles de ingreso y en la intensidad de su asociación con el gasto. La figura 7.4 muestra las rectas ajustadas para cada estrato bajo esta especificación.

```
ggplot(data = plot_data,
       aes(y = Income, x = Expenditure, colour = Stratum)) +
  geom_jitter() +
  geom_abline(data = model_coef,
              mapping = aes(slope = beta_1, intercept = beta_0, colour = Stratum)) +
  ggtitle(latex2exp::TeX(
    "$Income_{kj} \\sim \\hat{\\beta}_{0j} + \\hat{\\beta}_{1j}Expenditure_{kj} + \\epsilon_{kj
  )) +
  theme(
    legend.position = "none",
    plot.title = element_text(hjust = 0.5)
  )
```

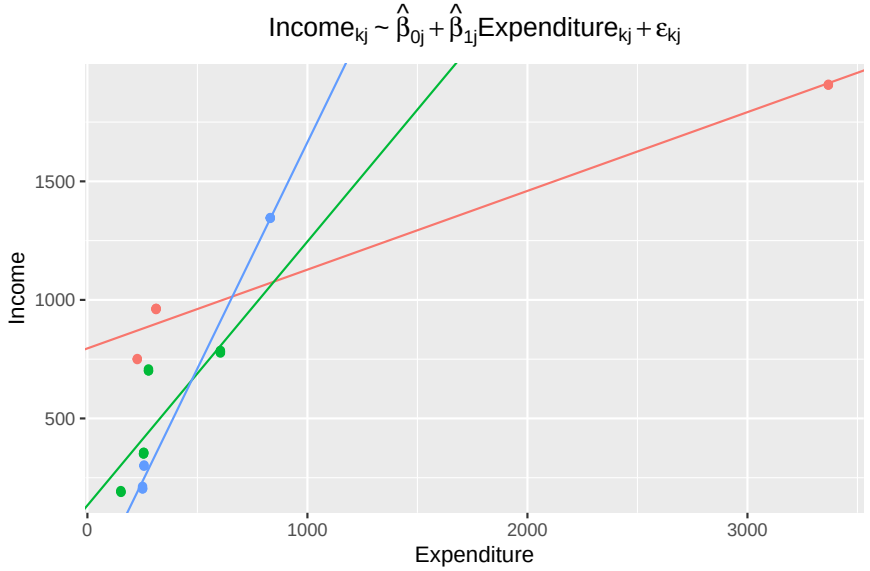

Figura 7.4.: Modelo con intercepto y pendiente específicos por estrato (ajuste independiente)

Aunque esta última especificación ofrece una representación más flexible de los datos y permite capturar las particularidades de cada estrato, presenta una limitación importante: los parámetros se estiman de manera completamente independiente para cada grupo. Como consecuencia, no se aprovecha la información compartida entre estratos que podrían presentar patrones similares, lo que puede generar estimaciones inestables, especialmente en aquellos grupos con un número reducido de observaciones.

Según Gelman y Hill (2006) y Goldstein (2011), los modelos multinivel superan esta dificultad mediante un mecanismo conocido como encogimiento (*shrinkage*), en el cual las estimaciones específicas de cada grupo se obtienen combinando la información propia del grupo con la información proveniente del conjunto de la población. De esta manera, se logra un equilibrio entre la representación de las diferencias entre estratos y la eficiencia estadística de las estimaciones.

7.2. Pesos *q-weighted*

De acuerdo con Pfeffermann et al. (1998), Asparouhov (2006), Rabe-Hesketh y Skrondal (2006), Carle (2009) y Cai (2013), en relación con la incorporación de pesos de muestreo en modelos multinivel, existen enfoques basados en pseudo-máxima verosimilitud, en particular mediante procedimientos de mínimos cuadrados generalizados ponderados, así como métodos basados en algoritmos EM para la estimación de modelos jerárquicos con ponderaciones. La incorporación de los pesos resulta importante cuando el muestreo es informativo, aunque no existe una única estrategia metodológica para su implementación.

Según Kreft y De Leeuw (1998) y Bates et al. (2015), desde una perspectiva general, el enfoque de máxima verosimilitud consiste en estimar los parámetros del modelo identificando aquellos valores que maximizan la probabilidad de observar los datos disponibles bajo la especificación asumida. Una extensión importante de este enfoque es la máxima verosimilitud restringida (REML), la cual mejora la estimación de los componentes de varianza al considerar la pérdida de grados de libertad debida a la estimación de los efectos fijos, produciendo en muchos casos estimaciones menos sesgadas que la máxima verosimilitud convencional.

En el contexto de los modelos multinivel con datos provenientes de encuestas, una dificultad adicional radica en la incorporación coherente de los pesos de muestreo a través de los distintos niveles de la jerarquía. La estructura agrupada de los datos implica que las observaciones no son independientes, por lo que la función de log-verosimilitud no puede descomponerse como una suma simple de contribuciones individuales. En su lugar, es necesario considerar explícitamente la dependencia entre los distintos niveles del diseño muestral para lograr inferencias adecuadas.

Como explica Pfeffermann (2011), para ajustar modelos multinivel con datos provenientes de encuestas complejas, una alternativa consiste en utilizar el enfoque de pesos *q-weighted*. La idea central de este método es eliminar de los pesos muestrales la componente sistemática explicada por las covariables incluidas en el modelo, de modo que los pesos resultantes conserven únicamente la parte residual atribuible al mecanismo de selección.

No obstante, este no es el único enfoque disponible para la estimación de modelos multinivel con datos de encuestas complejas y tampoco sustituye el uso de pesos condicionales o de pesos escalados en cada etapa del diseño, requeridos por algunos métodos de pseudo-verosimilitud multinivel. El procedimiento para construir los pesos *q-weighted* se resume en los siguientes pasos:

1. Ajustar un modelo de regresión para los factores de expansión finales de la encuesta, utilizando como variables explicativas el mismo conjunto de covariables que será empleado en el modelo multinivel. Siendo w_k el factor de expansión original de la unidad k y $\mathbf{x}_k$ el vector de covariables. Entonces se ajusta el modelo

$$w_k = f(\mathbf{x}_k, \beta) + \varepsilon_k,$$

   donde $f(\cdot)$ representa la relación funcional entre los pesos w_k y las covariables $\mathbf{x}_k$, a través de los coeficientes β.

2. Obtener los valores predichos del modelo para cada unidad de observación, de tal forma que:

$$\hat{w}_k = f(\mathbf{x}_k, \hat{\beta}).$$

Estos valores representan la componente de los factores de expansión explicada por las covariables incluidas en el modelo.

3. Construir los pesos *q-weighted* dividiendo los factores de expansión originales por la estimación de su esperanza condicional en la muestra,

$$q_k = \frac{w_k}{\widehat{E}_s(w_k \mid \mathbf{x}_k)}.$$

De esta forma, los nuevos pesos reflejan la variación residual de los factores de expansión una vez controlado el efecto de las covariables.

4. Definir el nuevo diseño muestral utilizando los pesos ajustados q_k en lugar de los pesos originales w_k, y emplear este diseño para la estimación de los parámetros del modelo multinivel.

El siguiente código implementa el procedimiento de construcción de pesos *q-weighted*. En primer lugar, ajusta un modelo de regresión lineal que explica los factores de expansión originales a partir de la variable explicativa gasto. De esta forma, los pesos resultantes conservan únicamente la componente de variación no explicada por las covariables incluidas en el modelo multinivel.

```
mod_qw <- lm(wk ~ Expenditure, data = survey_data)
predicted_weights <- predict(mod_qw)
stopifnot(all(is.finite(predicted_weights)), all(predicted_weights > 0))
survey_data$qk <- survey_data$wk / predicted_weights
```

Esta implementación es ilustrativa. La comprobación con `stopifnot()` evita continuar si los valores predichos no son finitos o estrictamente positivos, pero todavía debe evaluarse que el modelo represente adecuadamente $E_s(w_k \mid \mathbf{x}_k)$; una regresión lineal de los pesos puede producir predicciones no válidas. Además, el vector final `wk` no necesariamente permite reconstruir por sí solo los pesos condicionales de cada etapa requeridos por un modelo multinivel ponderado.

7.3. Modelo lineal multinivel

Los modelos multinivel más utilizados en la práctica corresponden a extensiones del modelo lineal clásico para datos con estructura jerárquica. En su formulación básica, estos modelos asumen que la variable respuesta es continua y sigue una distribución normal condicional a los efectos fijos y aleatorios incluidos en el modelo. Asimismo, se supone que los efectos aleatorios y los errores residuales son independientes entre sí y se distribuyen normalmente con media cero y varianzas constantes.

7.3.1. Modelo nulo

De acuerdo con Goldstein (2011) y Rabe-Hesketh y Skrondal (2012), en el análisis de regresión multinivel, se distinguen dos tipos de parámetros: los coeficientes de regresión, conocidos como parámetros fijos, y los componentes de varianza asociados a los efectos aleatorios. Una etapa inicial fundamental en este tipo de análisis consiste en descomponer la variabilidad total de la variable

dependiente en los distintos niveles de la estructura jerárquica. En el contexto del ejemplo anterior, esta descomposición permite separar la variación del ingreso en una componente atribuible a diferencias dentro de los estratos y otra asociada a diferencias entre estratos.

Según Snijders y Bosker (2011), el punto de partida para esta descomposición es, el denominado modelo nulo, el cual no incorpora variables explicativas y se expresa como

$$y_{kj} = \beta_{0j} + \epsilon_{kj}$$

En donde y_{kj} denota el valor observado del ingreso para la unidad k en el estrato j. El término β_{0j} representa el intercepto específico de cada estrato, capturando las diferencias en los niveles promedio de la variable entre grupos. Finalmente, ϵ_{kj} corresponde al error a nivel de unidad, el cual recoge la variabilidad no explicada dentro de cada estrato y se asume con media cero y varianza constante condicional al grupo.

A su vez, el intercepto por estrato puede descomponerse en una parte común a todos los estratos y una desviación específica de cada uno de ellos, de la siguiente forma:

$$\beta_{0j} = \gamma_{00} + \tau_{0j}$$

donde γ_{00} representa el intercepto promedio global y τ_{0j} es el efecto aleatorio para el intercepto que captura la desviación del estrato j, permitiendo modelar explícitamente la heterogeneidad entre estratos. Además, los componentes aleatorios del modelo se asumen distribuidos normalmente con media cero y varianzas específicas para cada nivel de la jerarquía. En particular, el efecto aleatorio asociado al intercepto por estrato cumple

$$\tau_{0j} \sim N\left(0, \sigma_\tau^2\right),$$

lo que implica que las desviaciones de cada estrato respecto al intercepto global se distribuyen alrededor de cero, con una variabilidad determinada por σ_τ^2. De manera análoga, el término de error a nivel de unidad sigue la distribución

$$\epsilon_{kj} \sim N\left(0, \sigma_\epsilon^2\right),$$

Donde σ_ϵ^2 representa la variabilidad residual dentro de los estratos. A partir de este modelo se define el coeficiente de correlación intraclase (ICC, por sus siglas en inglés), que cuantifica la proporción de la varianza total explicada por las diferencias entre estratos:

$$\rho = \frac{\sigma_\tau^2}{\sigma_\tau^2 + \sigma_\epsilon^2}$$

Un valor elevado del ICC indica que una proporción importante de la variabilidad total de la variable de interés se debe a diferencias entre estratos, lo que sugiere la presencia de heterogeneidad relevante entre grupos y la necesidad de incorporar explícitamente dicha estructura en el modelo. En contraste, un ICC bajo implica que la variabilidad se concentra principalmente dentro de los estratos, los cuales presentan una mayor homogeneidad relativa entre sí.

Siguiendo con el análisis de la encuesta de ejemplo, cuando la variable de interés es el ingreso, el modelo nulo constituye el punto de partida para cuantificar qué proporción de la variabilidad total se debe a diferencias entre estratos y qué proporción corresponde a diferencias entre hogares dentro de un mismo estrato.

De acuerdo con Bates et al. (2015), la estimación de modelos multinivel en `R` se realiza mediante el paquete `lme4`. En particular, la función `lmer()` permite especificar los efectos aleatorios mediante expresiones de la forma `(efecto | grupo)`. En este contexto, la expresión `(1 | Stratum)` indica que el modelo incluye un intercepto aleatorio para cada estrato. El valor `1` representa el intercepto del modelo, mientras que `Stratum` identifica la variable de agrupación.

Como advierte Carle (2009), en `lme4`, el argumento `weights` representa pesos previos que modifican la contribución residual, pero no implementa por sí solo la pseudo-verosimilitud multinivel con pesos de diseño por etapa ni la estimación de varianza del diseño. Por tanto, los modelos ponderados con `qk` que siguen deben interpretarse como una ilustración modelada, no como una solución general para inferencia de encuestas complejas; para ese propósito se requieren pesos escalados y software que los incorpore expresamente en la estimación multinivel.

```
mod_null <- lmer(
  Income ~ (1 | Stratum),
  data = survey_data,
  weights = qk
)
```

La tabla 7.1 presenta los interceptos estimados para cada estrato a partir del modelo nulo. Dado que este modelo no incorpora variables explicativas, cada intercepto puede interpretarse como el ingreso promedio esperado en el correspondiente estrato. Los resultados evidencian diferencias importantes entre grupos: mientras que el estrato `idStrt004` presenta el mayor ingreso promedio estimado (959.6), el estrato `idStrt009` registra el menor valor (207.6). Estas diferencias sugieren la existencia de una variabilidad sustancial entre estratos, justificando la incorporación de efectos aleatorios para modelar explícitamente la estructura jerárquica de los datos.

```
coef_mod_null <- coef(mod_null)$Stratum
coef_mod_null %>%
  slice(1:12L)
```

```
coef_mod_null <- coef(mod_null)$Stratum
coef_mod_null %>%
  slice(1:12L) %>%
  tabla_fmt()
```

Tabla 7.1.: Interceptos estimados por estrato bajo pesos q-weighted (modelo nulo, primeros 12 estratos)

	(Intercept)
idStrt001	635
idStrt002	507

Tabla 7.1.: Interceptos estimados por estrato bajo pesos q-weighted (modelo nulo, primeros 12 estratos)

	(Intercept)
idStrt003	486
idStrt004	960
idStrt005	518
idStrt006	439
idStrt007	477
idStrt008	377
idStrt009	218
idStrt010	594
idStrt011	590
idStrt012	362

De acuerdo con Lüdecke et al. (2026), a partir de los componentes de varianza estimados en el modelo nulo, la correlación intraclase se calcula mediante la función `icc()` del paquete `performance`. Este indicador cuantifica la proporción de la variabilidad total del ingreso que puede atribuirse a diferencias entre estratos, constituyendo una medida de la dependencia existente entre observaciones pertenecientes a un mismo grupo. Los resultados obtenidos se presentan en la tabla 7.2.

```
performance::icc(mod_null) %>%
  as.data.frame()
```

```
performance::icc(mod_null) %>%
  as.data.frame() %>%
  tabla_fmt()
```

Tabla 7.2.: Correlación intraclase del modelo nulo (ingreso por estrato)

ICC_adjusted	ICC_unadjusted	optional
0.329	0.329	FALSE

La correlación intraclase estimada de 32 % indica que aproximadamente un tercio de la variabilidad total observada en el ingreso se debe a diferencias entre estratos, mientras que el porcentaje restante corresponde a diferencias entre hogares dentro de los mismos estratos. Por último, como el modelo nulo no incluye ningún predictor, la estimación del ingreso dentro de cada estrato es constante e igual al intercepto estimado para ese estrato.

7.3.2. Modelo con intercepto aleatorio

De acuerdo con Goldstein (2011) y Rabe-Hesketh y Skrondal (2012), el modelo multinivel más simple que incorpora covariables corresponde al modelo de intercepto aleatorio. En esta especificación, se

asume que los efectos de las variables explicativas son comunes a todos los estratos, mientras que el nivel promedio de la variable respuesta puede variar entre ellos. El modelo se expresa como

$$y_{kj} = \beta_{0j} + \mathbf{x}_{kj}\beta + \epsilon_{kj}$$

En donde y_{kj} representa el valor observado de la variable respuesta para la unidad k perteneciente al estrato j, $\mathbf{x}_{kj}$ es el vector de covariables asociado a dicha unidad, β es el vector de coeficientes de regresión comunes a todos los estratos y ϵ_{kj} corresponde al error aleatorio a nivel de unidad.

Al igual que en el modelo nulo, el intercepto se modela como $\beta_{0j} = \gamma_{00} + \tau_{0j}$; en donde γ_{00} representa el intercepto promedio global y τ_{0j} es el efecto aleatorio asociado al estrato j, que mide la desviación de dicho estrato respecto al promedio general.

Los componentes aleatorios del modelo se asumen independientes y normalmente distribuidos, de manera que $\tau_{0j} \sim N(0, \sigma^2_\tau)$ y $\epsilon_{kj} \sim N(0, \sigma^2_\epsilon)$. Bajo estas condiciones, la variabilidad total de la variable respuesta puede descomponerse en una componente entre estratos, cuantificada por σ^2_τ, y una componente dentro de los estratos, representada por σ^2_ϵ.

El siguiente código ajusta un modelo multinivel de intercepto aleatorio, en donde el ingreso (`Income`) se explica a partir del gasto del hogar (`Expenditure`), incorporando además un efecto aleatorio asociado al estrato (`Stratum`) y utilizando los pesos *q-weighted* almacenados en la variable `qk`. Una vez ajustado el modelo, se estima la correlación intraclase (ICC) a partir de los componentes de varianza estimados.

```
random_intercept_model <- lmer(
  Income ~ Expenditure + (1 | Stratum),
  data = survey_data,
  weights = qk
)
performance::icc(random_intercept_model)
```

```
# Intraclass Correlation Coefficient

    Adjusted ICC: 0.203
  Unadjusted ICC: 0.109
```

Según Snijders y Bosker (2011), la correlación intraclase ajustada de 0.196 indica que, después de controlar por el gasto del hogar (`Expenditure`), aproximadamente el 20 % de la variabilidad residual del ingreso sigue siendo atribuible a diferencias entre estratos. Aunque este valor evidencia agrupamiento residual, el ICC por sí solo no justifica una especificación multinivel concreta; también deben considerarse la definición sustantiva de los grupos, el ajuste y los diagnósticos del modelo. Los coeficientes estimados por estrato se presentan en la tabla 7.3:

```
coef(random_intercept_model)$Stratum %>%
  slice(1:8L)
```

```
coef(random_intercept_model)$Stratum %>%
  slice(1:8L) %>%
  tabla_fmt()
```

Tabla 7.3.: Coeficientes del modelo con intercepto aleatorio por estrato (primeros 8 estratos)

	(Intercept)	Expenditure
idStrt001	250.4	1.19
idStrt002	156.8	1.19
idStrt003	142.9	1.19
idStrt004	296.8	1.19
idStrt005	-32.4	1.19
idStrt006	53.8	1.19
idStrt007	12.5	1.19
idStrt008	107.1	1.19

Con el fin de facilitar la visualización de los resultados del modelo, se utiliza la submuestra reducida presentada anteriormente, la cual contiene únicamente tres estratos seleccionados.

```
estimated_coef <- inner_join(
  coef(random_intercept_model)$Stratum %>%
    tibble::rownames_to_column(var = "Stratum"),
  plot_data %>%
    select(Stratum) %>%
    distinct()
)
```

La figura 7.5 muestra las rectas de regresión estimadas para cada estrato bajo el modelo de intercepto aleatorio. Se observa que las rectas comparten una misma pendiente, reflejando el efecto común del gasto sobre el ingreso, pero difieren en su posición vertical debido a los interceptos específicos estimados para cada estrato.

```
ggplot(data = plot_data,
       aes(y = Income, x = Expenditure, colour = Stratum)) +
  geom_jitter() +
  geom_abline(data = estimated_coef,
              mapping = aes(slope = Expenditure,
                            intercept = `(Intercept)`,
                            colour = Stratum)) +
  theme(
    legend.position = "none",
    plot.title = element_text(hjust = 0.5)
  )
```

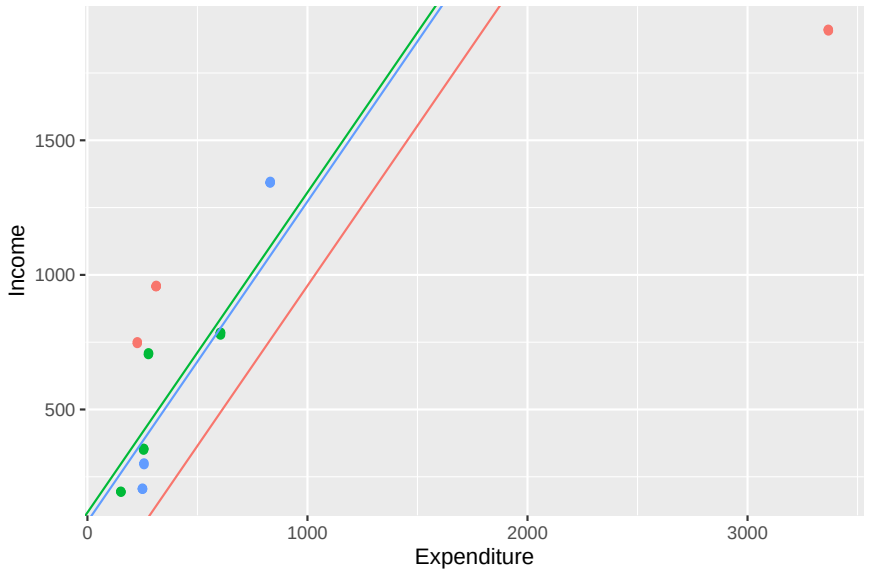

Figura 7.5.: Rectas de regresión estimadas por estrato: modelo con intercepto aleatorio

7.3.3. Modelo con intercepto y pendiente aleatoria

De acuerdo con Goldstein (2011) y Snijders y Bosker (2011), una extensión natural del modelo anterior es el modelo de intercepto y pendiente aleatorios. En esta especificación, no sólo se permite que el nivel promedio de la variable respuesta varíe entre estratos, sino también que el efecto de una o más covariables difiera entre grupos. De esta manera, cada estrato puede tener su propia recta de regresión, tanto en términos de intercepto como de pendiente. El modelo puede expresarse como

$$y_{kj} = \beta_{0j} + x_{kj}\beta_{1j} + \mathbf{z}_{kj}\beta + \epsilon_{kj},$$

donde y_{kj} representa el valor observado de la variable respuesta para la unidad k perteneciente al estrato j, x_{kj} corresponde a la covariable cuya pendiente se permite variar entre estratos, $\mathbf{z}_{kj}$ contiene las covariables con efectos fijos comunes a todos los grupos, y ϵ_{kj} es el término de error a nivel de unidad. En este modelo, tanto el intercepto como la pendiente asociada a x_{kj} se consideran aleatorios y se descomponen de la siguiente forma:

$$\beta_{0j} = \gamma_{00} + \tau_{0j}, \qquad \beta_{1j} = \gamma_{10} + \tau_{1j}$$

donde γ_{00} y γ_{10} representan, respectivamente, el intercepto y la pendiente promedio en la población, mientras que τ_{0j} y τ_{1j} capturan las desviaciones específicas del estrato j respecto a dichos promedios. Estos efectos aleatorios se asumen normalmente distribuidos como se indica a continuación:

$$\begin{pmatrix} \tau_{0j} \\ \tau_{1j} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2_{\tau_0} & \sigma_{\tau_{01}} \\ \sigma_{\tau_{01}} & \sigma^2_{\tau_1} \end{pmatrix} \right),$$

mientras que el error a nivel de unidad satisface $\epsilon_{kj} \sim N(0, \sigma^2_\epsilon)$. En consecuencia, el modelo permite cuantificar no sólo la variabilidad entre estratos en los niveles promedio de la variable respuesta,

sino también la heterogeneidad existente en el efecto de la covariable cuyo coeficiente se modela como aleatorio.

El siguiente código ajusta un modelo multinivel con intercepto y pendiente aleatorios en el que el ingreso (`Income`) se explica a partir del gasto del hogar (`Expenditure`); aunque la redacción anterior mencionaba `Zone`, esa variable no aparece en la fórmula estimada. En este caso, tanto el intercepto como el efecto del gasto pueden variar entre estratos mediante efectos aleatorios asociados a `Stratum`, empleando además los pesos *q-weighted* almacenados en la variable `qk`.

```
random_slope_model <- lmer(
  Income ~ 1 + Expenditure + (1 + Expenditure | Stratum),
  data = survey_data,
  weights = qk
)

performance::icc(random_slope_model)
```

```
# Intraclass Correlation Coefficient

    Adjusted ICC: 0.690
  Unadjusted ICC: 0.457
```

Nótese que en la especificación del código se incluye simultáneamente efectos fijos y efectos aleatorios para el intercepto y la variable `Expenditure`. Los términos ubicados fuera del paréntesis, `1 + Expenditure`, corresponden a los efectos fijos del modelo y permiten estimar el intercepto promedio global y la pendiente promedio global asociada al gasto. Por su parte, la expresión `(1 + Expenditure | Stratum)` incorpora las desviaciones específicas de cada estrato respecto a dichos promedios. La inclusión explícita de los efectos fijos es fundamental, ya que los efectos aleatorios se interpretan como desviaciones alrededor de una media poblacional. Los coeficientes del modelo por estrato se presentan en la tabla 7.4:

Como recomiendan Bates et al. (2015), los modelos con pendientes aleatorias requieren suficientes grupos y variación interna para estimar de manera estable la varianza y la covarianza de los efectos aleatorios; por ello, conviene revisar advertencias de convergencia y ajustes singulares antes de interpretar estos componentes.

```
coef(random_slope_model)$Stratum %>%
  slice(1:10L)
```

```
coef(random_slope_model)$Stratum %>%
  slice(1:10L) %>%
  tabla_fmt()
```

Tabla 7.4.: Coeficientes del modelo con intercepto y pendiente aleatoria por estrato (primeros 10 estratos)

	(Intercept)	Expenditure
idStrt001	-222.8	2.730
idStrt002	35.5	1.607
idStrt003	151.9	1.164
idStrt004	224.2	1.353
idStrt005	-89.7	1.286
idStrt006	28.6	1.217
idStrt007	41.0	1.079
idStrt008	163.8	0.928
idStrt009	16.6	0.838
idStrt010	89.9	1.830

Una vez más, con el propósito de facilitar la visualización de los resultados, se utiliza la submuestra reducida definida previamente, la cual considera únicamente tres estratos seleccionados.

```
estimated_coef <- inner_join(
  coef(random_slope_model)$Stratum %>%
    tibble::rownames_to_column(var = "Stratum"),
  plot_data %>%
    select(Stratum) %>%
    distinct()
)
```

La figura 7.6 presenta las rectas de regresión estimadas para cada estrato bajo el modelo de intercepto y pendiente aleatorios. A diferencia del modelo de intercepto aleatorio, en este caso las rectas difieren tanto en su posición vertical como en su inclinación, reflejando que los estratos no sólo presentan niveles promedio de ingreso distintos, sino también diferentes intensidades en la relación entre ingreso y gasto.

```
ggplot(data = plot_data,
       aes(y = Income, x = Expenditure, colour = Stratum)) +
  geom_jitter() +
  geom_abline(data = estimated_coef,
              mapping = aes(slope = Expenditure,
                            intercept = `(Intercept)`,
                            colour = Stratum)) +
  theme(
    legend.position = "none",
    plot.title = element_text(hjust = 0.5)
  )
```

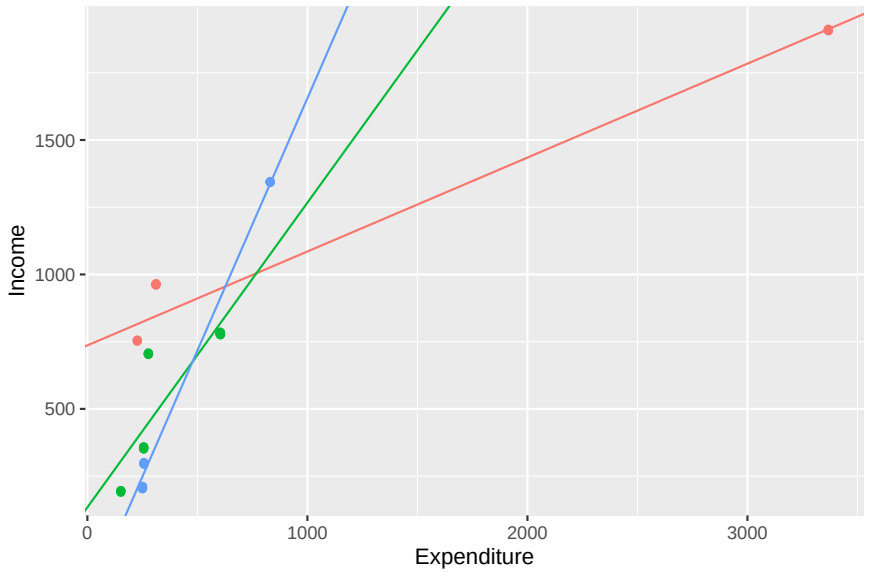

Figura 7.6.: Rectas de regresión estimadas por estrato: modelo con intercepto y pendiente aleatoria

7.4. Modelo logístico multinivel

De acuerdo con Snijders y Bosker (2011) y Rabe-Hesketh y Skrondal (2012), los modelos logísticos multinivel constituyen una extensión de los modelos multinivel para variables continuas al caso en que la variable respuesta es dicotómica. En lugar de modelar directamente el valor esperado de una variable continua, estos modelos estiman la probabilidad de ocurrencia de un evento, incorporando simultáneamente covariables individuales y efectos aleatorios asociados a los grupos de pertenencia. En el contexto de encuestas de hogares, esta formulación es especialmente útil cuando se analizan resultados binarios, como encontrarse o no en condición de pobreza, acceder o no a un servicio, participar o no en el mercado laboral, o presentar determinada característica sociodemográfica.

Al igual que en los modelos multinivel para variables continuas, el punto de partida consiste en reconocer que las unidades de observación no son independientes cuando pertenecen a un mismo grupo sustantivo. Como se advirtió al inicio, un estrato, conglomerado o dominio del diseño solo debe emplearse como nivel aleatorio cuando corresponda al contexto que se desea modelar y el supuesto de efectos intercambiables sea defendible. Sin embargo, en el caso logístico la respuesta no se representa mediante una distribución normal con un error residual aditivo, sino mediante una distribución Bernoulli condicional a la probabilidad de ocurrencia del evento. Esta probabilidad se vincula con los predictores a través de la función logit, lo que permite expresar el modelo en una escala lineal.

$$y_{kj} \mid \pi_{kj} \sim \text{Bernoulli}(\pi_{kj})$$

donde y_{kj} toma el valor uno si la unidad k del estrato j presenta el evento de interés y cero en caso contrario. La probabilidad condicional del evento se denota por $\pi_{kj} = Pr(y_{kj} = 1)$ y se relaciona con el predictor lineal mediante

$$\text{logit}(\pi_{kj}) = \log\left(\frac{\pi_{kj}}{1-\pi_{kj}}\right) = \eta_{kj}$$

En esta formulación, η_{kj} cumple un papel análogo al valor esperado lineal de los modelos para variables continuas. La diferencia central es que los efectos fijos y aleatorios actúan sobre la escala logit y no directamente sobre la probabilidad. Por tanto, las diferencias entre estratos se interpretan como desplazamientos en los *log-odds* del evento, los cuales luego se transforman a probabilidades mediante la función logística. La figura 7.7 presenta la relación entre gasto y probabilidad de pobreza, con una curva logística ajustada.

```
ggplot(data = survey_data,
       aes(y = poor, x = Expenditure)) +
  geom_point(alpha = 0.5) +
  geom_smooth(
    formula = y ~ x,
    method = "glm",
    se = FALSE,
    method.args = list(family = binomial(link = "logit"))
  ) +
  labs(y = "Poverty (1 = poor)", x = "Expenditure")
```

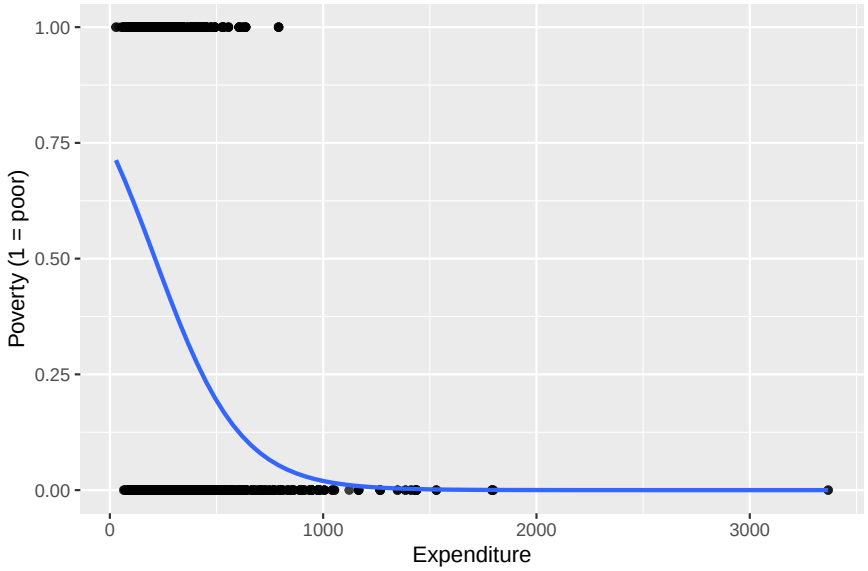

Figura 7.7.: Relación entre gasto y condición de pobreza con curva logística ajustada

La curva ajustada resume la asociación marginal entre el gasto y la condición de pobreza, sin incorporar todavía la estructura jerárquica de los estratos. En términos generales, la forma descendente de la curva indica que, a medida que aumenta el gasto del hogar, disminuye la probabilidad estimada de encontrarse en condición de pobreza. Sin embargo, esta relación promedio puede ocultar diferencias relevantes entre estratos, especialmente cuando los hogares pertenecen a contextos socioeconómicos distintos.

7.4.1. Modelo nulo logístico

Según Snijders y Bosker (2011), como en el caso de variables continuas, el modelo nulo logístico constituye el punto de partida para estudiar la estructura jerárquica de la variable respuesta. Este modelo no incorpora covariables y permite evaluar si la probabilidad promedio del evento varía entre estratos. En el ejemplo considerado, el evento de interés corresponde a la condición de pobreza, de modo que el modelo permite separar la heterogeneidad atribuible a diferencias entre estratos de la variación individual propia de una respuesta binaria.

En el modelo nulo, la variable de interés sigue la siguiente distribución Bernoulli:

$$y_{kj} \mid \pi_{kj} \sim \text{Bernoulli}(\pi_{kj})$$

En donde la probabilidad de éxito para la unidad k perteneciente al grupo j se modela mediante

$$\text{logit}(\pi_{kj}) = \beta_{0j},$$

donde β_{0j} representa el intercepto específico del estrato j en escala logit. A su vez, dicho intercepto se descompone como

$$\beta_{0j} = \gamma_{00} + \tau_{0j}$$

siendo γ_{00} el intercepto promedio global de la población y τ_{0j} el efecto aleatorio asociado al estrato j, el cual captura las desviaciones de cada grupo respecto al promedio general. Al igual que en el modelo nulo para variables continuas, este efecto aleatorio permite modelar explícitamente la heterogeneidad entre grupos:

$$\tau_{0j} \sim N(0, \sigma^2_\tau)$$

De acuerdo con Merlo et al. (2006), a diferencia del modelo lineal multinivel, aquí no se incorpora un término residual aditivo ϵ_{kj} en la ecuación del nivel individual. La variabilidad interna en el grupo está determinada por la distribución Bernoulli de la respuesta. Para cuantificar la dependencia entre unidades pertenecientes al mismo estrato, puede utilizarse el enfoque de variable latente, bajo el cual esta varianza residual se aproxima por la varianza de la distribución logística estándar, es decir, $\pi^2/3$. Así, la correlación intraclase del modelo logístico se expresa como

$$\rho = \frac{\sigma^2_\tau}{\sigma^2_\tau + \frac{\pi^2}{3}}$$

De acuerdo con Bates et al. (2015), un valor elevado de ρ indica que la probabilidad del evento presenta una estructura de agrupamiento importante, es decir, que dos unidades pertenecientes al mismo estrato tienden a parecerse más entre sí que dos unidades tomadas de estratos distintos. En contraste, un valor bajo sugiere que la mayor parte de la variación se concentra a nivel individual. Este ICC es una medida en escala latente y no equivale directamente a una correlación de respuestas binarias observadas. El ajuste en `R`, mediante `glmer()`, se realiza de la siguiente manera:

```
logistic_null_model <- glmer(
  poor ~ (1 | Stratum),
  data = survey_data,
  weights = qk,
  family = binomial(link = "logit")
)
```

Los coeficientes del modelo nulo logístico por estrato se presentan en la tabla 7.5. Los interceptos estimados en el modelo nulo muestran una heterogeneidad considerable entre estratos. En los primeros estratos reportados, algunos interceptos son fuertemente negativos, como `idStrt004` e `idStrt003`, lo que corresponde a probabilidades base muy bajas de pobreza. En contraste, estratos como `idStrt009` e `idStrt006` presentan interceptos positivos y elevados, asociados con probabilidades base mucho mayores.

```
coef(logistic_null_model)$Stratum %>%
  slice(1:12L)
```

```
coef(logistic_null_model)$Stratum %>%
  slice(1:12L) %>%
  tabla_fmt()
```

Tabla 7.5.: Interceptos estimados por estrato en el modelo logístico nulo (primeros 12 estratos)

	(Intercept)
idStrt001	-0.852
idStrt002	-0.038
idStrt003	-2.378
idStrt004	-2.618
idStrt005	-1.037
idStrt006	0.902
idStrt007	-1.018
idStrt008	0.156
idStrt009	1.913
idStrt010	-0.609
idStrt011	-1.275
idStrt012	0.203

La correlación intraclase asciende a 0.315. Dado que el modelo no incluye covariables, estas diferencias reflejan exclusivamente la variación entre estratos en la prevalencia del evento.

```
performance::icc(logistic_null_model)
```

```
# Intraclass Correlation Coefficient

    Adjusted ICC: 0.315
  Unadjusted ICC: 0.315
```

7.4.2. Modelo logístico con intercepto aleatorio

De acuerdo con Snijders y Bosker (2011) y Rabe-Hesketh y Skrondal (2012), el modelo logístico con intercepto aleatorio incorpora covariables individuales o del hogar, permitiendo que el nivel base del evento varíe entre estratos. Su lógica es paralela a la del modelo con intercepto aleatorio para variables continuas: los efectos de las covariables se consideran comunes a todos los grupos, mientras que cada estrato puede tener su propio intercepto. En este caso, las diferencias entre interceptos se interpretan como diferencias en los *log-odds* base del evento.

El modelo se expresa como $y_{kj} \mid \pi_{kj} \sim \text{Bernoulli}(\pi_{kj})$. En donde $\text{logit}(\pi_{kj}) = \beta_{0j} + \mathbf{x}_{kj}\beta$ y $\beta_{0j} = \gamma_{00} + \tau_{0j}$. En este caso, $\mathbf{x}_{kj}$ representa el vector de covariables de la unidad k en el estrato j, β es el vector de efectos fijos comunes a todos los estratos y τ_{0j} captura la desviación específica del estrato respecto al intercepto promedio global. Como antes, se asume que $\tau_{0j} \sim N(0, \sigma^2_\tau)$.

Bajo esta especificación, dos estratos con los mismos valores de las covariables pueden presentar probabilidades base distintas del evento debido a sus interceptos aleatorios. Sin embargo, el efecto de cada covariable sobre la escala logit se mantiene constante entre estratos. En el ejemplo, esto equivale a permitir que la probabilidad base de pobreza cambie entre estratos, mientras que la asociación entre gasto y pobreza se resume mediante una pendiente común.

```
random_intercept_logit_model <- glmer(
  poor ~ Expenditure + (1 | Stratum),
  data = survey_data,
  family = binomial(link = "logit"),
  weights = qk
)
performance::icc(random_intercept_logit_model)
```

```
# Intraclass Correlation Coefficient

    Adjusted ICC: 0.298
  Unadjusted ICC: 0.171
```

Los coeficientes estimados por estrato se recogen en la tabla 7.6. Al incorporar el gasto del hogar como covariable, el coeficiente fijo asociado a `Expenditure` es negativo, lo que indica que mayores niveles de gasto se asocian con menores log-odds de pobreza. Esto implica una reducción progresiva de la probabilidad estimada de pobreza a medida que aumenta el gasto, manteniendo constante el efecto del estrato. En este modelo, el efecto del gasto es común, pero cada estrato tiene un nivel base propio de pobreza. Así, los estratos con interceptos positivos elevados presentan una mayor probabilidad base de pobreza para un mismo nivel de gasto, mientras que los estratos con interceptos negativos muestran una probabilidad base menor.

Aun después de controlar por el gasto, la correlación intraclase ajustada permanece alta, alrededor de 0.298, lo que muestra que las diferencias entre estratos continúan siendo relevantes.

```
coef(random_intercept_logit_model)$Stratum %>%
  slice(1:10L)
```

```
coef(random_intercept_logit_model)$Stratum %>%
  slice(1:10L) %>%
  tabla_fmt()
```

Tabla 7.6.: Coeficientes del modelo logístico con intercepto aleatorio por estrato (primeros 10 estratos)

	(Intercept)	Expenditure
idStrt001	1.044	-0.007
idStrt002	1.918	-0.007
idStrt003	-0.454	-0.007
idStrt004	0.062	-0.007
idStrt005	1.760	-0.007
idStrt006	3.194	-0.007
idStrt007	0.658	-0.007
idStrt008	1.711	-0.007
idStrt009	3.721	-0.007
idStrt010	1.175	-0.007

La figura 7.8 presenta las curvas de probabilidad predichas mediante un modelo logístico con intercepto aleatorio por estrato. Se observa una relación inversa entre el gasto y la probabilidad de pobreza, de modo que hogares con mayores niveles de gasto presentan una menor probabilidad de ser clasificados como pobres. Asimismo, las diferencias entre las curvas reflejan la heterogeneidad existente entre estratos, capturada mediante los interceptos aleatorios del modelo.

```
prediction_data <- plot_data %>%
  group_by(Stratum) %>%
  summarise(
    Expenditure = list(seq(min(Expenditure), max(Expenditure), len = 100))
  ) %>%
  tidyr::unnest_legacy()

prediction_data <- prediction_data %>%
  mutate(
    probability = predict(
      random_intercept_logit_model,
      newdata = prediction_data,
      type = "response"
    )
  )
```

```
ggplot(data = prediction_data,
       aes(y = probability, x = Expenditure, colour = Stratum)) +
  geom_line() +
  geom_point(data = plot_data,
```

```
            aes(y = poor, x = Expenditure)) +
  labs(y = "Poverty probability", x = "Expenditure") +
  theme(legend.position = "none")
```

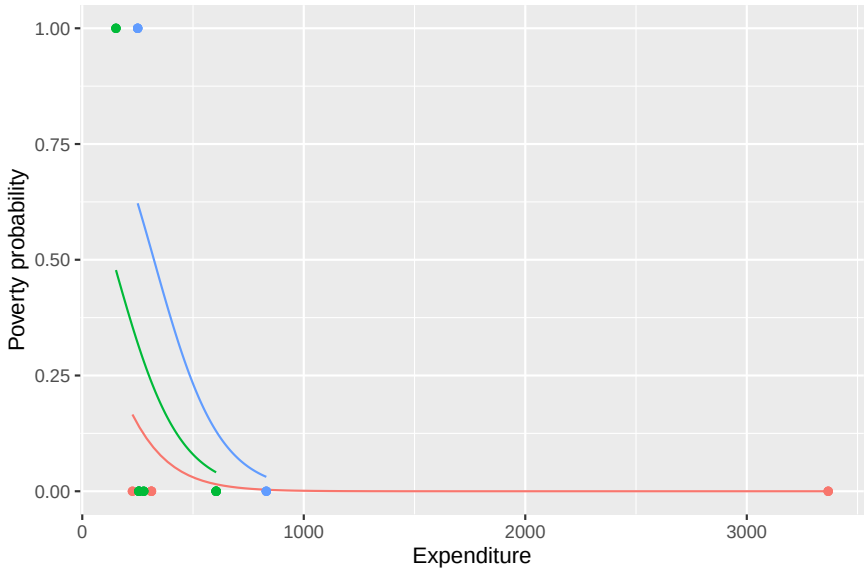

Figura 7.8.: Curvas de probabilidad predichas por estrato: modelo logístico con intercepto aleatorio

7.4.3. Modelo logístico con intercepto y pendiente aleatoria

El modelo logístico con intercepto y pendiente aleatoria extiende la especificación anterior permitiendo que no sólo varíe el nivel base del evento entre estratos, sino también el efecto de una covariable específica. Esta formulación es análoga al modelo de intercepto y pendiente aleatorios para variables continuas, con la salvedad de que las diferencias entre estratos se expresan en la escala logit y se traducen en curvas de probabilidad no lineales.

Sea x_{kj} la covariable cuyo efecto se permite variar entre estratos, y sea $\mathbf{z}_{kj}$ el conjunto de covariables con efectos fijos comunes. En este modelo, la probabilidad de éxito se asume de tal forma que $\text{logit}(\pi_{kj}) = \beta_{0j} + x_{kj}\beta_{1j} + \mathbf{z}_{kj}\beta$. En donde, β_{0j} es el intercepto específico del estrato j y β_{1j} es la pendiente específica asociada a la covariable x_{kj}.

Ambos parámetros se descomponen en una parte promedio poblacional y una desviación específica del estrato, de tal manera que $\beta_{0j} = \gamma_{00} + \tau_{0j}$ y $\beta_{1j} = \gamma_{10} + \tau_{1j}$. En donde los términos γ_{00} y γ_{10} representan, respectivamente, el intercepto promedio global y la pendiente promedio global en escala logit. Por su parte, τ_{0j} y τ_{1j} capturan cuánto se aparta el estrato j de esos promedios. La implementación en `R` es la siguiente:

```
random_slope_logit_model <- glmer(
  poor ~ 1 + Expenditure + (1 + Expenditure | Stratum),
  data = survey_data,
  weights = qk,
  family = binomial(link = "logit")
```

```
)
performance::icc(random_slope_logit_model)
```

```
# Intraclass Correlation Coefficient

    Adjusted ICC: 0.875
  Unadjusted ICC: 0.631
```

De acuerdo con Bates et al. (2015), la estimación del modelo con intercepto y pendiente aleatoria indica una heterogeneidad mucho más marcada entre estratos. En este caso, tanto los interceptos como las pendientes asociadas al gasto pueden cambiar entre grupos. Antes de interpretar una correlación intraclase muy alta en este modelo debe verificarse que el ajuste haya convergido y que la matriz de efectos aleatorios no sea singular; una estructura aleatoria demasiado compleja puede producir estimaciones de frontera. Los coeficientes del modelo por estrato se muestran en la tabla 7.7.

```
coef(random_slope_logit_model)$Stratum %>%
  slice(1:10L)
```

```
coef(random_slope_logit_model)$Stratum %>%
  slice(1:10L) %>%
  tabla_fmt()
```

Tabla 7.7.: Coeficientes del modelo logístico con intercepto y pendiente aleatoria por estrato (primeros 10 estratos)

	(Intercept)	Expenditure
idStrt001	4.865	-0.025
idStrt002	9.862	-0.035
idStrt003	-1.133	-0.007
idStrt004	1.899	-0.015
idStrt005	8.030	-0.026
idStrt006	-1.153	0.009
idStrt007	0.974	-0.012
idStrt008	1.488	-0.006
idStrt009	3.660	-0.005
idStrt010	4.133	-0.020

Los coeficientes por estrato evidencian que la relación entre gasto y pobreza no sólo cambia en el nivel base, sino también en la intensidad y dirección de la pendiente. En varios estratos la pendiente del gasto es negativa, lo que mantiene la interpretación esperada de que mayores niveles de gasto reducen la probabilidad de pobreza. Sin embargo, también aparecen pendientes cercanas a cero e incluso positivas en algunos estratos, lo que sugiere patrones locales distintos. Esta variabilidad es precisamente la que el modelo busca capturar mediante el efecto aleatorio de la pendiente. La figura 7.9 muestra que las curvas predichas son más heterogéneas entre estratos que en el modelo anterior:

```
prediction_data <- plot_data %>%
  group_by(Stratum) %>%
  summarise(
    Expenditure = list(seq(min(Expenditure), max(Expenditure), len = 100))
  ) %>%
  tidyr::unnest_legacy()

prediction_data <- prediction_data %>%
  mutate(
    probability = predict(
      random_slope_logit_model,
      newdata = prediction_data,
      type = "response"
    )
  )
```

```
ggplot(data = prediction_data,
       aes(y = probability, x = Expenditure, colour = Stratum)) +
  geom_line() +
  geom_point(data = plot_data,
             aes(y = poor, x = Expenditure)) +
  labs(y = "Poverty probability", x = "Expenditure") +
  theme(legend.position = "none")
```

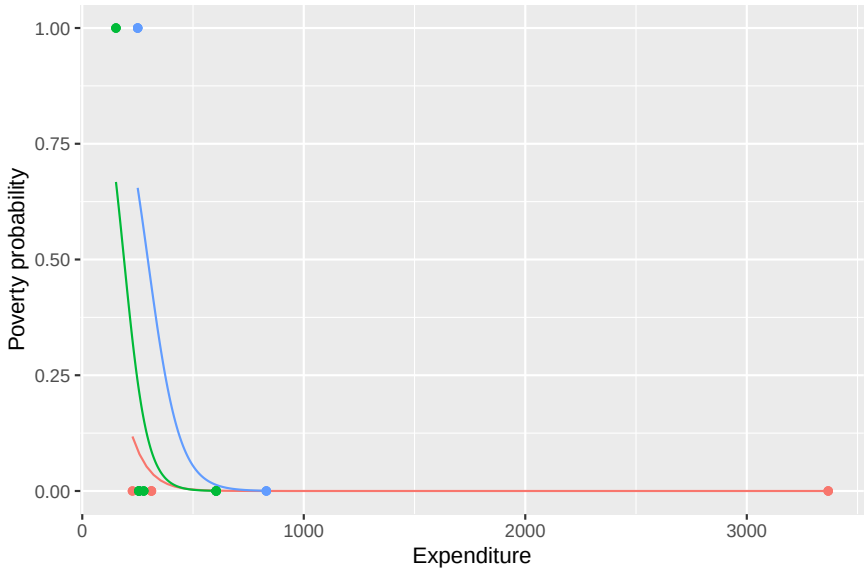

Figura 7.9.: Curvas de probabilidad predichas por estrato: modelo logístico con intercepto y pendiente aleatoria

Parte I.

Anexos

8. Manejo de bases de datos con R

De acuerdo con R Core Team (2026a), R, descrito en la literatura es un software estadístico ampliamente utilizado para el análisis de datos y la generación de gráficos y permite un ambiente integrado para el análisis estadístico, en particular de bases de datos. Esto significa que permite realizar procesos de análisis, programación, visualización y manejo de datos dentro de un mismo entorno.

En este anexo se utiliza el enfoque del `tidyverse`, cuya organización de datos se relaciona con los principios expuestos por Wickham (2014), y en particular el paquete `dplyr`, para organizar una secuencia lógica de trabajo con bases de datos: cargar la información, inspeccionarla, seleccionar variables, filtrar registros, ordenar filas, crear variables y producir resúmenes.

8.1. Preparación del entorno de trabajo

R es un lenguaje de programación de carácter colaborativo, lo que permite que su comunidad de usuarios y desarrolladores contribuya continuamente mediante la creación de funciones, paquetes y librerías especializadas. Gracias a este ecosistema abierto, es posible ampliar sus capacidades para distintos tipos de análisis estadístico y procesamiento de datos. Entre las librerías más utilizadas para el análisis de bases de datos se encuentran las siguientes:

- De acuerdo con Wickham, Averick, et al. (2026), la librería `tidyverse` corresponde a un conjunto de paquetes de R diseñados para apoyar distintas etapas del trabajo con datos, desde su importación y transformación hasta su exploración, análisis y visualización. Su uso se ha extendido ampliamente en ciencia de datos porque ofrece una forma de trabajo consistente y ordenada.
- Según Wickham, François, et al. (2026), el paquete `dplyr` ofrece herramientas para organizar y transformar bases de datos rectangulares. Su utilidad radica en que reúne, en una sintaxis simple y consistente, las operaciones más frecuentes del trabajo analítico con datos: seleccionar variables, filtrar registros, crear nuevas columnas, ordenar observaciones y resumir información.
- De acuerdo con Gutiérrez (2020), el paquete `TeachingSampling` ofrece funciones para seleccionar muestras probabilísticas y realizar inferencia sobre poblaciones finitas bajo diversos diseños de muestreo.

Antes de poder utilizar las diferentes funciones que cada librería tiene, es necesario descargarlas de antemano de la web. El comando `install.packages` permite realizar esta tarea. Note que algunas librerías pueden depender de otras, así que para poder utilizarlas es necesario instalar también las dependencias adicionales.

```
install.packages("dplyr")
install.packages("tidyverse")
install.packages("TeachingSampling")
```

Según Posit Software, PBC (2026), una vez instalados los paquetes, se cargan con `library()`. El primer comando limpia los objetos existentes en la sesión de trabajo, lo que ayuda a que los resultados del capítulo dependan solo del código que se ejecuta aquí. Debe tenerse en cuenta que un paquete solo puede cargarse si previamente ha sido instalado en el sistema. La opción `rm(list = ls())` no sustituye el inicio de una sesión limpia: para favorecer la reproducibilidad, conviene iniciar los proyectos con un espacio de trabajo vacío y ejecutar un guion completo y autosuficiente.

```
rm(list = ls())

library(tidyverse)
library(dplyr)
library(TeachingSampling)
```

Una vez se descargan e instalan las librerías o paquetes en `R`, el paso recomendado es que todos los procesamientos se realicen mediante la creación de proyectos. Un proyecto de RStudio asocia un archivo `.Rproj` con un directorio de trabajo y guarda opciones del proyecto; los archivos de origen, datos y resultados permanecen como archivos separados dentro de ese directorio. Adicionalmente, contiene información que permite la compilación de cada archivo de `R` que se va a utilizar, mantiene la información para integrarse con sistemas de control de código fuente y ayuda a organizar la aplicación de los procesamientos en componentes lógicos.

8.2. Carga e inspección inicial de la base

De acuerdo con Wickham, Çetinkaya-Rundel, y Grolemund (2023), es muy usual que al trabajar proyectos en `R` sea necesario importar bases de datos con información relevante para un estudio en particular. Los formatos que pueden importarse mediante `R` y sus paquetes son diversos; entre ellos se encuentran `.xlsx`, `.csv`, `.txt` y `.dta`, que es el formato utilizado por `Stata`. Una vez leída la base de datos en el formato pertinente es recomendable transformarla al formato nativo de `R`, es decir `.RDS`. Este es un formato más eficiente y propio de `R`, y su uso facilita la importación, la transformación y el almacenamiento dentro de un flujo reproducible de análisis.

Una vez se carga la base de datos se procede a utilizar las funciones en `R` para poder obtener resultados de los procesamientos agregados y gráficos de interés. Para ejemplificar el uso de funciones que permitan obtener resultados agregados, utilizaremos la base de datos `BigCity` del paquete `TeachingSampling`. Esta base corresponde a un conjunto de variables socioeconómicas de 150266 personas en un año en particular.

```
data(BigCity)
bigcity_data <- BigCity
```

El objeto `bigcity_data` contiene una copia de `BigCity`. La función `head()` permite observar las primeras filas y obtener una impresión rápida de la estructura de la base.

```
head(bigcity_data)
```

```
       HHID PersonID   Stratum     PSU  Zone    Sex Age MaritalST Income
1 idHH00001  idPer01 idStrt001 PSU0001 Rural   Male  38   Married    555
2 idHH00001  idPer02 idStrt001 PSU0001 Rural Female  40   Married    555
3 idHH00001  idPer03 idStrt001 PSU0001 Rural Female  20    Single    555
4 idHH00001  idPer04 idStrt001 PSU0001 Rural   Male  19    Single    555
5 idHH00001  idPer05 idStrt001 PSU0001 Rural   Male  18    Single    555
6 idHH00002  idPer01 idStrt001 PSU0001 Rural   Male  35   Married    298
  Expenditure Employment  Poverty
1         488   Employed  NotPoor
2         488   Employed  NotPoor
3         488   Inactive  NotPoor
4         488   Employed  NotPoor
5         488   Inactive  NotPoor
6         217   Employed Relative
```

Una vez cargada la base de datos en R, se puede empezar a realizar los procesamientos según las necesidades de cada investigador. En este sentido, una de las primeras revisiones que se realizan al cargar una base de datos es verificar su dimensión; es decir, identificar la cantidad de filas y columnas que contiene. Lo anterior se puede hacer con la función `nrow`, que identifica el número de registros en la base de datos, y con la función `ncol`, que muestra el número de variables. Los códigos computacionales son los siguientes:

```
nrow(bigcity_data)
```

```
[1] 150266
```

```
ncol(bigcity_data)
```

```
[1] 12
```

Una forma resumida de revisar la cantidad de filas y columnas de una base de datos es mediante la función `dim`, la cual devuelve un vector cuya primera componente corresponde al número de filas y cuya segunda componente indica el número de columnas.

```
dim(bigcity_data)
```

```
[1] 150266     12
```

En algunas ocasiones, las bases de datos pueden ser extensas, ya sea porque contienen un número considerable de variables observadas o porque el número de registros es grande. Por esta razón, para visualizarlas adecuadamente, puede ser conveniente abrirlas en una ventana independiente y explorar su contenido de forma tabular. Esto se realiza mediante la función `View`, como se muestra a continuación.

```
View(bigcity_data)
```

Otra verificación importante al cargar una base de datos en R consiste en identificar las variables que contiene. Para ello, se puede utilizar la función names, la cual muestra los nombres de las variables incluidas en la base de datos.

```
names(bigcity_data)
```

```
 [1] "HHID"        "PersonID"    "Stratum"     "PSU"         "Zone"
 [6] "Sex"         "Age"         "MaritalST"   "Income"      "Expenditure"
[11] "Employment"  "Poverty"
```

La salida anterior muestra los nombres de las variables contenidas en la base de datos. En este caso, la base cuenta con 12 variables. Las primeras variables corresponden a identificadores y elementos del diseño muestral, como HHID, PersonID, Stratum, PSU y Zone. Luego aparecen variables sociodemográficas, como Sex, Age y MaritalST. Finalmente, se incluyen variables económicas y de condición laboral o social, como Income, Expenditure, Employment y Poverty.

Si se desea profundizar en el tipo de información que almacena cada variable, puede utilizarse la función str, la cual muestra de manera compacta la estructura de un objeto y sus componentes. Para nuestra base de datos, esta función se utilizaría de la siguiente manera:

```
str(bigcity_data)
```

```
'data.frame':   150266 obs. of  12 variables:
 $ HHID       : chr  "idHH00001" "idHH00001" "idHH00001" "idHH00001" ...
 $ PersonID   : chr  "idPer01" "idPer02" "idPer03" "idPer04" ...
 $ Stratum    : chr  "idStrt001" "idStrt001" "idStrt001" "idStrt001" ...
 $ PSU        : chr  "PSU0001" "PSU0001" "PSU0001" "PSU0001" ...
 $ Zone       : chr  "Rural" "Rural" "Rural" "Rural" ...
 $ Sex        : chr  "Male" "Female" "Female" "Male" ...
 $ Age        : int  38 40 20 19 18 35 29 14 13 6 ...
 $ MaritalST  : Factor w/ 6 levels "Partner","Married",..: 2 2 5 5 5 2 2 5 5 NA ...
 $ Income     : num  555 555 555 555 555 ...
 $ Expenditure: num  488 488 488 488 488 ...
 $ Employment : Factor w/ 3 levels "Unemployed","Inactive",..: 3 3 2 3 2 3 3 NA NA NA ...
 $ Poverty    : Factor w/ 3 levels "NotPoor","Extreme",..: 1 1 1 1 1 3 3 3 3 3 ...
```

Como se observa en la salida anterior, la función str permite identificar la clase de cada variable incluida en la base de datos. Por ejemplo, la variable HHID es de tipo carácter, al igual que la variable Sex, mientras que las variables Income y Expenditure son de tipo numérico. Los demás atributos asociados con las variables también pueden revisarse directamente en la salida del código. Esta función resulta especialmente útil para obtener una visión general del contenido y la estructura de una base de datos.

8.3. Encadenamiento de operaciones

Según Bache y Wickham (2026), una de las contribuciones más útiles de R es la incorporación de las tuberías (*pipelines*), las cuales permiten construir consultas, transformaciones y nuevos objetos a partir de una base de datos de manera más intuitiva, ordenada y fácil de interpretar. El operador de encadenamiento `%>%`, parte del paquete `magrittr` y cargado automáticamente al utilizar la librería `tidyverse`, permite enlazar varias operaciones sucesivas, de modo que el resultado obtenido en un paso se convierte en el insumo del paso siguiente. A continuación, se presenta un ejemplo sencillo de su uso con la base de datos `BigCity`, en el cual se calcula el número total de elementos que contiene la base mediante la función `count`.

```
bigcity_data %>%
  count()
```

```
      n
1 150266
```

La línea de código anterior puede interpretarse de forma sencilla: se toma la base de datos como punto de partida y, sobre ella, se realiza un conteo de sus registros. Además de la función `count`, existe una amplia variedad de funciones que pueden combinarse con el operador `%>%` para realizar consultas, transformaciones y resúmenes sobre los datos. El paquete `dplyr` toma ventaja de este operador, permitiendo organizar las operaciones más comunes sobre una base rectangular. En las siguientes subsecciones se presentan algunas de ellas, siguiendo un orden típico de trabajo: reducir observaciones, conservar variables, ordenar la información, crear nuevas columnas y producir resultados agregados.

8.3.1. `filter`: selección de registros

`filter()`, una de las funciones de transformación de `dplyr`, conserva las filas que cumplen una o varias condiciones. En el siguiente ejemplo se crean dos subconjuntos de `bigcity_data`: uno con los hombres y otro con las mujeres. El resultado son dos objetos nuevos. `male_data` contiene únicamente registros con `Sex == "Male"`, mientras que `female_data` contiene registros con `Sex == "Female"`.

```
male_data <- bigcity_data %>%
  filter(Sex == "Male")
female_data <- bigcity_data %>%
  filter(Sex == "Female")
```

También puede filtrarse por condición de pobreza. En este caso se conservan solo las personas clasificadas como no pobres. El objeto `nonpoor_data` contiene el subconjunto de registros cuya variable `Poverty` toma el valor `"NotPoor"`.

```
nonpoor_data <- bigcity_data %>%
  filter(Poverty == "NotPoor")
```

La función `filter()` también permite conservar registros cuyos valores pertenecen a un conjunto específico. En los siguientes ejemplos se filtra por valores particulares de ingreso. Note que `%in%` evalúa si `Income` pertenece al conjunto indicado. Así, `working_data` conserva ingresos iguales a 265 o 600, y `income_filter_data` conserva ingresos iguales a 1000 o 2000.

```
working_data <- bigcity_data %>%
  filter(Income %in% c(265, 600))

income_filter_data <- bigcity_data %>%
  filter(Income %in% c(1000, 2000))
```

8.3.2. `select`: seleccionar variables

La función `select()` conserva columnas específicas. Para continuar con los ejemplos, se seleccionan variables de identificación de la persona y el hogar y otras como zona, sexo, edad e ingreso. El resultado es una base más reducida en columnas. Se mantiene `Age` porque será utilizada posteriormente para ordenar registros por edad.

```
working_data <- bigcity_data %>%
  select(PersonID, HHID, PSU, Zone, Sex, Age, Income)
```

Adicionalmente, `select()` también permite excluir variables. Para eliminar columnas, se antepone el signo menos (-) al nombre de cada variable. El objeto `compact_data` conserva las variables de `working_data`, excepto `HHID` y `PersonID`. Esta operación es útil cuando se desea trabajar con una base más compacta.

```
compact_data <- working_data %>%
  select(-HHID, -PersonID)
```

8.3.3. `arrange`: ordenamiento de filas

La función `arrange()` ordena las filas de una base según una o más variables. En el primer ejemplo, `working_data` se ordena de menor a mayor ingreso, mientras que `sorted_data` contiene las mismas columnas que `working_data`, pero sus filas quedan ordenadas por `Income`. La función `head()` muestra los primeros registros después del ordenamiento.

```
sorted_data <- working_data %>%
  arrange(Income)

sorted_data %>%
  head()
```

```
  PersonID      HHID     PSU  Zone    Sex Age Income
1  idPer01 idHH01522 PSU0112 Rural Female  82   10.0
2  idPer01 idHH22167 PSU0112 Rural Female  82   10.0
```

```
3  idPer01 idHH10745 PSU0893 Rural   Male  68   11.9
4  idPer01 idHH31390 PSU0893 Rural   Male  68   11.9
5  idPer01 idHH17632 PSU1432 Rural   Male  58   12.1
6  idPer02 idHH17632 PSU1432 Rural Female  50   12.1
```

El ordenamiento también puede realizarse considerando más de una variable. En este ejemplo, la base se organiza primero según la variable `Sex` y, posteriormente, dentro de cada grupo, de acuerdo con la variable `Age`. La siguiente salida permite revisar los primeros registros después de aplicar este criterio de ordenamiento.

```
working_data %>%
  arrange(Sex, Age) %>%
  head()
```

```
  PersonID      HHID     PSU  Zone    Sex Age Income
1  idPer04 idHH00009 PSU0001 Rural Female   0    152
2  idPer04 idHH20654 PSU0001 Rural Female   0    152
3  idPer06 idHH00042 PSU0004 Rural Female   0    528
4  idPer06 idHH20687 PSU0004 Rural Female   0    528
5  idPer04 idHH00191 PSU0014 Urban Female   0    355
6  idPer04 idHH20836 PSU0014 Urban Female   0    355
```

Para ordenar los registros de mayor a menor, se utiliza la opción `desc()` dentro de la función `arrange()`. En el siguiente ejemplo, la base se organiza de manera que las personas con mayor edad aparezcan en los primeros registros. Este tipo de ordenamiento permite identificar rápidamente los valores más altos de una variable dentro de la base reducida.

```
working_data %>%
  arrange(desc(Age)) %>%
  head()
```

```
  PersonID      HHID     PSU  Zone    Sex Age Income
1  idPer02 idHH01024 PSU0076 Urban Female  98    135
2  idPer02 idHH21669 PSU0076 Urban Female  98    135
3  idPer01 idHH02916 PSU0219 Rural Female  98    137
4  idPer01 idHH23561 PSU0219 Rural Female  98    137
5  idPer01 idHH03863 PSU0290 Rural   Male  98    245
6  idPer01 idHH24508 PSU0290 Rural   Male  98    245
```

8.3.4. `mutate`: creación o transformación de variables

La función `mutate()` crea nuevas variables o modifica variables existentes. Dado que `BigCity` no incluye una variable de región geográfica explícita, se construye `Region` a partir de `Stratum`. Primero se extrae la parte numérica del estrato con `gsub("\\D", "", Stratum)`, luego se convierte a número con `as.numeric()` y finalmente se divide en cinco intervalos con `cut()`. El resultado es una nueva

columna `Region` dentro de `bigcity_data`, en la que sus categorías permiten producir resúmenes regionales en los pasos siguientes. Esta variable es únicamente una recodificación didáctica de los identificadores numéricos de `Stratum`; no debe interpretarse como una región geográfica observada ni emplearse como tal en un análisis sustantivo.

```
bigcity_data <- bigcity_data %>%
  mutate(
    Region = cut(
      as.numeric(gsub("\\D", "", Stratum)),
      breaks = 5,
      labels = c("North", "South", "Center", "West", "East")
    )
  )
```

En el análisis de bases de datos también es frecuente recodificar variables categóricas. El siguiente código crea `region_id`, una versión codificada de `Region` que conserva el orden de las regiones y asigna códigos de dos dígitos. Esta forma de recodificación facilita la presentación o el cruce con otras fuentes de información.

```
bigcity_data <- bigcity_data %>%
  mutate(
    region_id = factor(
      Region,
      levels = c("North", "South", "Center", "West", "East"),
      labels = c("01", "02", "03", "04", "05")
    )
  )
```

El siguiente ejemplo crea `log_income`, que corresponde al logaritmo natural del ingreso (transformación de la escala del ingreso). Esta operación se usa con frecuencia cuando una variable monetaria presenta alta dispersión. Luego se muestran las primeras filas de las variables relevantes. Como el logaritmo no está definido para ingresos iguales o menores que cero, esos casos se convierten explícitamente en valores faltantes antes de calcular la transformación.

```
log_income_data <- working_data %>%
  mutate(
    log_income = if_else(
      Income > 0,
      log(Income),
      NA_real_
    )
  )

log_income_data %>%
  select(PersonID, Income, log_income) %>%
  head()
```

```
  PersonID Income log_income
1  idPer01    555       6.32
2  idPer02    555       6.32
3  idPer03    555       6.32
4  idPer04    555       6.32
5  idPer05    555       6.32
6  idPer01    298       5.70
```

Note que `mutate()` puede crear más de una variable en el mismo paso. Además, las variables creadas al inicio de `mutate()` pueden ser utilizadas en cálculos posteriores dentro del mismo llamado. Por ejemplo, `income_2` duplica el ingreso original e `income_4` duplica `income_2`. La siguiente salida muestra que `income_4` equivale a cuatro veces el ingreso inicial.

```
income_transform_data <- log_income_data %>%
  mutate(
    income_2 = 2 * Income,
    income_4 = 2 * income_2
  )

income_transform_data %>%
  select(PersonID, income_2, income_4) %>%
  head()
```

```
  PersonID income_2 income_4
1  idPer01     1110     2220
2  idPer02     1110     2220
3  idPer03     1110     2220
4  idPer04     1110     2220
5  idPer05     1110     2220
6  idPer01      597     1193
```

También se puede crear una variable categórica mediante condiciones. En este ejemplo, la función `case_when()` clasifica la edad en rangos y luego convierte el resultado en un factor ordenado. El resultado es `age_group`, una variable de grupos etarios ordenados. Esta transformación facilita la producción de tablas por tramos de edad.

```
bigcity_data <- bigcity_data %>%
  mutate(
    age_group = case_when(
      Age <= 5 ~ "0-5",
      Age <= 15 ~ "6-15",
      Age <= 30 ~ "16-30",
      Age <= 45 ~ "31-45",
      Age <= 60 ~ "46-60",
      TRUE ~ "Over 60"
    ),
```

```
    age_group = factor(
      age_group,
      levels = c(
        "0-5", "6-15", "16-30",
        "31-45", "46-60", "Over 60"
      ),
      ordered = TRUE
    )
  )
```

8.3.5. `summarise`: resúmenes descriptivos

De acuerdo con Heeringa, West, y Berglund (2017), la función `summarise()` permite construir una tabla resumen a partir de una o varias variables de la base de datos. En el siguiente ejemplo se calculan dos estadísticas descriptivas de la variable `Income`: el total y la media. Como resultado, se obtiene una tabla con una sola fila que contiene la suma de los ingresos observados y el ingreso promedio de los registros incluidos en la base. Estos cálculos no incorporan pesos, estratos ni conglomerados; por tanto, describen los registros de `BigCity`, pero no deben interpretarse como estimaciones de una población obtenidas bajo un diseño muestral complejo.

```
bigcity_data %>%
  summarise(
    total_income = sum(Income),
    mean_income = mean(Income)
  )
```

```
  total_income mean_income
1     87893117         585
```

También pueden calcularse medidas de localización para describir la posición de los valores dentro de la distribución. En este caso, se obtiene la mediana, el primer decil, el noveno decil y la diferencia entre estos dos últimos. El resultado permite resumir distintos puntos de la distribución del ingreso, mientras que el rango decílico muestra la distancia entre el 10 % inferior y el 10 % superior de los ingresos observados.

```
bigcity_data %>%
  summarise(
    median_income = median(Income),
    decile_1 = quantile(Income, 0.1),
    decile_9 = quantile(Income, 0.9),
    decile_range = decile_9 - decile_1
  )
```

```
  median_income decile_1 decile_9 decile_range
1           449      165     1126          961
```

Con `summarise()` también es posible calcular medidas de dispersión, tales como la varianza, la desviación estándar, el valor mínimo, el valor máximo, el rango y el rango intercuartílico. Estas estadísticas permiten describir el grado de variabilidad de los ingresos observados en la muestra. En particular, el rango se obtiene como la diferencia entre el valor máximo y el valor mínimo, mientras que el rango intercuartílico resume la dispersión del 50 % central de la distribución.

```
bigcity_data %>%
  summarise(
    variance_income = var(Income),
    sd_income = sd(Income),
    min_income = min(Income),
    max_income = max(Income),
    range_income = max_income - min_income,
    iqr_income = IQR(Income)
  )
```

```
  variance_income sd_income min_income max_income range_income iqr_income
1          332463       577         10      32920        32910        468
```

8.3.6. `group_by`: agrupación de casos

La función `group_by()` permite agrupar la base de datos de acuerdo con una o más variables. Al combinarla con `summarise()`, las estadísticas se calculan de manera independiente dentro de cada grupo definido. En el siguiente ejemplo, se cuenta el número de registros por región y luego se ordena el resultado de mayor a menor. La siguiente salida permite identificar cuántos registros están asociados con cada región de la base de datos.

```
bigcity_data %>%
  group_by(Region) %>%
  summarise(n = n()) %>%
  arrange(desc(n))
```

```
# A tibble: 5 x 2
  Region     n
  <fct>  <int>
1 East   39160
2 West   33868
3 Center 25944
4 South  25898
5 North  25396
```

Para los conteos simples, `count()` ofrece una forma abreviada de escribir la combinación entre `group_by()` y `summarise(n = n())`. En la práctica, ambas alternativas producen el mismo resultado, pues agrupan la base de datos según una variable categórica y calculan cuántos registros pertenecen a cada grupo. Por ejemplo, el siguiente código reproduce los mismos resultados que el código anterior.

```
bigcity_data %>%
  count(Region) %>%
  arrange(desc(n))
```

```
  Region     n
1   East 39160
2   West 33868
3 Center 25944
4  South 25898
5  North 25396
```

Para obtener el número de registros por sexo, se aplica la misma lógica de agrupación y conteo. En este caso, la base se agrupa según la variable `Sex` y luego se calcula el número de registros asociados con cada una de sus categorías. El resultado permite identificar cuántas observaciones corresponden a cada sexo dentro de la base de datos.

```
bigcity_data %>%
  count(Sex) %>%
  arrange(desc(n))
```

```
     Sex     n
1 Female 79190
2   Male 71076
```

El conteo también puede realizarse según la zona geográfica. En este caso, la base se agrupa por la variable `Zone` y se calcula el número de registros correspondiente a cada una de sus categorías. La siguiente salida permite conocer cómo se distribuyen las observaciones entre las distintas zonas de la base de datos.

```
bigcity_data %>%
  count(Zone) %>%
  arrange(desc(n))
```

```
   Zone     n
1 Urban 78164
2 Rural 72102
```

Además de los conteos, también es posible calcular medidas resumen por grupo, como la media. En este caso, para obtener el ingreso promedio por región junto con el número total de registros, se agrupa la base según la variable `Region` y luego se aplica `summarise()`. El resultado es una tabla regional que presenta, para cada región, el total de observaciones y el ingreso medio correspondiente.

```
bigcity_data %>%
  group_by(Region) %>%
  summarise(
    n = n(),
    mean_income = mean(Income)
  )
```

```
# A tibble: 5 x 3
  Region     n mean_income
  <fct>  <int>       <dbl>
1 North  25396        509.
2 South  25898        644.
3 Center 25944        701.
4 West   33868        571.
5 East   39160        530.
```

El mismo procedimiento puede aplicarse para calcular el ingreso medio por sexo. En este caso, la base se agrupa según la variable `Sex` y luego se obtiene el promedio de `Income` para cada categoría. La tabla resultante permite comparar el ingreso medio observado entre los distintos grupos definidos por el sexo.

```
bigcity_data %>%
  group_by(Sex) %>%
  summarise(
    n = n(),
    mean_income = mean(Income)
  )
```

```
# A tibble: 2 x 3
  Sex        n mean_income
  <chr>  <int>       <dbl>
1 Female 79190        579.
2 Male   71076        591.
```

Por último, se calculan la media, la desviación estándar y el rango intercuartílico de los ingresos según la condición de ocupación. Para ello, la base se agrupa por la variable `Employment` y luego se obtienen las estadísticas descriptivas de `Income` dentro de cada grupo. La siguiente salida presenta un resumen del comportamiento de los ingresos para cada categoría de ocupación.

```
bigcity_data %>%
  group_by(Employment) %>%
  summarise(
    n = n(),
    mean_income = mean(Income),
    sd_income = sd(Income),
    iqr_income = IQR(Income)
  )
```

```
# A tibble: 4 x 5
  Employment     n mean_income sd_income iqr_income
  <fct>      <int>       <dbl>     <dbl>      <dbl>
1 Unemployed  4630        429.      375.       392.
2 Inactive   44104        532.      553.       439.
3 Employed   62188        661.      606.       529.
4 <NA>       39344        541.      558.       407.
```

9. Inferencia en poblaciones finitas

De acuerdo con Skinner, Holt, y Smith (1989) y Kish (1987), una cuestión fundamental en el análisis de encuestas es identificar cuál es la medida de probabilidad que rige la inferencia. En la inferencia estadística clásica se suele asumir que las observaciones disponibles son variables aleatorias independientes e idénticamente distribuidas (IID). Bajo ese enfoque, la aleatoriedad proviene del modelo que genera los datos. Sin embargo, este supuesto no describe adecuadamente la información proveniente de encuestas con diseños muestrales complejos, donde las observaciones son seleccionadas mediante estratificación, conglomeración y probabilidades de inclusión desiguales.

9.1. Dos posibles inferencias

Según Särndal, Swensson, y Wretman (2003), en encuestas de hogares conviene distinguir tres objetos. Primero, la población finita, denotada como $U = \{1, 2, \dots, N\}$, la cual contiene todas las unidades de interés. Segundo, la variable de estudio y_k, observada o definida para cada unidad $k \in U$. Tercero, el diseño muestral $p(s)$, que asigna probabilidades a las posibles muestras $s \subset U$. En la notación utilizada a lo largo de este documento, una unidad puede identificarse como k dentro del conglomerado o UPM i, perteneciente al estrato h, de modo que los estimadores ponderados suelen escribirse como sumas de la forma

$$\hat{t}_y = \sum_h \sum_i \sum_k w_{hik}\, y_{hik}$$

donde w_{hik} representa el factor de expansión o peso muestral asociado con la unidad observada. En la teoría de muestreo, los valores y_k se consideran fijos una vez definida la población. La aleatoriedad no está en los valores de la variable, sino en los indicadores de inclusión

$$I_k = \begin{cases} 1, & \text{si } k \in s, \\ 0, & \text{si } k \notin s, \end{cases}$$

que inducen la probabilidad de inclusión $\pi_k = \Pr_p(I_k = 1)$ y, a su vez, el correspondiente peso de muestreo, definido como $w_k = \frac{1}{\pi_k}$.

Un estimador que ignora los pesos de muestreo puede no representar adecuadamente a la población objetivo, especialmente cuando la muestra proviene de un diseño complejo. En la inferencia basada en el diseño, la población se considera fija y la fuente de aleatoriedad proviene del mecanismo de selección de la muestra; por ello, sus elementos centrales son las probabilidades de inclusión π_k y los pesos de muestreo w_k.

Como demuestran Horvitz y Thompson (1952), por el contrario, la inferencia basada en modelos considera que los valores de la variable de interés en toda la población se conciben como realizaciones

de un proceso estocástico, usualmente denotado por ξ. Así, mientras la inferencia basada en los modelos estudia propiedades como $E_\xi(Y_k)$ y $Var_\xi(Y_k)$, la inferencia basada en el diseño se orienta a incorporar las características del plan muestral en los estimadores. La inferencia basada en el diseño no garantiza que todo estimador sea exactamente insesgado; esa propiedad debe verificarse para cada estimador y diseño. En particular, el total de Horvitz–Thompson satisface $E_p(\hat{t}_y) = t_y$ bajo probabilidades de inclusión positivas.

De acuerdo con David A. Binder (2011), esta distinción permite entender por qué, aun cuando exista un modelo razonable para la variable de interés, el diseño muestral debe incorporarse explícitamente si se desea obtener inferencias válidas para la población objetivo. Para comprender la distinción entre inferencia basada en el modelo e inferencia basada en el diseño, partimos de un ejemplo sencillo. Supóngase que se generan $N = 100$ realizaciones independientes de una variable Bernoulli con parámetro $\theta = 0.3$, donde θ representa la proporción esperada de personas desempleadas en una población generada por un modelo. Si $Y_k \sim Bernoulli(\theta)$, entonces la esperanza bajo el modelo es $E_\xi(Y_k) = \theta$, mientras que la varianza bajo el modelo es $Var_\xi(Y_k) = \theta(1-\theta)$.

En la literatura especializada, el modelo ξ recibe el nombre de modelo de superpoblación, pues genera poblaciones finitas. De esta forma, para una población finita generada por ese modelo, la media poblacional es $\bar{Y}_U = \frac{1}{N}\sum_{k\in U} Y_k$. Nótese que, bajo el modelo ξ, esta media es insesgada para θ, puesto que

$$E_\xi(\bar{Y}_U) = E_\xi\left(\frac{1}{N}\sum_{k\in U} Y_k\right) = \frac{1}{N}\sum_{k\in U} E_\xi(Y_k) = \theta.$$

La siguiente simulación de Monte Carlo reproduce este proceso. En cada repetición se genera una población completa de tamaño $N = 100$ y se calcula su media. El promedio de esas medias poblacionales debe aproximarse a θ.

```
set.seed(2026)
population_size <- 100
theta <- 0.3
n_sim_model <- 1000
model_estimates <- rep(NA, n_sim_model)

for (sim in seq_len(n_sim_model)) {
  outcome <- rbinom(population_size, 1, theta)
  model_estimates[sim] <- mean(outcome)
}

cbind(
  theta,
  expected_model_mean = mean(model_estimates),
  bias = mean(model_estimates) - theta
)
```

```
     theta expected_model_mean    bias
[1,]   0.3               0.302 0.00225
```

El código fija una semilla para que el ejercicio sea reproducible. Luego genera `n_sim_model` poblaciones independientes mediante `rbinom()`. La salida reporta el parámetro verdadero `theta`, el promedio de las medias simuladas `expected_model_mean` y el sesgo Monte Carlo. Como se espera bajo el modelo Bernoulli IID, `expected_model_mean` queda muy cerca de 0.3 y el sesgo es cercano a cero. Las pequeñas diferencias se deben a que se usan 1000 simulaciones, no un número infinito de repeticiones.

En la teoría de muestreo, en cambio, las características de interés son parámetros fijos de la población. Si una persona está desempleada, su estado se considera un valor fijo de la población finita. Lo aleatorio es el mecanismo mediante el cual se selecciona la muestra. El total poblacional se define como $t_y = \sum_{k \in U} y_k$, mientras que la media poblacional es $\bar{y}_U = \frac{t_y}{N}$. Si se observa una muestra s, se propone estimar el total como

$$\hat{t}_y = \sum_{k \in s} \frac{y_k}{\pi_k} = \sum_{k \in s} w_k y_k.$$

Como se mencionó en los capítulos anteriores, cuando el tamaño poblacional N es desconocido, la media poblacional puede estimarse mediante el estimador de razón ponderado, definido como

$$\hat{\bar{y}} = \frac{\sum_{k \in s} w_k y_k}{\sum_{k \in s} w_k}$$

Según Cochran (1977), estos estimadores incorporan las probabilidades de inclusión en su forma funcional y esa es la diferencia central frente al promedio simple no ponderado de una muestra. Para ilustrar el impacto del diseño, supóngase que la población anterior se divide en N_I UPMs. La UPM i tiene tamaño N_i, total $t_{y_i} = \sum_{k \in U_i} y_{ik}$ y media $\bar{y}_i = t_{y_i}/N_i$. Bajo un diseño de tamaño fijo que asigna probabilidades de inclusión de primer orden proporcionales a N_i, siempre que todas sean menores o iguales que uno, estas probabilidades toman la siguiente forma:

$$\pi_{Ii} = \frac{n_I N_i}{N}$$

Cuando todos los miembros de las UPM seleccionadas son observados, el estimador de Horvitz–Thompson de la media poblacional es exactamente insesgado bajo las probabilidades de inclusión anteriores y puede escribirse como

$$\hat{\bar{y}} = \frac{1}{N} \sum_{i \in S_I} \frac{t_{y_i}}{\pi_{Ii}} = \frac{1}{N} \sum_{i \in S_I} \frac{N_i \bar{y}_i}{n_I N_i / N} = \frac{1}{n_I} \sum_{i \in S_I} \bar{y}_i.$$

Esta expresión ejemplifica por qué, bajo este diseño específico de selección proporcional al tamaño, el promedio de las medias de los conglomerados brinda insesgamiento. En contraste, el promedio simple de las personas observadas en las UPMs seleccionadas, sería:

$$\bar{y}_s = \frac{\sum_{i \in S_I} t_{y_i}}{\sum_{i \in S_I} N_i},$$

el cual trata la muestra como si fuera autoponderada. Este estimador puede ser sesgado cuando las probabilidades de selección son desiguales, especialmente si el tamaño del conglomerado se relaciona con la variable de interés. La siguiente simulación fija una población finita y repite muchas veces la selección de hogares con probabilidades proporcionales al tamaño mediante `S.piPS()` del paquete `TeachingSampling`. En cada muestra se calculan dos estimadores: `design_estimates`, que corresponde a $\hat{\bar{y}}$, y `simple_estimates`, que corresponde al promedio simple de las personas observadas en las UPMs seleccionadas, $\bar{y}_s$. Además, estos estimadores se comparan con el parámetro del modelo `theta_population`, correspondiente a $\theta = 0.3$.

```
library(TeachingSampling)

set.seed(2026)
population_size <- 100
theta <- 0.3
n_sim_sampling <- 1000
design_estimates <- simple_estimates <- rep(NA, n_sim_sampling)

cluster_size <- rep(2:6, each = 5)
n_clusters <- length(cluster_size)
cluster_id <- rep(seq_len(n_clusters), cluster_size)
size_center <- weighted.mean(cluster_size, cluster_size)
prob_person <- theta + rep((cluster_size - size_center) / 12, cluster_size)
outcome <- rbinom(population_size, 1, prob_person)
theta_population <- mean(outcome)
cluster_totals <- tapply(outcome, cluster_id, sum)
cluster_means <- tapply(outcome, cluster_id, mean)

sampled_cluster_count <- floor(n_clusters * 0.3)
for (sim in seq_len(n_sim_sampling)) {
  pps_sample <- S.piPS(sampled_cluster_count, cluster_size)
  sampled_clusters <- pps_sample[, 1]
  sampled_cluster_size <- cluster_size[sampled_clusters]
  sampled_cluster_total <- cluster_totals[sampled_clusters]
  sampled_cluster_mean <- cluster_means[sampled_clusters]
  design_estimates[sim] <- mean(sampled_cluster_mean)
  simple_estimates[sim] <-
    sum(sampled_cluster_total) / sum(sampled_cluster_size)
}

cbind(
  theta_population,
  expected_pps = mean(design_estimates),
  bias_pps = mean(design_estimates) - theta_population,
  expected_simple = mean(simple_estimates),
  bias_simple = mean(simple_estimates) - theta_population
)
```

```
     theta_population expected_pps bias_pps expected_simple bias_simple
```

```
[1,]          0.3        0.304  0.00368          0.358       0.0585
```

El primer bloque del código genera hogares de tamaños desiguales, entre 2 y 6 personas. Para hacer visible el efecto del diseño, la probabilidad individual de desempleo varía suavemente con el tamaño del hogar, pero se centra de manera que el promedio esperado de la población siga siendo cercano a `theta`. Luego se calcula `theta_population`, que es la proporción real de desempleo en esa población específica. En cada repetición, `S.piPS()` selecciona UPMs con probabilidades proporcionales al tamaño; después se calculan la media de las medias de las UPMs seleccionadas y el promedio simple de las personas observadas. La salida compara ambos estimadores con `theta_population`. El sesgo de Monte Carlo de `expected_pps` es cercano a cero, mientras que el de `expected_simple` no lo es, mostrando resultado de ignorar el diseño en este ejercicio.

9.2. La integración de ambas inferencias

No obstante, en muchas aplicaciones ambos marcos inferenciales no se presentan como enfoques excluyentes, sino como componentes complementarios de un mismo problema. En particular, puede considerarse que la muestra ha sido obtenida mediante un diseño probabilístico complejo y que, al mismo tiempo, los valores de la variable de interés responden a un modelo de superpoblación. David A. Binder (2011) plantea esta formulación como un esquema de inferencia doble, en el cual la incertidumbre no proviene de una única fuente, sino de la combinación entre el mecanismo que genera los valores poblacionales (modelo de superpoblación) y el mecanismo que determina qué unidades son observadas (diseño de muestreo).

Formalmente, intervienen dos medidas de probabilidad: ξ, asociada al proceso generador de los valores Y_k, y $p(\cdot)$, asociada al diseño muestral que selecciona la muestra s. Por tanto, las propiedades de los estimadores deben evaluarse considerando simultáneamente ambas fuentes de variación. Bajo este marco, la esperanza combinada puede expresarse como

$$E_{\xi p}(\hat{\theta}) = E_{\xi}\{E_p(\hat{\theta} \mid U)\},$$

Donde primero se evalúa el comportamiento del estimador bajo el diseño, condicionado a la población finita generada, y luego se promedia sobre el modelo de superpoblación. En este contexto, la verosimilitud conjunta completa puede resultar intratable, especialmente cuando las probabilidades de inclusión dependen de variables relacionadas con la respuesta. Las ecuaciones de estimación ponderadas de David A. Binder (1983) fundamentan la máxima pseudo-verosimilitud (MPV), que pondera las contribuciones de cada unidad por el inverso de su probabilidad de inclusión; Pfeffermann (1993) examina las condiciones y los objetivos bajo los cuales resulta pertinente incorporar esos pesos en modelos para datos de encuestas.

9.2.1. Método de la máxima verosimilitud

El método de máxima verosimilitud, presentado formalmente por Casella y Berger (2002), parte de una muestra aleatoria $y_1, \ldots, y_n$ generada por una distribución con densidad o función de probabilidad $f(y;\theta)$. Si las observaciones son independientes e idénticamente distribuidas (IID), la función de verosimilitud se define como

$$L(\theta) = \prod_{k=1}^{n} f(y_k; \theta).$$

El límite superior correcto es el tamaño muestral n, no el tamaño poblacional N. Como los productos pueden ser difíciles de manipular, se utiliza la log-verosimilitud:

$$\ell(\theta) = \sum_{k=1}^{n} \log f(y_k; \theta)$$

El estimador de máxima verosimilitud $\hat{\theta}$ es el valor de θ que maximiza $\ell(\theta)$. Si la función es diferenciable, este valor satisface las ecuaciones de score:

$$U(\theta) = \frac{\partial\ \ell(\theta)}{\partial\ \theta} = \sum_{k=1}^{N} u_k(\theta) = 0$$

donde $u_k(\theta) = \frac{\partial}{\partial\theta} \log f(y_k; \theta)$ es la contribución de la unidad k al score total. Para una distribución Bernoulli con parámetro θ, la función de probabilidad es $f(y_k; \theta) = \theta^{y_k}(1-\theta)^{1-y_k}$, para $y_k \in \{0, 1\}$. La log-verosimilitud es

$$\ell(\theta) = \sum_{k=1}^{N} [y_k \log(\theta) + (1 - y_k) \log(1 - \theta)]$$

Al derivar e igualar a cero y despejar, se obtiene el estimador MV, definido por:

$$\hat{\theta}_{MV} = \frac{1}{n} \sum_{k=1}^{n} y_k$$

Así, la proporción muestral es el estimador natural del parámetro Bernoulli. El punto clave es que este resultado depende del supuesto de que las observaciones tienen igual distribución y no provienen de un diseño con probabilidades desiguales.

Por otro lado, en un modelo de regresión lineal múltiple con errores normales, se tiene que $\mathbf{y} \sim N(\mathbf{X}\beta, \sigma^2 I)$. En este caso, la log-verosimilitud, omitiendo constantes, puede escribirse como

$$\ell(\beta, \sigma^2) = -\frac{n}{2} \log(\sigma^2) - \frac{1}{2\sigma^2} (\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta)$$

Maximizar esta expresión respecto de β equivale a minimizar la suma de cuadrados de los residuos. Por ello, el estimador de máxima verosimilitud de β coincide con el estimador de mínimos cuadrados ordinarios $\hat{\beta}_{MV} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$. Este resultado es importante porque muestra que muchos procedimientos estadísticos usuales pueden interpretarse como máxima verosimilitud bajo supuestos de modelo específicos. Sin embargo, si los datos provienen de una encuesta compleja, la matriz $\mathbf{X}'\mathbf{X}$ no refleja por sí sola el mecanismo de selección de la muestra.

9.2.2. Método de máxima pseudo-verosimilitud

Cuando el diseño muestral es complejo, los datos observados no satisfacen necesariamente los supuestos de independencia e idéntica distribución. Algunas unidades pueden representar a muchas personas de la población y otras a pocas. Además, dos unidades seleccionadas dentro del mismo conglomerado pueden estar correlacionadas. En este contexto, usar directamente la log-verosimilitud ordinaria puede producir estimaciones que describen la muestra, pero no la población objetivo.

La máxima pseudo-verosimilitud, desarrollada a partir de los resultados de David A. Binder (1983), incorpora los pesos de muestreo en la log-verosimilitud. Para unidades observadas $k \in s$, se define

$$\ell_p(\theta) = \sum_{k \in s} w_k \log f(y_k; \theta)$$

En la notación de esta sección, el peso $w_k = 1/\pi_k$ indica cuántas unidades de la población representa la observación y_k. Por eso, la contribución de una unidad con baja probabilidad de inclusión se amplifica en la pseudo-verosimilitud. Las ecuaciones de estimación resultantes son

$$U_p(\theta) = \sum_{k \in s} w_k \; u_k(\theta) = 0$$

La estimación MPV conserva la lógica de máxima verosimilitud, pero reemplaza el *score* ordinario por un *score* ponderado. La solución $\hat{\theta}_{MPV}$ es el valor que hace que la suma ponderada de contribuciones individuales sea igual a cero. Para una distribución Bernoulli, la pseudo-log-verosimilitud es

$$\ell_p(\theta) = \sum_{k \in s} w_k \left[y_k \log(\theta) + (1 - y_k) \log(1 - \theta) \right].$$

Al derivar e igualar a cero, se obtiene que:

$$\frac{\partial \ell_p(\theta)}{\partial \theta} = \sum_{k \in s} w_k \left[\frac{y_k}{\theta} - \frac{1 - y_k}{1 - \theta} \right] = 0$$

Cuya solución es:

$$\hat{\theta}_{MPV} = \frac{\sum_{k \in s} w_k y_k}{\sum_{k \in s} w_k} = \hat{p}_d$$

Por tanto, para una variable binaria, la MPV conduce al estimador ponderado de la proporción, equivalente al estimador de Hájek. Este resultado conecta directamente la teoría de pseudo-verosimilitud con los estimadores de proporciones presentados en los capítulos de variables categóricas.

Asimismo, para un modelo de regresión lineal múltiple, la pseudo-log-verosimilitud ponderada implica minimizar una suma de cuadrados ponderada:

$$(\mathbf{y} - \mathbf{X}\beta)'\mathbf{W}(\mathbf{y} - \mathbf{X}\beta),$$

donde

$$\mathbf{W} = \begin{pmatrix} w_1 & 0 & \cdots & 0 \\ 0 & w_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & w_n \end{pmatrix}.$$

La solución es el estimador de mínimos cuadrados ponderados:

$$\hat{\beta}_{MPV} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{y}$$

De acuerdo con Heeringa, West, y Berglund (2017), este resultado extiende de manera natural el estimador ordinario al contexto de encuestas complejas: cada observación contribuye al ajuste del modelo en proporción a su peso muestral. No obstante, para la inferencia no basta con ponderar la estimación puntual; también se requiere estimar correctamente la varianza considerando estratos, conglomerados y pesos.

La siguiente simulación de Monte Carlo anidada ilustra la inferencia doble. En el ciclo externo se genera una nueva población bajo el modelo Bernoulli. En el ciclo interno se seleccionan muchas muestras por diseño y se comparan el estimador que incorpora el diseño y el promedio simple de la muestra.

```
set.seed(2026)
population_size <- 100
theta <- 0.3
n_sim_population <- 100
expected_design_estimates <- expected_simple_estimates <- population_means <-
  rep(NA, n_sim_population)

cluster_size <- rep(2:6, each = 5)
n_clusters <- length(cluster_size)
cluster_id <- rep(seq_len(n_clusters), cluster_size)
size_center <- weighted.mean(cluster_size, cluster_size)
prob_person <- theta + rep((cluster_size - size_center) / 12, cluster_size)
sampled_cluster_count <- floor(n_clusters * 0.3)

for (population_sim in seq_len(n_sim_population)) {
  outcome <- rbinom(population_size, 1, prob_person)
  population_means[population_sim] <- mean(outcome)
  cluster_totals <- tapply(outcome, cluster_id, sum)
  cluster_means <- tapply(outcome, cluster_id, mean)

  design_estimates <- simple_estimates <- rep(NA, 100)
  for (sample_sim in seq_len(100)) {
    pps_sample <- S.piPS(sampled_cluster_count, cluster_size)
```

```
    sampled_clusters <- pps_sample[, 1]
    sampled_cluster_size <- cluster_size[sampled_clusters]
    sampled_cluster_total <- cluster_totals[sampled_clusters]
    sampled_cluster_mean <- cluster_means[sampled_clusters]
    design_estimates[sample_sim] <- mean(sampled_cluster_mean)
    simple_estimates[sample_sim] <-
      sum(sampled_cluster_total) / sum(sampled_cluster_size)
  }
  expected_design_estimates[population_sim] <- mean(design_estimates)
  expected_simple_estimates[population_sim] <- mean(simple_estimates)
}

cbind(
  theta,
  mean_population = mean(population_means),
  expected_pps = mean(expected_design_estimates),
  bias_pps = mean(expected_design_estimates) - theta,
  expected_simple = mean(expected_simple_estimates),
  bias_simple = mean(expected_simple_estimates) - theta
)
```

```
     theta mean_population expected_pps bias_pps expected_simple bias_simple
[1,]   0.3           0.304        0.302  0.00248           0.332      0.0317
```

El código reproduce las dos fuentes de aleatoriedad. El ciclo externo representa el modelo ξ, porque genera nuevas poblaciones finitas con `rbinom()`. El ciclo interno representa el diseño p, porque para cada población se seleccionan múltiples muestras con `S.piPS()`. La salida incluye `mean_population`, que es el promedio de las medias poblacionales generadas; `expected_pps`, que resume el comportamiento del estimador que incorpora el diseño; y `expected_simple`, que resume el promedio simple de la muestra seleccionada. Cuando el diseño se incorpora correctamente, el sesgo Monte Carlo respecto de `theta` debe ser pequeño. La comparación con `expected_simple` muestra por qué la inferencia basada en encuestas requiere pesos y varianzas de diseño, incluso cuando se parte de un modelo estadístico sencillo.

Como explica Pfeffermann (1993), en síntesis, la máxima verosimilitud ordinaria es adecuada cuando las observaciones pueden tratarse como IID bajo un modelo probabilístico. En encuestas complejas, la máxima pseudo-verosimilitud ofrece una extensión natural: conserva la estructura de las ecuaciones de estimación, pero pondera cada contribución individual por su peso muestral. La ponderación protege frente a ciertos diseños informativos o a determinadas especificaciones incorrectas, pero su necesidad y sus propiedades dependen del diseño y del modelo. Así, los modelos estadísticos se ajustan de manera más coherente con la población que la encuesta busca representar.

Referencias

Agresti, Alan. 2019. *An Introduction to Categorical Data Analysis*. 3.ª ed. Hoboken, NJ: John Wiley & Sons. https://doi.org/10.1002/9781119405283.

Asparouhov, Tihomir. 2006. «General Multi-Level Modeling with Sampling Weights». *Communications in Statistics, Theory and Methods* 35 (3): 439-60. https://doi.org/10.1080/03610920500476598.

Bache, Stefan Milton, y Hadley Wickham. 2026. *magrittr: A Forward-Pipe Operator for R*. https://doi.org/10.32614/CRAN.package.magrittr.

Bates, Douglas, Martin Mächler, Ben Bolker, y Steve Walker. 2015. «Fitting Linear Mixed-Effects Models Using lme4». *Journal of Statistical Software* 67 (1): 1-48. https://doi.org/10.18637/jss.v067.i01.

Binder, David A. 1983. «On the variances of asymptotically normal estimators from complex surveys». *International Statistical Review/Revue Internationale de Statistique*, 279-92.

Binder, David A. 2011. «Estimating model parameters from a complex survey under a model-design randomization framework». *Pakistan Journal of Statistics* 27 (4).

Binder, David A., y Milorad S. Kovačević. 1995. «Estimating Some Measures of Income Inequality from Survey Data: An Application of the Estimating Equations Approach». *Survey Methodology* 21 (2): 137-45. https://www150.statcan.gc.ca/n1/en/catalogue/12-001-X199500214396.

Browne, William J., y David Draper. 2006. «A Comparison of Bayesian and Likelihood-Based Methods for Fitting Multilevel Models». *Bayesian Analysis* 1 (3): 473-514. https://doi.org/10.1214/06-BA117.

Bruch, Christian, Ralf Münnich, y Stefan Zins. 2011. «Variance estimation for complex surveys». *Deliverable D3* 1.

Cai, Tianji. 2013. «Investigation of Ways to Handle Sampling Weights for Multilevel Model Analyses». *Sociological Methodology* 43 (1): 178-219. https://doi.org/10.1177/0081175013490035.

Carle, Adam C. 2009. «Fitting Multilevel Models in Complex Survey Data with Design Weights: Recommendations». *BMC Medical Research Methodology* 9: 49. https://doi.org/10.1186/1471-2288-9-49.

Casella, George, y Roger L. Berger. 2002. *Statistical Inference*. 2.ª ed. Pacific Grove, CA: Duxbury.

CEPAL. 2023. *Diseño y análisis estadístico de las encuestas de hogares de América Latina*. Metodologías de la CEPAL 5. Santiago: Naciones Unidas. https://repositorio.cepal.org/server/api/core/bitstreams/14d443ec-8f1e-44a7-b22b-9f124f205abe/content.

CEPAL. 2025. *Panorama Social de América Latina y el Caribe, 2025: cómo salir de la trampa de alta desigualdad, baja movilidad social y débil cohesión social*. LC/PUB.2025/23-P/-*. Santiago: Comisión Económica para América Latina y el Caribe. https://repositorio.cepal.org/server/api/core/bitstreams/61e9dfa1-de2a-4af6-b670-0e2ed9b00cfe/content.

Chambers, Raymond L., y Chris J. Skinner, eds. 2003. *Analysis of Survey Data*. Chichester: John Wiley & Sons. https://doi.org/10.1002/0470867205.

Cochran, William Gemmell. 1977. *Sampling techniques*. John Wiley & Sons.

Deaton, Angus. 1997. *The Analysis of Household Surveys: A Microeconometric Approach to Development Policy*. Baltimore, MD: World Bank; The Johns Hopkins University Press.

Deville, Jean-Claude, y Carl-Erik Särndal. 1992. «Calibration Estimators in Survey Sampling». *Journal of the American Statistical Association* 87 (418): 376-82. https://doi.org/10.1080/01621459.1992.10475217.

Díaz Monroy, Luis Guillermo, Mario Alfonso Morales Rivera, y Leidy Rocío León Dávila. 2018. *Análisis estadístico de datos categóricos*. 1.ª ed. Bogotá: Editorial Universidad Nacional de Colombia. https://portaldelibros.unal.edu.co/gpd-anyalisis-estadyustico-de-datos-categy-ricos-9789587834376-667e35367b60f.html.

Dorfman, Alan H., y Richard Valliant. 1993. «Quantile Variance Estimators in Complex Surveys». En *Proceedings of the Survey Research Methods Section*, 866-71. American Statistical Association.

Efron, Bradley. 1979. «Bootstrap methods: Another look at the jackknife». *The Annals of Statistics* 7 (1): 1-26. https://doi.org/10.1214/aos/1176344552.

FAO. 2026. «Indicador 2.1.1: Prevalencia de la subalimentación». https://www.fao.org/sustainable-development-goals-data-portal/data/indicators/2.1.1-prevalence-of-undernourishment/es.

Fay, Robert E. 1979. «On Adjusting the Pearson Chi-Square Statistic for Cluster Sampling». *Proceedings of the Social Statistics Section, American Statistical Association*, 402-5.

Fellegi, Ivan P. 1980. «Approximate Tests of Independence and Goodness of Fit Based on Stratified Multistage Samples». *Journal of the American Statistical Association* 75 (370): 261-68. https://doi.org/10.2307/2287405.

Ferreira, Juan Pablo. 2026. *weightflow: Declarative API for Staged Survey Weights*. https://doi.org/10.32614/CRAN.package.weightflow.

Finch, W. Holmes, Jocelyn E. Bolin, y Ken Kelley. 2019. *Multilevel Modeling Using R*. 2.ª ed. Boca Raton, FL: CRC Press.

Fox, John, y Sanford Weisberg. 2019. *An R Companion to Applied Regression*. 3.ª ed. Thousand Oaks, CA: SAGE Publications.

Freedman Ellis, Greg, y Ben Schneider. 2026. *srvyr: 'dplyr'-Like Syntax for Summary Statistics of Survey Data*. https://doi.org/10.32614/CRAN.package.srvyr.

Fuller, Wayne A. 1975. «Regression analysis for sample survey». *Sankhya, Series C* 37: 117-32.

———. 2002. «Regression estimation for survey samples». *Survey Methodology* 28 (1): 5-23.

———. 2009. *Sampling Statistics*. Wiley Series en Survey Methodology. Hoboken, New Jersey: John Wiley & Sons. https://doi.org/10.1002/9780470523551.

Gambino, Jack G, y Pedro Luis do Nascimento Silva. 2009. «Sampling and estimation in household surveys». En *Handbook of Statistics*, 29:407-39. Elsevier.

Gelman, Andrew, y Jennifer Hill. 2006. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. New York: Cambridge University Press.

Goldstein, Harvey. 2011. *Multilevel Statistical Models*. 4.ª ed. Chichester: John Wiley & Sons.

Groves, Robert M., Floyd J. Fowler, Mick P. Couper, James M. Lepkowski, Eleanor Singer, y Roger Tourangeau. 2009. *Survey Methodology*. 2.ª ed. Hoboken, NJ: John Wiley & Sons.

Gutiérrez, Andrés. 2016. *Estrategias de muestreo: diseño de encuestas y estimación de parámetros*. Segunda edición. Ediciones de la U.

———. 2020. *TeachingSampling: Selection of Samples and Parameter Estimation in Finite Population*. https://doi.org/10.32614/CRAN.package.TeachingSampling.

Gutiérrez, Andrés, y Giovany Babativa-Márquez. 2023. «Efectos de diseño para indicadores sociales en América Latina: función generalizada de varianza para estimadores directos provenientes de encuestas de hogares». LC/TS.2023/95 106. CEPAL - Serie Estudios Estadísticos. Santiago: CEPAL. https://repositorio.cepal.org/server/api/core/bitstreams/feb8c8be-6434-41f2-8fd9-f12f71a0739a/content.

Hansen, Morris H, William N Hurwitz, William G Madow, et al. 1953. «Sample survey methods

and theory».
Heeringa, Steven G., Brady T. West, y Patricia A. Berglund. 2017. *Applied Survey Data Analysis*. 2.ª ed. Boca Raton: Chapman & Hall/CRC. https://doi.org/10.1201/9781315153278.
Henry, Lionel, Hadley Wickham, y Winston Chang. 2024. *ggstance: Horizontal 'ggplot2' Components*. https://doi.org/10.32614/CRAN.package.ggstance.
Horvitz, Daniel G., y Donovan J. Thompson. 1952. «A Generalization of Sampling Without Replacement from a Finite Universe». *Journal of the American Statistical Association* 47 (260): 663-85. https://doi.org/10.1080/01621459.1952.10483446.
Jacob, Guilherme, Djalma Pessoa, y Anthony Damico. 2024. *convey: Income Concentration Analysis with Complex Survey Samples*. https://doi.org/10.32614/CRAN.package.convey.
Kalton, Graham, y Ismael Flores-Cervantes. 2003. «Weighting methods». *Journal of official statistics* 19 (2): 81.
Kish, Leslie. 1965. «Survey sampling.»
———. 1987. *Statistical Design for Research*. New York: John Wiley & Sons.
Kish, Leslie, y Martin R. Frankel. 1974. «Inference from complex samples». *Journal of the Royal Statistical Society, Series B* 36: 1-37.
Korn, Edward L., y Barry I. Graubard. 1999. *Analysis of Health Surveys*. New York: John Wiley & Sons. https://doi.org/10.1002/9781118032619.
Korn, Edward L, y Barry I Graubard. 1995. «Analysis of large health surveys: accounting for the sampling design». *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 158 (2): 263-95.
Kovačević, Milorad S., y David A. Binder. 1997. «Variance Estimation for Measures of Income Inequality and Polarization: The Estimating Equations Approach». *Journal of Official Statistics* 13 (1): 41-58.
Kovar, JG, JNK Rao, y CFJ Wu. 1988. «Bootstrap and other methods to measure errors in survey estimates». *Canadian Journal of Statistics* 16 (S1): 25-45.
Kreft, Ita G. G., y Jan De Leeuw. 1998. *Introducing Multilevel Modeling*. London: Sage Publications.
Langel, Matti, y Yves Tillé. 2013. «Variance Estimation of the Gini Index: Revisiting a Result Several Times Published». *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 176 (2): 521-40. https://doi.org/10.1111/j.1467-985X.2012.01048.x.
Li, Jianzhu, y Richard Valliant. 2015. «Linear Regression Diagnostics in Cluster Samples». *Journal of Official Statistics* 31 (1): 61-75. https://doi.org/10.1515/jos-2015-0003.
Long, Jacob A. 2026. *jtools: Analysis and Presentation of Social Scientific Data*. https://doi.org/10.32614/CRAN.package.jtools.
Lorenz, Max O. 1905. «Methods of Measuring the Concentration of Wealth». *Publications of the American Statistical Association* 9 (70): 209-19. https://doi.org/10.2307/2276207.
Lüdecke, Daniel, Mattan S. Ben-Shachar, Indrajeet Patil, Philip Waggoner, y Dominique Makowski. 2026. *performance: Assessment of Regression Models Performance*. https://doi.org/10.32614/CRAN.package.performance.
Lumley, Thomas. 2010. *Complex Surveys: A Guide to Analysis Using R: A Guide to Analysis Using R*. John Wiley & Sons.
———. 2026. *svyVGAM: Design-Based Inference in Vector Generalised Linear Models*. https://doi.org/10.32614/CRAN.package.svyVGAM.
Lumley, Thomas, Peter Gao, Ben Schneider, y Stas Kolenikov. 2026. *survey: Analysis of Complex Survey Samples*. https://doi.org/10.32614/CRAN.package.survey.
McCullagh, Peter, y John A. Nelder. 1989. *Generalized Linear Models*. 2.ª ed. London: Chapman & Hall. https://doi.org/10.1007/978-1-4899-3242-6.
Merlo, Juan, Basile Chaix, Henrik Ohlsson, Anders Beckman, Kristina Johnell, Per Hjerpe, Lennart

Råstam, y Klaus Larsen. 2006. «A Brief Conceptual Tutorial of Multilevel Analysis in Social Epidemiology: Using Measures of Clustering in Multilevel Logistic Regression to Investigate Contextual Phenomena». *Journal of Epidemiology & Community Health* 60 (4): 290-97. https://doi.org/10.1136/jech.2004.029454.

Molina, Elizabeth A., y Chris J. Skinner. 1992. «Pseudo-likelihood and quasi-likelihood estimation for complex sampling schemes». *Computational Statistics & Data Analysis* 13: 395-405.

Naciones Unidas. 2005. *Designing Household Survey Samples: Practical Guidelines*. Studies en Methods, Series F, No. 98. New York: United Nations.

———. 2009. *Diseño de Muestras para Encuestas de Hogares: Directrices Prácticas.* Estudios de Métodos, Serie F 98. Nueva York: Naciones Unidas. https://unstats.un.org/unsd/publication/seriesf/seriesf_98s.pdf.

———. 2021. «Report of the Intersecretariat Working Group on Household Surveys». E/CN.3/2021/16. United Nations Statistical Commission.

———. 2026. *Handbook of Surveys on Households and Individuals Foundations and Emerging Approaches*. Naciones Unidas.

Nelder, John A., y Robert W. M. Wedderburn. 1972. «Generalized Linear Models». *Journal of the Royal Statistical Society: Series A (General)* 135 (3): 370-84. https://doi.org/10.2307/2344614.

OECD. 2013. *OECD Framework for Statistics on the Distribution of Household Income, Consumption and Wealth*. Paris: OECD Publishing. https://doi.org/10.1787/9789264194830-en.

Osier, Guillaume. 2009. «Variance estimation for complex indicators of poverty and inequality using linearization techniques». En *Survey research methods*, 3:167-95. 3.

Park, Inho, Marianne Winglee, Jay Clark, Keith Rust, Andrea Sedlak, y David Morganstein. 2003. «Design Effects and Survey Planning». En *Proceedings of the Section on Survey Research Methods*, 3179-86. https://www.asasrms.org/Proceedings/y2003/Files/JSM2003-000156.pdf.

Pebesma, Edzer. 2018. «Simple Features for R: Standardized Support for Spatial Vector Data». *The R Journal* 10 (1): 439-46. https://doi.org/10.32614/RJ-2018-009.

Pebesma, Edzer, Roger Bivand, Etienne Racine, Michael Sumner, Ian Cook, Tim Keitt, Robin Lovelace, Hadley Wickham, Jeroen Ooms, y Kirill Müller. 2026. *sf: Simple Features for R*. https://doi.org/10.32614/CRAN.package.sf.

Pedersen, Thomas Lin. 2025. *patchwork: The composer of plots*. https://doi.org/10.32614/CRAN.package.patchwork.

Pfeffermann, Danny. 1993. «The Role of Sampling Weights When Modeling Survey Data». *International Statistical Review* 61 (2): 317-37. https://doi.org/10.2307/1403631.

———. 2011. «Modelling of complex survey data: Why model? Why is it a problem? How can we approach it?» *Survey Methodology* 37 (2): 115-36.

Pfeffermann, Danny, Chris J. Skinner, David J. Holmes, Harvey Goldstein, y Jon Rasbash. 1998. «Weighting for Unequal Selection Probabilities in Multilevel Models». *Journal of the Royal Statistical Society: Series B* 60 (1): 23-40. https://doi.org/10.1111/1467-9868.00106.

Posit Software, PBC. 2026. *RStudio Projects*. Posit Software, PBC. https://docs.posit.co/ide/user/ide/guide/code/projects.html.

Prener, Christopher, Timo Grossenbacher, y Angelo Zehr. 2025. *biscale: Tools and Palettes for Bivariate Thematic Mapping*. https://doi.org/10.32614/CRAN.package.biscale.

R Core Team. 2026a. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

———. 2026b. *Serialization Interface for Single Objects*. Vienna: R Foundation for Statistical Computing. https://stat.ethz.ch/R-manual/R-devel/library/base/html/readRDS.html.

Rabe-Hesketh, Sophia, y Anders Skrondal. 2006. «Multilevel Modelling of Complex Survey Data». *Journal of the Royal Statistical Society: Series A* 169 (4): 805-27. https://doi.org/10.1111/j.1467-

985X.2006.00426.x.

———. 2012. *Multilevel and Longitudinal Modeling Using Stata*. 3.ª ed. College Station, TX: Stata Press.

Rao, J. N. K., y Alastair J. Scott. 1981. «The Analysis of Categorical Data From Complex Sample Surveys: Chi-Squared Tests for Goodness of Fit and Independence in Two-Way Tables». *Journal of the American Statistical Association* 76 (374): 221-30. https://doi.org/10.1080/01621459.1981.10477633.

Rao, J. N. K., C. F. J. Wu, y K. Yue. 1992. «Some recent work on resampling methods for complex surveys». *Survey Methodology* 18: 209-17.

Rao, Jon N. K., y Alastair J. Scott. 1984. «On chi-squared tests for multiway contingency tables with cell proportions estimated from survey data». *The Annals of Statistics* 12 (1): 46-60.

Robinson, David, Alex Hayes, Simon Couch, y Emil Hvitfeldt. 2026. *broom: Convert Statistical Objects into Tidy Tibbles*. https://doi.org/10.32614/CRAN.package.broom.

Rust, Keith F., y J. N. K. Rao. 1996. «Variance Estimation for Complex Surveys Using Replication Techniques». *Statistical Methods in Medical Research* 5 (3): 283-310. https://doi.org/10.1177/096228029600500305.

Särndal, Carl-Erik, y Sixten Lundström. 2005. *Estimation in surveys with nonresponse*. John Wiley & Sons.

Särndal, Carl-Erik, Bengt Swensson, y Jan Wretman. 2003. *Model assisted survey sampling*. Springer Science & Business Media.

SAS Institute Inc. 2012. «Estimating the Standard Deviation of a Variable in a Finite Population». SAS Support, Sample 45701. https://support.sas.com/kb/45/701.html.

Schulze Waltrup, Linda, y Göran Kauermann. 2016. «A Short Note on Quantile and Expectile Estimation in Unequal Probability Samples». *Survey Methodology* 42 (1): 179-87. https://www150.statcan.gc.ca/n1/en/catalogue/12-001-X201600114545.

Shah, Babubhai V., Maurice M. Holt, y Ralph F. Folsom. 1977. «Inference about regression models from sample survey data». *Bulletin of the International Statistical Institute* 41 (3): 43-57.

Skinner, Chris J., David Holt, y Tom M. F. Smith, eds. 1989. *Analysis of complex surveys*. New York: John Wiley & Sons.

Snijders, Tom AB, y Roel Bosker. 2011. «Multilevel analysis: An introduction to basic and advanced multilevel modeling».

Solt, Frederick, y Yue Hu. 2025. *dotwhisker: Dot-and-Whisker Plots of Regression Results*. https://doi.org/10.32614/CRAN.package.dotwhisker.

Tennekes, Martijn. 2018. «tmap: Thematic Maps in R». *Journal of Statistical Software* 84 (6): 1-39. https://doi.org/10.18637/jss.v084.i06.

———. 2026. *tmap: Thematic Maps*. https://doi.org/10.32614/CRAN.package.tmap.

Thomas, D. Roland, y Jon N. K. Rao. 1987. «Small-sample comparisons of level and power for simple goodness-of-fit statistics under cluster sampling». *Journal of the American Statistical Association* 82 (398): 630-36.

Tukey, John W. 1977. *Exploratory Data Analysis*. Reading, MA: Addison-Wesley.

UNECE. 2011. *Canberra Group Handbook on Household Income Statistics*. 2.ª ed. Geneva: United Nations Economic Commission for Europe.

Valliant, Richard. 2024. *svydiags: Regression Model Diagnostics for Survey Data*. https://doi.org/10.32614/CRAN.package.svydiags.

Valliant, Richard, y Jill A. Dever. 2017. *Survey Weights: A Step-by-step Guide to Calculation*. 1 edition. Stata Press.

Valliant, Richard, Jill A. Dever, y Frauke Kreuter. 2018. *Practical Tools for Designing and Weighting Survey Samples*. Statistics for Social y Behavioral Sciences. Cham: Springer International

Publishing. https://doi.org/10.1007/978-3-319-93632-1.

Warnes, Gregory R., Ben Bolker, Thomas Lumley, Arni Magnusson, Bill Venables, Genei Ryodan, y Steffen Moeller. 2023. *gtools: Various R Programming Tools*. https://doi.org/10.32614/CRAN.package.gtools.

Wickham, Hadley. 2014. «Tidy Data». *Journal of Statistical Software* 59 (10): 1-23. https://doi.org/10.18637/jss.v059.i10.

———. 2016. *ggplot2: Elegant Graphics for Data Analysis*. 2.ª ed. Cham: Springer. https://doi.org/10.1007/978-3-319-24277-4.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2026. *tidyverse: Easily Install and Load the Tidyverse*. https://doi.org/10.32614/CRAN.package.tidyverse.

Wickham, Hadley, Mine Çetinkaya-Rundel, y Garrett Grolemund. 2023. *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*. 2.ª ed. Sebastopol, CA: O'Reilly Media. https://r4ds.hadley.nz/.

Wickham, Hadley, Winston Chang, Lionel Henry, Thomas Lin Pedersen, Kohske Takahashi, Claus Wilke, Kara Woo, Hiroaki Yutani, Dewey Dunnington, y Teun van den Brand. 2026. *ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics*. https://doi.org/10.32614/CRAN.package.ggplot2.

Wickham, Hadley, Romain François, Lionel Henry, Kirill Müller, y Davis Vaughan. 2026. *dplyr: A Grammar of Data Manipulation*. https://doi.org/10.32614/CRAN.package.dplyr.

Wilke, Claus O. 2025. *cowplot: Streamlined Plot Theme and Plot Annotations for 'ggplot2'*. https://doi.org/10.32614/CRAN.package.cowplot.

Wilkinson, Leland. 2005. *The grammar of graphics*. 2.ª ed. New York: Springer.

Wolter, Kirk M. 2007. *Introduction to Variance Estimation*. 2.ª ed. Springer Series en Statistics. New York, NY: Springer. https://doi.org/10.1007/978-0-387-35099-8.

Woodruff, Ralph S. 1952. «Confidence Intervals for Medians and Other Position Measures». *Journal of the American Statistical Association* 47 (260): 635-46. https://doi.org/10.1080/01621459.1952.10483443.

———. 1971. «A Simple Method for Approximating the Variance of a Complicated Estimate». *Journal of the American Statistical Association* 66 (334): 411-14. https://doi.org/10.1080/01621459.1971.10482279.