Falta de respuesta e imputación

Andrés Gutiérrez

Comisión Económica para América Latina y el Caribe (CEPAL) - andres.gutierrez@cepal.org

Tabla de contenidos I

Tabla de contenidos II

Representatividad y ausencia de respuesta

Introducción

- La ausencia de respuesta es un problema común en las encuestas de hogares que puede comprometer la calidad de las estadísticas.
- Existen dos tipos de ausencia de respuesta: ignorable y no ignorable, donde la primera no afecta las estimaciones una vez que se considera en un modelo, mientras que la segunda sí tiene un impacto en la calidad de las estimaciones.
- La ausencia de respuesta no ignorable puede depender de la característica de interés, lo que dificulta su manejo y puede llevar a sesgos en las estimaciones.

Introducción

- Se han desarrollado diferentes enfoques para abordar la ausencia de respuesta, como el ajuste de subcobertura, técnicas de imputación, modelos asistidos y el uso de indicadores para evaluar la efectividad de la información auxiliar.
- La literatura ha profundizado en el análisis de la ausencia de respuesta, con autores como Lumley (2010, cap. 9), Fuller (2009, cap. 5) y C. Särndal (2011), quienes proponen métodos y técnicas para mitigar los efectos negativos de este problema.
- Los enfoques van desde ajustes en los pesos muestrales hasta la utilización de modelos probabilísticos y conjuntos balanceados para lograr una representatividad adecuada en las estimaciones.

Introducción

- La implementación de estrategias efectivas para abordar la ausencia de respuesta es crucial para garantizar la precisión y confiabilidad de los resultados de las encuestas de hogares.
- El seguimiento y la planificación cuidadosa durante el diseño de la encuesta son fundamentales para minimizar los efectos negativos de la ausencia de respuesta.
- Las técnicas y enfoques utilizados deben adaptarse a las características específicas de cada encuesta y población objetivo para obtener resultados válidos y confiables.

El concepto de representatividad

El concepto de representatividad

- El concepto de representatividad en encuestas a menudo carece de una definición clara y precisa.
- Autores como Kruskal y Mosteller han tratado de explicar lo que significa una muestra representativa.
- Bethlehem, Cobben, y Schouten (2009) menciona que algunos de estos conceptos son muy vagos e imprecisos; por ejemplo:
  - Reconocimiento general de los datos.
  - Ausencia de fuerzas selectivas en la muestra.
  - Una muestra que sea una miniatura de la población.
  - Cobertura suficiente de la población,
  - Suficientemente bueno para un propósito particular.

El concepto de representatividad

En términos de notación:

- En una muestra probabilística s de tamaño n sin reemplazo de una población finita U de tamaño N, la ausencia de respuesta se modela mediante las probabilidades de respuesta.
- Cada elemento k en la población tiene una probabilidad desconocida ϕ_k de responder cuando se selecciona en la muestra.
- Esto se representa con el vector de indicadores $D = (D_1, D_2, ..., D_N)$, donde $D_k = 1$ indica que el elemento k fue seleccionado en la muestra y respondió, mientras que $D_k = 0$ indica lo contrario.
- La probabilidad de respuesta ϕ_k se define como la probabilidad de que el elemento k responda cuando es seleccionado en la muestra, es decir,

$$\phi_k = P\left(D_k = 1 \mid I_k = 1\right)$$

El concepto de representatividad

- El indicador de representatividad se define como la *ausencia de fuerzas selectivas* en una muestra.
- Una muestra es considerada "fuertemente representativa" si todas las probabilidades de respuesta de los elementos de la población son iguales y la respuesta de un elemento es independiente de los demás.
- Esto se representa mediante la igualdad de las probabilidades de respuesta para todos los elementos de la población, es decir,

$$\phi_k = P\left(D_k = 1 \mid I_k = 1\right) = \phi$$

para $k = 1, 2, \ldots, N$.

El concepto de representatividad

- La representatividad fuerte se logra cuando el mecanismo de datos faltantes es MCAR para cada variable objetivo en el estudio, evitando sesgos en los estimadores.
- En presencia de una variable auxiliar categórica X con H categorías que divide la población en estratos, es *débilmente representativa* cuando la probabilidad de respuesta promedio es igual en cada estrato.
- La probabilidad de respuesta del elemento k en el estrato h se denota como ϕ_{hk}.

El concepto de representatividad

- La representatividad débil se verifica mediante la igualdad de la probabilidad de respuesta promedio en cada estrato:

$$\bar{\phi}_h = \frac{1}{N_h} \sum_{k=1}^{N_h} \phi_{hk} = \phi$$

  para $h = 1, 2, \ldots, H$.

- Este concepto implica que la respuesta no permite distinguir entre encuestados y no encuestados utilizando solo la información de la variable auxiliar X.

Indicadores de representatividad

Indicadores de representatividad

- La falta de respuesta en las encuestas puede causar sesgos en las estimaciones si ciertos grupos de la población están sobre-representados o sub-representados, especialmente si estos grupos se comportan de manera diferente en las variables de la encuesta.
- Los Institutos Nacionales de Estadística (INE) a menudo utilizan la tasa de respuesta como indicador de calidad de la encuesta, pero una baja tasa de respuesta no garantiza una precisión deficiente en las estimaciones.

Indicadores de representatividad

Un ejemplo ilustrativo es la Encuesta Integrada de Condiciones de Vida de los Hogares en los Países Bajos en 1998, donde la tasa de respuesta aumentó significativamente después de un mes, pero esto no se tradujo en mejores estimaciones debido al aumento del sesgo en los estimadores.

- Además de la tasa de no respuesta, se requieren indicadores de calidad de la encuesta para evaluar el riesgo de estimadores sesgados.

Indicadores de representatividad

- Los Indicadores de Representatividad (Indicadores R), estudiados por Shlomo, Skinner, y Schouten (2012), son herramientas que evalúan qué tan bien la muestra de respondientes representa a la población y cómo la composición de la respuesta difiere de la población.
- El proyecto RISQ (Representativity Indicators for Survey Quality) en Europa se basa en estos indicadores para desarrollar y probar métricas que evalúen la calidad de las encuestas.

Indicadores de representatividad

- Los indicadores R determinan la desviación entre la composición de la respuesta y la muestra original, siendo útiles para medir el impacto del sesgo por ausencia de respuesta en la encuesta.
- La función de distancia asociada a los indicadores R cuantifica la discrepancia entre las probabilidades de respuesta individuales y la probabilidad de respuesta promedio, proporcionando una medida objetiva del sesgo en la composición de la respuesta.

Indicadores de representatividad

Supongamos que se conocen las probabilidades de respuesta individual $\phi_1, \phi_2, \ldots, \phi_N$ de todos los elementos de la población. Entonces la desviación estándar es

$$S(\phi) = \sqrt{\frac{1}{N-1}\sum_{k=1}^{N}\left(\phi_k - \bar{\phi}\right)^2}$$

- Nótese que $S(\phi) = 0$ si todas las probabilidades de respuesta son iguales
- El valor máximo de $S(\phi)$ es igual a 0.5.

Indicador de representatividad R

El indicador R se define como:

$$R(\phi) = 1 - 2 \cdot S(\phi)$$

$R(\phi)$ cuantifica la representatividad de la muestra de respondientes en relación con la población, con un valor máximo de 1 cuando la muestra es completamente representativa y un mínimo de -1 cuando la muestra es totalmente no representativa.

Indicador de representatividad R

La probabilidad de respuesta media estimada se calcula mediante:

$$\hat{\bar{\phi}} = \frac{1}{N} \sum_{k=1}^{n} \frac{\hat{\phi}_k}{\pi_k}$$

Y el indicador de representatividad estimado $\hat{R}(\phi)$ se obtiene con la fórmula:

$$\hat{R}(\phi) = 1 - 2\sqrt{\frac{1}{N-1} \sum_{k=1}^{n} \frac{(\hat{\phi}_k - \hat{\bar{\phi}})^2}{\pi_k}}$$

Donde: π_k es la probabilidad de inclusión de la unidad k en la muestra.

Indicador de representatividad R

El indicador $R_x(\phi)$ mide la desviación de la representatividad débil considerando las clases definidas por la variable auxiliar X. Se calcula mediante la fórmula:

$$R_x(\phi) = 1 - 2\sqrt{\frac{1}{n-1}\sum_{h=1}^{H} N_h(\bar{\phi}_h - \bar{\phi})^2}$$

- $R_x(\phi)$ mide la variación de las probabilidades de respuesta entre clases X.
- Si se supone que la variación dentro de la clase es cero en todas las clases, entonces $R_x(\phi) = R(\phi)$.

Ejemplo del Indicador de representatividad R

- En un estudio de Statistics Netherlands, se utilizó un seguimiento a gran escala entre los no encuestados en la Encuesta de Población Activa (EPA) de Holanda.
- Se abordaron dos muestras de personas no respondidas en la EPA, una con el enfoque de devolución de llamada y otra con el enfoque de preguntas básicas.

Ejemplo del Indicador de representatividad R

1. La respuesta inicial de la EPA tenía un valor de R de 0.8, indicando que no era fuertemente representativa. Tras aplicar el enfoque de devolución de llamada, la tasa de respuesta aumentó del 62.2% al 76.9%, y el valor de R aumentó a 0.85, sugiriendo una mejora en la composición de los datos.
2. El enfoque de preguntas básicas resultó en un aumento de la tasa de respuesta del 62.2% al 75.6%, pero el valor de R disminuyó a 0.78, indicando que no mejoró la composición del conjunto de datos. Los intervalos de confianza para ambos enfoques se superpusieron.

Observaciones

- La elección del conjunto de variables auxiliares para el indicador R es crucial para comparar diferentes conjuntos de datos en el tiempo o en dominios diferentes.
- Fijar el conjunto de variables auxiliares de antemano puede mitigar el sesgo, aunque puede aumentar el error estándar debido al sobreajuste.
- Utilizar técnicas de selección de modelos para encontrar el mejor modelo para cada conjunto de datos puede ser beneficioso, especialmente con muestras grandes donde más variables del modelo contribuyen significativamente.

Observaciones

- Esta metodología permite determinar si la muestra de respondientes efectivos difiere de la muestra inicial, lo que puede guiar la estrategia de recolección de datos para grupos específicos.
- Los cambios en la composición de los datos observados durante la encuesta pueden requerir ajustes en la estrategia de recolección, como el enfoque en grupos subrepresentados o la implementación de diseños receptivos.

Clasificación de la ausencia de respuesta

Clasificación de la ausencia de respuesta

- Existe una extensa literatura sobre la ausencia de respuesta, que se centra en la prevención y las técnicas de estimación para abordarla en el proceso de una encuesta.
- Los aspectos clave son la prevención antes de que ocurra la ausencia de respuesta y el ajuste adecuado para la ausencia de respuesta en el proceso de inferencia.
- Se reconocen tres tipos de ausencia de respuesta según Little y Rubin (2002), incluyendo la ausencia de respuesta completamente aleatoria (MCAR - *missing completely at random*).
- La MCAR se presenta cuando la probabilidad de que un individuo responda no depende de la característica de interés, ni de alguna otra covariable auxiliar.

Patrón de respuesta MCAR

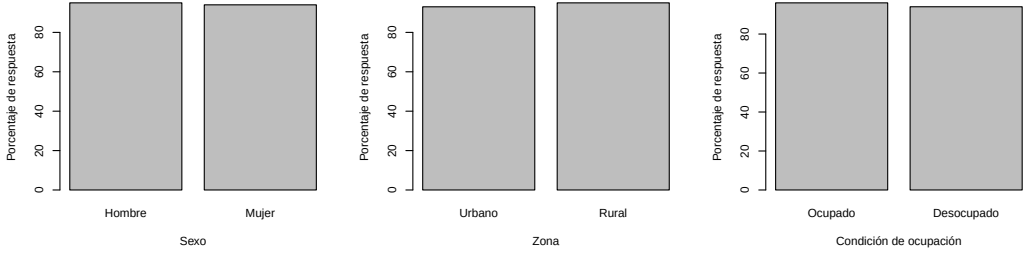

Figura 1: Patrón de respuesta MCAR

Ausencia de Respuesta Aleatoria (MAR)

- La ausencia de respuesta aleatoria (MAR) ocurre cuando la probabilidad de que alguien responda depende de ciertas covariables auxiliares, pero no de la característica de interés en sí misma.
- Por ejemplo, en una encuesta laboral, la ausencia de respuesta puede depender de variables como la edad, el sexo o el nivel económico del encuestado, pero no de su estado laboral.

Patrón de respuesta MAR

El patrón de ausencia de respuesta MAR se muestra en el gráfico donde ciertas variables auxiliares como el sexo y la zona del encuestado influyen en la tasa de respuesta, mientras que el estado de ocupación no lo hace.

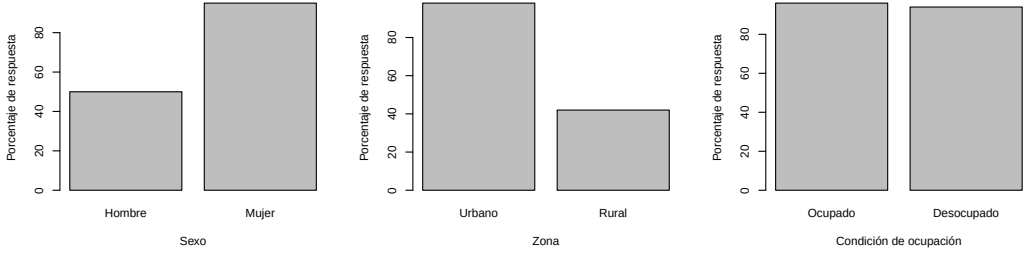

Figura 2: Patrón de respuesta MAR

Ausencia de Respuesta no Aleatoria (NMAR)

- La ausencia de respuesta no aleatoria (NMAR) ocurre cuando la falta de respuesta depende directamente de la característica de interés.
- Por ejemplo, en una encuesta laboral, la ausencia de respuesta puede depender específicamente del estado de ocupación del encuestado.

Ausencia de Respuesta no Aleatoria (NMAR)

El patrón de ausencia de respuesta NMAR se evidencia en el gráfico, donde la condición de ocupación influye en la tasa de respuesta, lo que dificulta mitigar el sesgo generado por esta falta de respuesta.

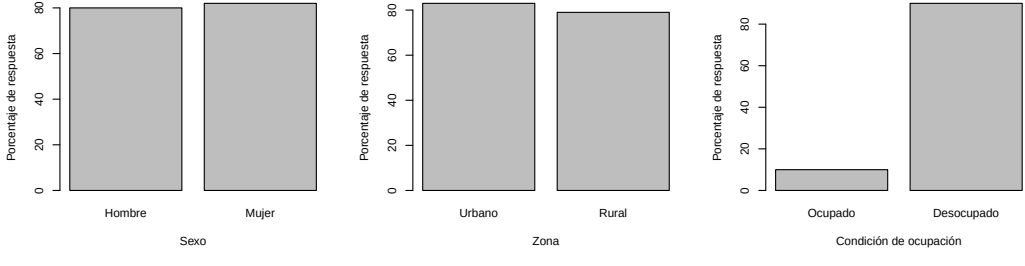

Figura 3: Patrón de respuesta MNAR

Ausencia de respuesta de registro y de unidad

Ausencia de respuesta de registro y de unidad

- La literatura especializada aborda la ausencia de respuesta desde dos enfoques complementarios: la prevención previa y las técnicas de estimación posterior a la recolección de datos.
- Si se asume que la ausencia de respuesta sigue un mecanismo completamente aleatorio (MCAR), se puede considerar únicamente a las unidades con registros completos en el proceso de inferencia, eliminando a las unidades que no respondieron.

Ausencia de respuesta de registro y de unidad

- Este enfoque puede inducir sesgos y reducir la eficiencia de la inferencia, por lo que se necesita un ajuste adicional de los factores de expansión.
- En la mayoría de las encuestas, no se asume el mecanismo MCAR y se implementan ajustes adicionales después de que ha ocurrido la ausencia de respuesta para evitar sesgos y aumentar la precisión de los estimadores.

Ausencia de respuesta de registro y de unidad

- Dos tipos de ausencia de respuesta son la ausencia de respuesta de unidad y la ausencia de respuesta por registro, que afectan a la información de la encuesta.
- Las técnicas principales para abordar la ausencia de respuesta son el ajuste a los pesos de muestreo y la imputación, que implican compensar los valores perdidos o sustituirlos por valores artificiales.
- La ausencia de respuesta puede manifestarse como la falta total de información de una unidad de observación o como información faltante en algunos registros de las unidades.
- Las causas de la ausencia de respuesta pueden variar, desde la imposibilidad de establecer contacto con el hogar hasta la sensibilidad de algunas preguntas en el cuestionario.

Ausencia de respuesta de registro y de unidad

Los siguientes son algunos puntos de vista para enfrentar la ausencia de respuesta:

- *Ignorancia*: Algunos investigadores optan por ignorar la ausencia de respuesta y realizar inferencias con los datos de las unidades respondientes, sin realizar ajustes estadísticos.
- *Prevención*: Diseñar la encuesta de manera que se reduzca la ausencia de respuesta, a través de capacitación del equipo encuestador, redacción adecuada de preguntas, longitud del cuestionario, visitas adicionales y programación de entrevistas.
- *Reacción*: Utilizar herramientas para analizar la encuesta y corregir los sesgos causados por la ausencia de respuesta, como ajustar los ponderadores de las unidades o implementar procedimientos de imputación en los registros.

Observaciones

- Ignorar la ausencia de respuesta puede llevar a subestimar datos clave en encuestas, como el ingreso medio y el número total de desempleados, lo que podría llevar a decisiones erróneas en políticas públicas.
- La ausencia de respuesta introduce sesgos significativos en las estimaciones de calidad de los resultados de la encuesta, lo que requiere una estrategia cuidadosa para abordar sus consecuencias.

Observaciones

- Aumentar el tamaño de la muestra para enfrentar la ausencia de respuesta puede resultar en una mayor homogeneidad de los respondientes, lo que aumenta el sesgo y malgasta recursos que podrían usarse de manera más efectiva para remediar la ausencia de respuesta.
- Es crucial considerar métodos adecuados para abordar la ausencia de respuesta y mitigar sus efectos en la calidad de los resultados de la encuesta.

Posibles soluciones

Posibles soluciones

- **Imputación total**: Consiste en imputar todos los valores faltantes para individuos con al menos un valor perdido, considerándose como la única forma de tratar la ausencia de respuesta.
- **Ponderación total**: se trata de ponderar cada una de las variables de interés, así sea de manera diferenciada. No se utiliza la imputación y existirán tantos conjuntos de factores de expansión como variables con valores perdidos.

Posibles soluciones

- **Eliminación total**: se trata de eliminar todos los registros con algún valor perdido y hacer el análisis con el conjunto restante de valores respondidos.
- **Enfoque combinado**: se trata de imputar únicamente en los elementos que tienen al menos un registro (no todos) perdido y modificar los factores de expansión en aquellos casos en donde hay omisión de todos los registros del cuestionario.

Notación

- En la notación de C.-E. Särndal y Lundström (2006), una muestra de unidades se denota como s, donde r representa los respondientes que han contestado al menos una de las I variables de interés.
- Las unidades que no responden a ninguna variable pertenecen al conjunto $s - r$, mientras que r_i es el conjunto de unidades que han respondido a una variable en particular. Nótese que

$$r_i \subseteq r \subseteq s$$

- Si un valor es faltante y se imputa, se denota como $\hat{y}_k$ para representar su valor imputado.

Ejemplo

La figura ilustra cómo después de la recolección de datos, algunos individuos no responden a algunas o todas las variables de la encuesta, donde las celdas en blanco indican registros respondidos y las celdas en negro indican registros no respondidos y faltantes.

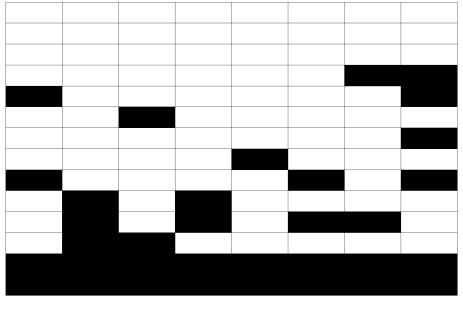

Figura 4: Un conjunto de datos después del proceso de observación.

Ejemplo

Para este ejemplo particular, se observa que:

- El número de variables de interés en la encuesta de hogares es $I = 8$.
- El número de unidades incluidas en la muestra s es $n = \#(s) = 14$.
- El número de respondientes efectivos en la primera variable es $\#(r_1) = 10$, en la segunda variable es $\#(r_2) = 9$, y así sucesivamente hasta notar que el número de respondientes efectivos en la última variable de la base de datos es de $\#(r_8) = 8$.

Imputación total

- En este enfoque se imputarían todos los valores y_k que están perdidos.
- Esto resulta en un conjunto completo de datos con valores imputados $\{y_{\circ\ k} : k \in s\}$, donde $y_{\circ\ k}$ se define como:

$$y_{\circ\ k} = \begin{cases} y_k, \text{ for } k \in r_i \\ \hat{y}_k, \text{ for } k \in s - r_i \end{cases}$$

- El estimador del total bajo este enfoque se calcula como:

$$\hat{t}_{y,\pi} = \sum_s d_k y_{\circ\ k} = \sum_{r_i} d_k y_k + \sum_{s-r_i} d_k \hat{y}_k$$

Donde d_k representa los factores de expansión de las unidades de la muestra.

Ejemplo de Imputación total

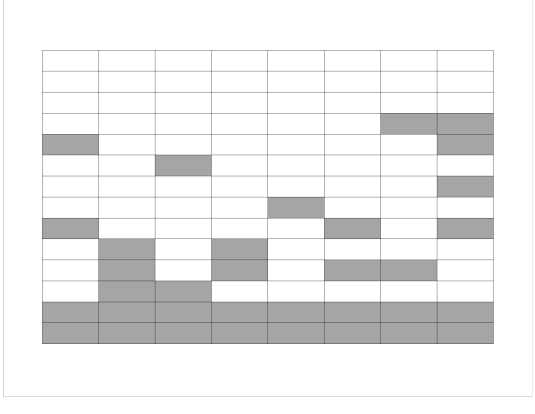

Figura 5: Imputación total: todas las unidades que no respondieron son imputadas (las celdas en gris indican los valores que fueron imputados).

Ponderación total

- Se utilizan pesos de calibración específicos $w_k = d_k F_{ik}$ para compensar la ausencia de respuesta de unidad y de registro.
- El estimador del total se calcula como $

$$\hat{t}_{y,cal} = \sum_{r_i} w_k y_k = \sum_{r_i} d_k F_{ik} y_k$$

- Si todos los r_i son diferentes, cada variable de estudio requerirá un conjunto de ponderadores diferentes.
- Este enfoque induce un número no uniforme de casos por variable y se utilizan pesos $w_k^{(i)}$ para cada variable $i \in I$ que compensan la ausencia de respuesta de la unidad.

Ejemplo de Ponderación total

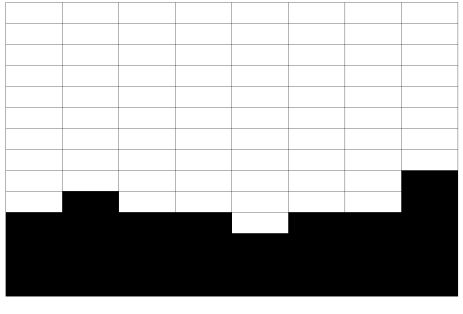

Figura 6: Ponderación total: cada variable tendrá un conjunto de pesos diferente. No se utiliza ningún método de imputación.

Eliminación total

- Se eliminan todas las unidades de la base de datos que contienen al menos un registro perdido.
- Este enfoque induce un solo conjunto de ponderadores, pero conlleva un fuerte decrecimiento en el tamaño de muestra.
- Se recomienda abstenerse de este enfoque debido a la pérdida de información, aumento del sesgo y disminución de la precisión de los estimadores.
- Solo las unidades del conjunto de respondientes efectivos en todas las variables se consideran para el análisis posterior, lo que puede generar problemas de sesgo y eficiencia estadística.

Ejemplo de Eliminación total

Figura 7: Enfoque de eliminación: únicamente se consideran las unidades que respondedieron a todas las varaibles.

Enfoque combinado

- El enfoque combinado combina la imputación para la ausencia de respuesta por registro y el ajuste a los factores de ponderación para la ausencia de respuesta por unidad.
- Utiliza información auxiliar externa y un enfoque de calibración para producir pesos finales.
- La ecuación para obtener el conjunto de datos completo es $y_{\circ\, k} = y_k$ para

$$y_{\circ\, k} = \begin{cases} y_k, \text{ for } k \in r_i \\ \hat{y}_k, \text{ for } k \in r - r_i \end{cases}$$

En donde $\hat{y}_k$ es el valor imputado.

Ejemplo de Enfoque combinado

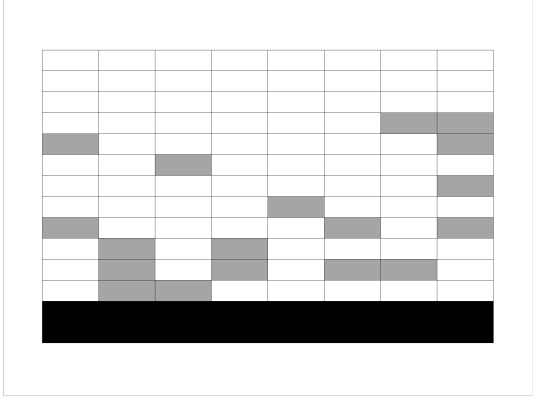

Figura 8: Enfoque combinado: las unidades que no respondieron a ningún ítem son eliminadas del análisis y los respondientes parciales son imputados.

Ausencia de respuesta de unidad

Introducción

- La información auxiliar desempeña un papel crucial en el diseño de muestreo, permitiendo la construcción de estratos, la asignación de tamaños de muestra y la definición de probabilidades de selección desiguales.
- En la etapa de estimación, la información auxiliar se utiliza para ajustar ponderaciones, imponer restricciones de consistencia y garantizar que la muestra expandida refleje adecuadamente las características poblacionales.

Introducción

La ausencia de respuesta a nivel de unidad puede causar sesgos significativos en la inferencia de las encuestas, especialmente si los respondientes y no respondientes tienen características diferentes, lo que afecta la estimación de los parámetros de interés.

Sesgo sobre los estimadores

- Considere la siguiente forma de estimar (ingenuamente) el promedio poblacional $\bar{y}_U$ mediante el estimador de Hájek

$$\tilde{y}_s = \frac{\sum_{s_r} d_k y_k}{\sum_{s_r} d_k} = \frac{\hat{t}_y}{\hat{N}}$$

Sesgo sobre los estimadores

El sesgo generado por la ausencia de respuesta en el estimador de Hájek para el promedio poblacional $\bar{y}_U$ puede cuantificarse mediante la siguiente fórmula:

$$B\left(\tilde{y}_s\right) = \frac{1}{N\bar{\phi}} \sum_U (y_k - \bar{y}_U)(\phi_k - \bar{\phi}) = \frac{Cov(\bar{y},\phi)}{\bar{\phi}} = \frac{Cor(Y,\phi)S(Y)S(\phi)}{\bar{\phi}}$$

Donde:

- ϕ_k es la probabilidad de respuesta para la unidad k.
- $Cov\left(Y,\phi\right)$ es la covarianza poblacional entre Y (la característica de interés) y ϕ (las probabilidades de respuesta).
- $Cor\left(Y,\phi\right)$ es la correlación entre Y y ϕ.
- $S\left(Y\right)$ es la desviación estándar de Y.
- $S\left(\phi\right)$ es la desviación estándar de ϕ.

Sesgo sobre los estimadores

Dado que el valor del coeficiente de correlación está restringido al intervalo $[-1, 1]$, el valor máximo del sesgo absoluto será igual a

$$|B(\tilde{y}_s)| \leq \frac{S(\phi) S(y)}{\bar{\phi}} = \frac{(1 - R(\phi)) S(y)}{2\bar{\phi}}$$

Sesgo sobre los estimadores

- El sesgo por ausencia de respuesta se puede estimar usando datos de la muestra y probabilidades de respuesta estimadas, siendo MCAR o homogeneidad poblacional condiciones en las que no habría sesgo en el estimador.
- Sarndal (2011) propone un indicador Δ_A para evaluar el sesgo por ausencia de respuesta en estimadores de calibración, donde valores grandes sugieren preferencia por ciertas variables en la calibración.
- Este indicador se basa en la distancia entre el estimador de expansión $\hat{t}_y$ y el estimador de calibración $\hat{t}_{y,cal}$, normalizado por el tamaño de la muestra,

$$\Delta_A = \frac{(\hat{t}_{y,cal} - \hat{t}_y)}{N}$$

.

Sesgo sobre los estimadores

- Δ_A se descompone en coeficiente de variación de pesos, coeficiente de determinación de regresión entre variables de estudio y calibración, y coeficiente de determinación en una regresión ponderada de desviaciones y covarianzas.

$$\frac{\Delta_A}{S_y} = cv_g \times R_{y,\mathbf{x}} \times R_{D,C}$$

- El primer factor representa la variabilidad de los pesos g_k utilizados en la calibración.
- El segundo factor, al cuadrado, es el coeficiente de determinación de una regresión múltiple entre la variable de estudio y las variables del vector de calibración.
- El último factor, al cuadrado, es la proporción de varianza explicada en una regresión ponderada que relaciona las desviaciones de las covariables con las covarianzas de la variable de estudio y las covariables.

Soluciones

- El ajuste de factores de expansión mediante modelos de *propensity score* es una opción que busca eliminar el sesgo causado por la ausencia de respuesta al incorporar información auxiliar.
- Los estimadores de calibración son otra alternativa que se basa en el paradigma de la inferencia basada en el diseño de muestreo y que también busca eliminar el sesgo en la estimación causado por la ausencia de respuesta.

Propensity Score

- Uno de los ajustes que se debe realizar en la generación de los ponderadores finales es la corrección por ausencia de respuesta. En donde

$$d_{4k} = \frac{d_{3k}}{\widehat{\phi}_k}$$

- Si el patrón de ausencia de respuesta es NMAR y no se pueden acceder a los determinantes de la respuesta, habrá sesgo constante. En este caso, el sesgo persistirá ya que la probabilidad de respuesta depende de la variable de interés, como se expresa en la ecuación:

$$\phi_k = f(\mathbf{y}_k, \beta)$$

Propensity Score

Si el patrón es MCAR o MAR y se pueden estimar las probabilidades de respuesta, el sesgo se anula en el estimador de Horvitz-Thompson. Esto se debe a que las probabilidades de respuesta se calculan utilizando las covariables $\mathbf{x}_k$, como se muestra en la ecuación:

$$\hat{\phi}_k = f(\mathbf{x}_k, \hat{\beta})$$

Propensity Score

El sesgo se anula en el estimador de Horvitz-Thompson, como se muestra en la ecuación:

$$
\begin{aligned}
E(\hat{t}_y) &= E\left(\sum_{k\in s_r} d_{3k} y_k\right) \\
&= E\left(\sum_{k\in s_r} \frac{y_k}{\pi_k \widehat{\phi}_k}\right) \\
&= E\left(E\left(\sum_{k\in U} \frac{y_k}{\pi_k \widehat{\phi}_k} I_k D_k | I_k\right)\right) \\
&= \sum_{k\in U} \frac{y_k}{\pi_k \widehat{\phi}_k} E\left(I_k\right) E\left(D_k | I_k\right) \\
&= \sum_{k\in U} \frac{y_k}{\pi_k \widehat{\phi}_k} \pi_k \phi_k = t_y
\end{aligned}
$$

Propensity Score

- Si el modelo está bien establecido, habrá concordancia directa entre $\hat{\phi}_k$ y ϕ_k, lo que anularía la última igualdad de la ecuación anterior.
- El insesgamiento del estimador está supeditado a la relación entre la probabilidad de respuesta y la ausencia de respuesta, expresado por la ecuación:

$$E(I_k D_k) = E\left(E(I_k D_k | I_k)\right) = E(I_k) E(D_k | I_k) = \pi_k \phi_k$$

Propensity Score

Es necesario que el modelo satisfaga las siguientes dos condiciones:

1. **Soporte común**: Es crucial asegurar que ninguna combinación de covariables genere un estado de forma determinística. Se expresa como:

$$0 < Pr(D_{1,k} = 1|\mathbf{x}_1) < 1$$

2. **Balanceo**: Dado que la respuesta de las unidades no proviene de un estudio aleatorizado, es esencial que la distribución de las probabilidades de respuesta sea similar entre respondientes y no respondientes.

$$\hat{\phi}_{1,k} = Pr[D_{1,k}|I_{1,k} = 1, \mathbf{x}]$$

Propensity Score

La Figura ilustra el soporte común entre respondientes y no respondientes para un modelo de propensity score

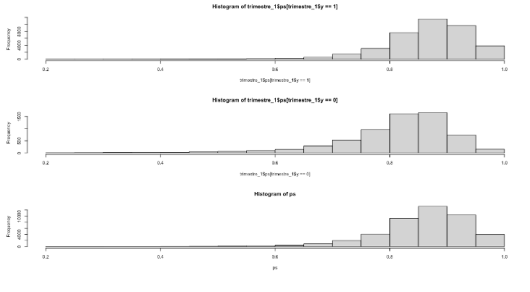

Figura 9: Distribución de las probabilidades estimadas de respuesta: respondientes (arriba), no respondientes (medio), ambos (abajo).

Propensity Score

La Figura muestra la propiedad de balanceo en el modelo, evidenciando que ambas distribuciones se alejan de los extremos (ceros y unos) y presentan una caracterización similar entre respondientes y no respondientes.

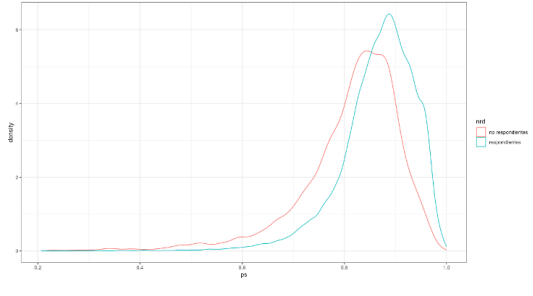

Figura 10: Balanceo entre respondientes y no respondientes

Calibración

- La calibración es un método efectivo para incorporar información auxiliar en la estimación de encuestas.
- Puede aplicarse en encuestas con información auxiliar disponible en diferentes niveles, como en muestreos de dos etapas.
- Ayuda a mejorar la precisión de los estimativos al corregir sesgos causados por la ausencia de respuesta de unidad en encuestas de hogares.
- Aunque los modelos de *propensity score* son comunes, la calibración ofrece una perspectiva novedosa para abordar estos sesgos.

Calibración

- El estimador tradicional en encuestas toma la forma:

$$\widehat{t}_y^* = \sum_{s_r} w_k \; y_k = \sum_{s_r} \frac{d_k}{\phi_k} y_k$$

- Se emplea un procedimiento en dos etapas que calcula los pesos básicos y ajusta un modelo de *propensity score* para estimar las probabilidades de respuesta.
- La práctica común es asumir que este estimador es insesgado, pero esto no es cierto debido a la imposibilidad de conocer todos los determinantes del mecanismo de respuesta.

Calibración

- Realizar ajustes basados únicamente en modelos de *propensity score* puede introducir sesgos en la estimación de los parámetros en encuestas de hogares.
- El enfoque de calibración doble se utiliza para corregir sesgos en encuestas mediante información auxiliar en dos niveles: poblacional y de muestra original.
- Se requiere información poblacional usual para calibrar factores de expansión, denotada como x_{1k}, que puede incluir variables como región, edad, sexo, y área (urbano/rural).

Calibración

- Se necesita información auxiliar en la muestra original, x_{2k}, que puede provenir de encuestas tipo panel y abarcar variables como ocupación, ingresos, u otras medidas en la primera oleada del panel.
- Este enfoque es particularmente útil en encuestas tipo panel o panel rotativo, donde se cuenta con información histórica de los individuos.

Calibración

- El enfoque de calibración doble implica calibrar los pesos en la muestra de respondientes (s_r) con la información auxiliar de la muestra original (s) y luego a nivel nacional (U) o por estratos.
- Si el mecanismo de ausencia de respuesta es MAR o MCAR, los ponderadores de calibración pueden eliminar el sesgo en las estimaciones finales si las variables que generan este mecanismo se han calibrado en alguno de los niveles mencionados.

Calibración

- C.-E. Särndal y Lundström (2006) proponen que, en la primera etapa, se calculen pesos calibrados intermedios w_{1k} sujetos a la restricción

$$\sum_s w_{1k} x_{1k} = \sum_U x_{1k}$$

.

- En la segunda etapa, se utilizan estos pesos intermedios para calcular los pesos finales de calibración w_k de la muestra de respondientes efectivos, sujetos a la restricción

$$\sum_{s_r} w_k x_{2k} = \sum_s w_{1k} x_k = \begin{pmatrix} \sum_U x_{1k} \\ \sum_{s_r} w_{1k} x_{2k} \end{pmatrix}$$

.

- La forma funcional de los pesos de calibración doble resultantes es:

$$w_k = d_k \times g_k \cong d_k \times \hat{\phi}_k$$

Calibración

- Los pesos g_k en el enfoque de calibración pueden entenderse como estimaciones de las probabilidades de respuesta ϕ_k.
- El sesgo en los estimadores sin corrección se propaga a través de las variables de la encuesta, especialmente en aquellas correlacionadas con los determinantes de la ausencia de respuesta.

Ejemplo de Calibración

Se realizó un experimento para comparar el efecto de la ausencia de respuesta en las estimaciones finales.

1. Se generó una población compuesta por individuos con diferente propensión de respuesta MCAR.
2. Se utilizaron metodologías de calibración y se comparó, de forma empírica, el efecto de la ausencia de respuesta sobre las estimaciones finales.

Ejemplo de Calibración

La figura muestra el comportamiento de ambas estimaciones: la línea roja representa el parámetro desconocido, los puntos negros son las estimaciones del estimador de calibración en cada iteración de la simulación, y los puntos grises son las estimaciones del estimador de Horvitz-Thompson.

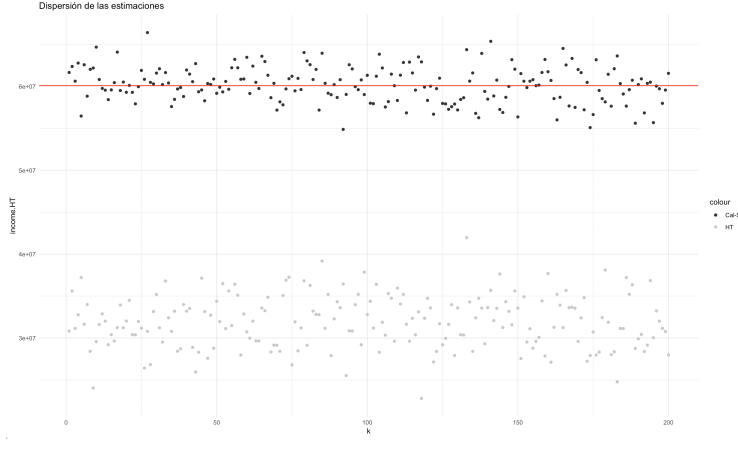

Figura 11: Estimaciones de Horvtiz-Thompson y de calibración.

Ejemplo de Calibración

- Las variables auxiliares utilizadas en la calibración deben ser capaces de explicar la variación en la probabilidad de respuesta, estar correlacionadas con las variables de interés y identificar los dominios de estimación más importantes.
- Al introducir otras covariables en la calibración, como grupo de edad, escolaridad, región y área, además de corregir el sesgo, se observa un aumento en la precisión de las nuevas estimaciones.

Ejemplo de Calibración

La figura muestra las distribuciones de tres estimadores: el estimador de Horvitz-Thompson (gris claro), el estimador de calibración con restricción de sexo (negro) y el estimador de calibración con todas las restricciones (gris oscuro).

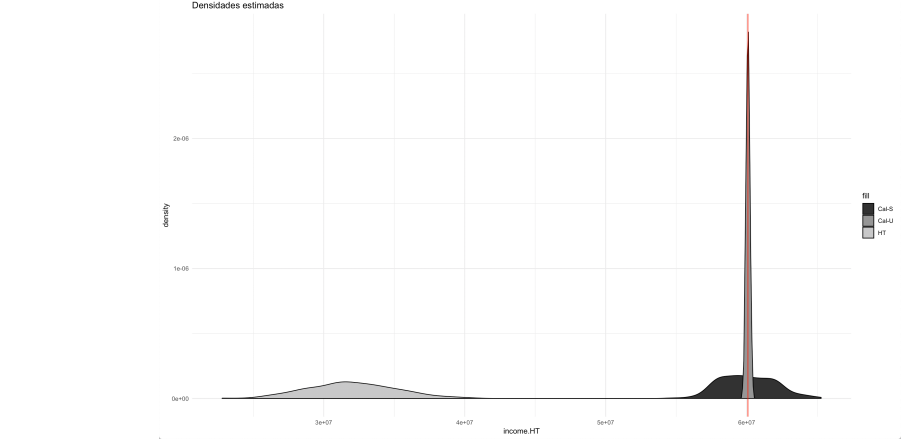

Figura 12: Distribuciones del estimador de Horvtiz-Thompson y de dos estimadores de calibración

Las consecuencias de la pandemia por COVID-19 en las encuestas de la región

- Las restricciones de movilidad debido al COVID-19 llevaron a los INE a cambiar la recolección presencial de datos por seguimiento telefónico, enfrentando desafíos en la corrección del sesgo de selección en las encuestas de hogares.
- La adopción de un seguimiento continuo a través de un panel seleccionado y el uso de contacto telefónico permitieron a los INE seguir produciendo estadísticas oficiales relevantes y oportunas durante la pandemia.

Las consecuencias de la pandemia por COVID-19 en las encuestas de la región

- La muestra maestra proporcionó un punto de partida valioso para los INE durante las restricciones de movilidad, pero no todos los hogares seleccionados proporcionaron su información de contacto telefónico.
- A pesar de ser contactables, algunos hogares no respondieron al cuestionario de la encuesta, lo que planteó desafíos adicionales para la recolección de datos durante la pandemia.

Ejemplo

- Si suponemos que la cobertura de la submuestra que sí proveyó datos de contacto asciende al 85 %
- La probabilidad de que un hogar contactado responda toda la encuesta es del 80 %
- Ajustando estos datos, solo se cuenta con el 68% de la muestra original.
- Se debe considerar que la muestra efectiva puede tener sesgos y es necesario explorar su magnitud y minimizarlos con técnicas estadísticas adecuadas.

Ejemplo

La figura muestra tres escenarios posibles encontrados por los INE al buscar sesgos en la muestra efectiva, destacando la ausencia de sesgo en uno de los diagramas y la presencia significativa de sesgo en los otros dos.

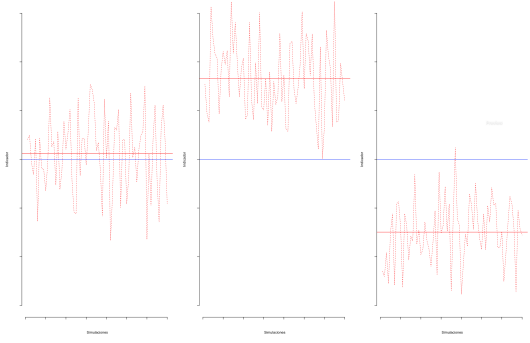

Figura 13: Distribuciones del estimador de Horvtiz-Thompson en tres escenarios de interés.

Ejemplo

La figura muestra un escenario simulado en donde se contempla el uso del estimador ajustado con la técnica de *propensity score* (línea verde) y el estimador de calibración en dos etapas (línea azul) comparado con el estimador sin ningún tipo de ajuste (línea negra).

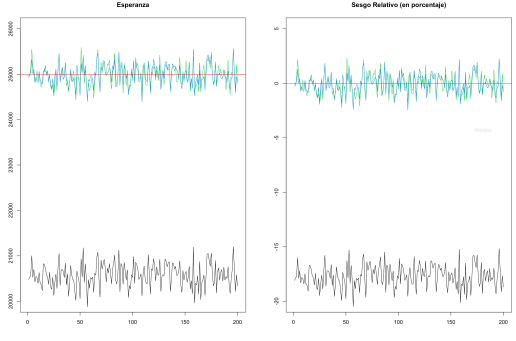

Figura 14: Distribuciones del estimador de Horvtiz-Thompson y de dos estimadores ajustados.

Observaciones

- La presencia o ausencia de sesgo determinará los siguientes pasos a seguir, siendo crucial la estrategia de levantamiento de información adoptada por los países.
- En caso de detectar sesgo, se recomienda seguir una posición parsimoniosa y ajustar los factores de expansión de manera diferencial, aprovechando el seguimiento mensual telefónico a la muestra maestra.
- La disponibilidad de covariables en la muestra maestra permitió estimar el patrón de ausencia de respuesta en la muestra de respondientes efectivos, lo que condujo a ajustar los factores de expansión en función de este patrón, mediante el factor $w_k = \frac{d_k}{\hat{\phi}_k}$.

Observaciones

- Calibración de pesos con variables demográficas y socioeconómicas minimiza sesgo y aumenta precisión.
- Uso de *propensity score* con edad, educación, área y sexo reduce sesgo en encuestas telefónicas.
- Estimadores de calibración aseguran consistencia entre cifras oficiales y encuestas telefónicas.
- Variables como edad, ingreso y ocupación eliminan sesgo en modelos de *propensity score*.

Ejemplo

Simulación

Para abordar el sesgo de selección en esta situación hipotética durante la pandemia por COVID-19, se podrían seguir los siguientes pasos:

- Supongamos un conjunto de datos artificial que define una población finita U de tamaño $N = 50000$
- supongamos que estamos interesados en observar la situación laboral (*empleada* o *desempleada*) de cada persona en U.
- Se observa esta característica de interés en dos periodos diferentes t_0 y t_1. Suponga que t_0 corresponde a un período antes de la pandemia y t_1 corresponde al período debido a la pandemia.

Simulación

- Si tuviéramos acceso a toda la población, nos encontraríamos con que en t_0, el 80% de las personas estaría empleada.
- Debido al impacto de la pandemia en los indicadores sociales (e.g., pobreza y mercado laboral), en t_1, observaríamos que muchas personas perdieron su trabajo, y la mitad de la población está desempleada.

Un vistazo a la población del ejemplo

Tabla 1: Un vistazo a la población del ejemplo. 10 primeras filas de un total de 50000.

y0	y1
Ocupado	Ocupado
Ocupado	Desocupado
Ocupado	Ocupado
Ocupado	Ocupado
Ocupado	Desocupado
Ocupado	Desocupado
Ocupado	Ocupado
Desocupado	Desocupado
Ocupado	Desocupado
Ocupado	Desocupado

Flujos netos verdaderos en la población

y_0 representa la característica de interés en el período previo a la pandemia.

Tabla 2: Flujos netos verdaderos en la población del ejemplo antes de la pandemia por COVID-19

y0	n	prop
Desocupado	10000	0.2
Ocupado	40000	0.8

y_1 representa la característica de interés en el período de la pandemia por COVID-19.

Tabla 3: Flujos netos verdaderos en la población del ejemplo en medio de la pandemia por COVID-19

y1	n	prop
Desocupado	25000	0.5
Ocupado	25000	0.5

Flujos brutos verdaderos del cambio en el estado laboral

Se puede observar, 25000 personas permanecieron ocupadas en los dos periodos, y 15000 personas cambiaron su situación laboral de ocupadas a desocupadas.

Tabla 4: Flujos brutos verdaderos del cambio en el estado laboral en la población de ejemplo.

y0	Desocupado	Ocupado
Desocupado	10000	0
Ocupado	15000	25000

La medición y observación

- La medición y observación de la situación laboral se realiza a través de una encuesta por muestreo en ambos períodos.
- Se selecciona una muestra aleatoria simple sin reemplazo de tamaño 4000 en ambos períodos.
- Para simplificar, supongamos que se pretende observar la misma muestra en ambos períodos (tipo panel).

La medición y observación

- La recolección cambió de presencial a telefónica debido a las restricciones por la pandemia.
- Se utilizan registros anteriores para obtener números de teléfono y realizar la encuesta por teléfono.
- Las tasas de muestreo difieren entre los períodos debido a la disponibilidad y respuesta de los hogares contactados.

La medición y observación

- La muestra telefónica es más pequeña (2305) que la muestra realizada cara a cara (4000).
- Los investigadores sospechan que los sesgos de selección no son despreciables en la muestra telefónica.

La medición y observación

Las tablas muestran los resultados basados en las muestras (no ponderados) para la encuesta cara a cara y la encuesta telefónica, respectivamente.

Tabla 5: Resultados observados en la muestra presencial del ejemplo.

	Estado	n	prop
Desocupado	Desocupado	820	0.205
Ocupado	Ocupado	3180	0.795

Tabla 6: Resultados observados en la muestra telefónica del ejemplo.

	Estado	n	prop
Desocupado	Desocupado	909	0.3944
Ocupado	Ocupado	1396	0.6056

Tasas de respuesta en la medición y ausencia de respuesta

- La tasa de respuesta de la muestra telefónica es del 58 %, lo que indica una baja tasa de respuesta.
- Se debe evaluar la naturaleza de la ausencia de respuesta para determinar si sigue una estructura MCAR o MAR.
- Bajo el supuesto MCAR, no se esperan patrones fuertes en las covariables entre los encuestados y no encuestados.

Tasas de respuesta en la medición y ausencia de respuesta

- Bajo el supuesto MAR, se pueden encontrar patrones fuertes en una o múltiples covariables entre los encuestados y no encuestados.
- Para verificar qué supuesto se ajusta mejor a las observaciones de la muestra seleccionada durante la pandemia, se necesita acceso a los datos de la muestra pre-pandemia para identificar a los individuos encuestados y no encuestados en la muestra actual.

Patrón de ausencia de respuesta MAR

En la Tabla, de los 2305 encuestados en la encuesta telefónica, 3.8178% estaban empleados en el período anterior, y aproximadamente 96.1822% estaban desempleados, lo que podría ser indicio de un patrón de ausencia de respuesta MAR.

Tabla 7: Proporción observada del estado de ocupación en la muestra telefónica del ejemplo.

	Estado	n	prop
Desocupado	Desocupado	88	0.0382
Ocupado	Ocupado	2217	0.9618

Patrón de ausencia de respuesta MAR

- De los 1695 individuos no encuestados en el período anterior, casi el 43.1858% estaban empleados, mientras que el 56.8142% estaban desempleados.
- Las proporciones de empleados y desempleados entre los no encuestados no son similares, lo que sugiere la presencia de un posible sesgo en la muestra telefónica.

Tabla 8: Proporción observada del estado de ocupación en la muestra telefónica del ejemplo.

	Estado	n	prop
Desocupado	Desocupado	732	0.4319
Ocupado	Ocupado	963	0.5681

Asociación en la encuesta

- Utilizar herramientas de inferencia clásica como la estadística de Ji-cuadrado de Pearson y el estadístico V de Cramer para verificar la asociación entre la respuesta en la encuesta telefónica y la situación laboral en la encuesta presencial.
- La tabla proporciona un resumen del comportamiento de la respuesta en la encuesta telefónica en relación con la situación laboral en la encuesta presencial.

Tabla 9: Asociación entre la respuesta telefónica y el estado de ocupación en el periodo anterior en la muestra del ejemplo.

Estado	Respuesta	Freq
Desocupado	0	732
Ocupado	0	963
Desocupado	1	88
Ocupado	1	2217

Prueba de bondad de ajuste de Pearson

El sistema de hipótesis es el siguiente:

- H0: Las dos variables son independientes.
- H1: Las dos variables se relacionan entre sí.

1. La prueba de bondad de ajuste de Pearson reveló una correlación significativa
2. La estadística V de Cramer mostró una asociación considerable entre estas dos variables, con un valor cercano a 0.5.

Calulando ϕ_k

Para simplificar el ejemplo, asumimos que la probabilidad de ser un encuestado depende de la situación laboral anterior. De esta forma, ϕ_k puede escribirse como una función de ese estado laboral anterior, incluido en las covariables $\mathbf{z}$.

$$\phi_k = f(\mathbf{z}_k, \beta)$$

- Se identificó una asociación significativa entre la respuesta en la encuesta telefónica y la situación laboral en la encuesta presencial, indicando un enfoque de respuesta MAR.
- Se ajustó un modelo de *propensity-score* para estimar las probabilidades de respuesta ($\widehat{\phi}_k$) basadas en las covariables disponibles, especialmente la situación laboral anterior.

Histograma de los *propensity-score*

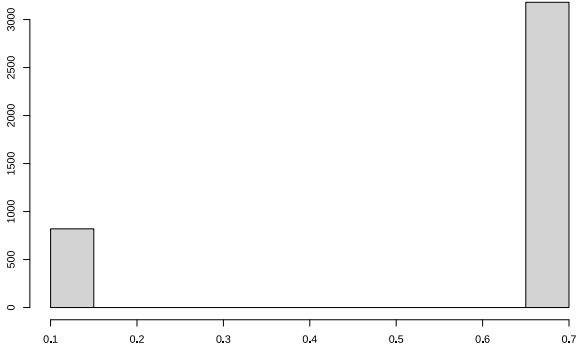

Figura 15: Histograma de los *propensity-score*

Estimación del número de empleados

- Utilizando los datos telefónicos y el nuevo conjunto de ponderaciones d_{4k}, ajustado por el puntaje de propensión estimado, tenemos que el número estimado de empleados en el período COVID es $\widehat{t}_y = \sum_{k \in s_{ER}} d_{4k} y_{1k} = 24970$.
- También se considera la posibilidad de calibrar los pesos en la muestra telefónica utilizando información auxiliar disponible en la muestra presencial y a nivel nacional para eliminar el sesgo, sujeto a

$$\sum_{s_0} w_{0k} x_k = \sum_{U} x_k = \mathbf{t_x}$$

Donde $\mathbf{t_x}$ puede representar conteos nacionales provenientes de censos o proyecciones demográficas.

Estimación del número de empleados

Estos pesos intermedios w_{0k} deben utilizarse para calcular los pesos finales de calibración w_{1k} de la muestra de encuestados efectivos que están sujetos a la siguiente restricción:

$$\sum_{s_0} w_{1k} x_k = \begin{pmatrix} \sum_U x_k \\ \sum_{s_1} w_{0k} z_k \end{pmatrix} = \begin{pmatrix} \mathbf{t_x} \\ \hat{\mathbf{t}}_\mathbf{z} \end{pmatrix}$$

En donde, $\hat{\mathbf{t}}_\mathbf{z}$ representa las cifras estimadas provenientes de la encuesta presencial.

Estimación del número de empleados

- El estimador de calibración se puede escribir de la siguiente manera:

$$\hat{t}_y^{cal} = \sum_{k \in s_1} w_{1k} y_{1k}$$

- Después de realizar la calibración en dos etapas, obtenemos que el número estimado de empleados en el período COVID es $\hat{t}_y^{cal} = \sum_{k \in s_1} w_{1k} y_{1k} = 24813$.
- La forma funcional de los pesos de calibración doble resultantes se puede expresar como:

$$w_{1k} = d_k \times g_{0k} \times g_{1k} \cong d_k \times \hat{\phi}_k$$

Histograma de los pesos ajustados

La figura muestra el histograma de los puntajes de propensión pronosticados, los cuales toman dos valores (0.6928 y 0.11), uno para cada categoría del estatus laboral en la encuesta cara a cara.

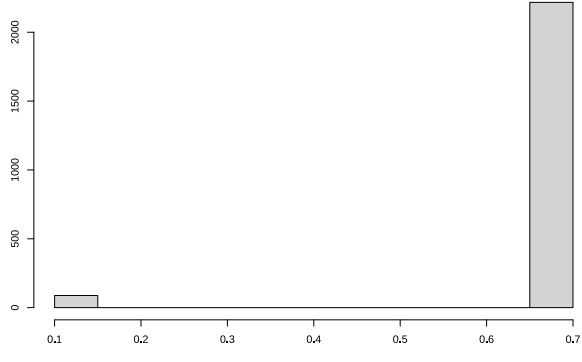

Figura 16: Histograma de los pesos ajustados

Observaciones

- Es importante notar que si no se considera el mecanismo de ausencia de respuesta, se pueden obtener estimaciones engañosas y sesgadas.
- El estimador sesgado $\hat{t}_y^{exp} = \sum_{k \in s_1} d_{3k} y_{1k}$ puede conducir a estimaciones incorrectas, donde d_{1k} representa los pesos muestrales no ajustados de la encuesta telefónica.
- Bajo este escenario, el número estimado de empleados en el período COVID es $\hat{t}_y^{exp} = \sum_{k \in s_1} d_{1k} y_{1k} = 19718$.

Ausencia de respuesta por registro

Introducción

- El diseño y ejecución de una encuesta puede presentar situaciones que generen sesgos en las estimaciones finales, desde la recolección hasta el análisis de datos.
- Es responsabilidad del estadístico identificar y abordar todas las instancias que podrían causar sesgos, minimizando tanto el error humano como el error estadístico en todas las etapas de la encuesta.
- La imputación implica estimar valores plausibles para los datos faltantes basándose en la información disponible en la base de datos, una práctica que se ha utilizado desde el siglo XIX.
- La imputación se utiliza no solo para tratar la ausencia de respuesta, sino también para identificar y manejar valores atípicos en la base de datos, contribuyendo así a obtener resultados más confiables en el análisis de la encuesta.

Modelos para la imputación

- La imputación es un proceso esencial para reemplazar valores faltantes en una base de datos, permitiendo un análisis completo y robusto de los datos.
- En la imputación, se buscan donantes apropiados en la muestra que compartan características similares con los individuos que no respondieron, utilizando la información de estos donantes para completar los valores faltantes.
- La imputación introduce un nuevo elemento de error llamado error de imputación, debido a la incertidumbre que surge al reemplazar valores no observados con información plausible.

Métodos de Imputación más Usados en Encuestas

- Imputación promedio (*mean value imputation*): Consiste en reemplazar los valores faltantes por la media de la variable en un subconjunto de datos apropiado, como las Unidades Primarias de Muestreo (UPM).
- Imputación por paquete caliente (*hot deck imputation*): Se utiliza un donante respondiente de la misma encuesta para reemplazar los valores faltantes, basándose en la información del individuo escogido previamente.

Métodos de Imputación más Usados en Encuestas

- Imputación por paquete frío (*cold deck imputation*): Similar al paquete caliente, pero los donantes son individuos de encuestas anteriores en lugar de la misma encuesta actual.
- Imputación estadística basada en modelos: Se emplean modelos estadísticos donde la variable dependiente es la que se desea imputar, utilizando las covariables del resto de los datos para hacer la predicción del valor faltante.

Métodos de Imputación más Usados en Encuestas

- Imputación de la unidad completa: Se refiere a la imputación de toda la información de un individuo cuando no hay datos disponibles para él.
- Imputación de registros: Se da cuando no todos los valores de un individuo están presentes, pero algunos sí, por lo que se imputan los valores faltantes a nivel de los registros.
- Los grupos de imputación $g = 1, \dots, G$ se utilizan para realizar la imputación, donde la unión de $s_1, \dots, s_G$ forma la muestra completa s.
- Se pueden usar diferentes métodos de imputación para cada grupo, pero dentro de cada grupo se debe usar el mismo método para garantizar la consistencia.

Métodos de Imputación más Usados en Encuestas

Considerar una jerarquía de métodos de imputación puede ser útil cuando la disponibilidad de covariables es limitada, utilizando métodos más sofisticados para grupos con más información auxiliar y métodos más simples para grupos con menos información.

Imputación por regresión

En este método determinístico, el valor imputado para el valor faltante y_k se calcula utilizando una regresión lineal.

$$\hat{y}_k = \mathbf{x}_k \hat{\beta}_i$$

Donde,

$$\hat{\beta}_i = \left(\sum_{r_i} a_k \mathbf{x}_k \mathbf{x}'_k \right)^{-1} \sum_{r_i} a_k \mathbf{x}_k y_k$$

El vector de coeficientes de regresión $\hat{\beta}_i$ se produce a partir de un ajuste de regresión múltiple utilizando los datos $(y_k, \mathbf{x}_k)$ disponibles para cada unidad $k \in r_i$ con pesos a_k especificados adecuadamente.

Imputación de razón

Un caso especial del anterior método se da cuando solo se tiene acceso a una sola covariable (positiva) $\mathbf{x}_k = x_k$, y definiendo $a_k = \frac{1}{x_k}$. En este caso, la estimación del coeficiente de regresión será

$$\hat{\beta}_i = \frac{\sum_{r_i} y_k}{\sum_{r_i} x_k} = R_i$$

Y por tanto, la imputación para el valor faltante se convierte en

$$\hat{y}_k = x_k \hat{\beta}_i = x_k \frac{\sum_{r_i} y_k}{\sum_{r_i} x_k} = x_k R_i$$

Este método se utiliza a menudo cuando la misma variable se mide en dos momentos diferentes en la misma encuesta.

Imputación de promedio

El caso más sencillo de la imputación por regresión se da cuando $a_k = x_k = 1$ para todo $k \in r_i$. En este escenario, el valor imputado se convierte en

$$\hat{y}_k = \frac{\sum_{r_i} y_k}{\sum_{r_i} 1} = \bar{y}_{r_i}$$

Por lo tanto, todos los valores faltantes recibirán el mismo valor imputado, que es justamente el promedio de la variable en el conjunto de respondientes.

Método del vecino más cercano

- Utiliza valores similares de una variable x para imputar valores faltantes de otra variable y mediante la búsqueda del "vecino" más cercano con valores similares en x.
- La imputación se realiza mediante la fórmula $\widehat{y}_k = y_{l(k)}$, donde $l(k)$ es el "elemento donante" determinado al minimizar una ecuación de distancia, como $D_{lk} = |x_k - x_l|$ para una sola covariable de imputación x_k.
- En casos con múltiples covariables de imputación, la distancia se puede calcular como $D_{lk} = \left(\sum_{j=1}^{J} h_j(x_{jk} - x_{jl})^2\right)$, donde h_j pondera adecuadamente cada covariable en la matriz de imputación.

Imputación por paquete caliente (Hot-Deck)

- Este método se utiliza cuando no es posible establecer una relación fuerte entre la variable de interés y y las covariables $\mathbf{x}$, y tampoco es posible validar los supuestos de modelización necesarios para otros métodos.
- En este enfoque, el valor imputado para el individuo k se obtiene de un donante seleccionado aleatoriamente del conjunto de datos de la variable de interés, es decir, $\widehat{y}_k = y_{l(k)}$.
- Este método no se recomienda cuando existen mejores opciones disponibles, ya que no se utiliza información auxiliar para determinar un sustituto adecuado.

Imputación múltiple

- Este método se utiliza cuando se dispone de información auxiliar que permite establecer mejores modelos entre las covariables y la variable de interés.
- Consiste en completar los datos utilizando información de respondientes en la encuesta (o encuestas anteriores en un diseño rotativo) y datos a nivel de la población para predecir los valores faltantes mediante un modelo de regresión probabilístico.

Imputación múltiple

- El modelo se formula como $y_k = f(\mathbf{x_k}, \beta) + \varepsilon_k$, donde ε_k es un término de error aleatorio.
- Se generan múltiples realizaciones de la variable de interés para los registros faltantes al simular M valores del término de error. Posteriormente, se obtienen M conjuntos de datos completos y se promedian las estimaciones de interés para obtener una estimación puntual más precisa.

Ejemplo de imputación en una encuesta de ingresos y gastos

- El levantamiento de encuestas de ingresos y gastos implica visitas masivas al hogar en múltiples ocasiones.
- Se solicita a los respondientes que completen cuestionarios sobre gastos e ingresos durante al menos dos semanas.
- La colaboración activa de todos los miembros del hogar es crucial para el éxito de la encuesta.
- En ocasiones, a pesar del seguimiento exhaustivo, algunas categorías de la encuesta pueden quedar sin información, lo que resulta en celdas vacías o ausencia de respuesta.

Ejemplo de imputación en una encuesta de ingresos y gastos

- El ejemplo ilustra el proceso de imputación en una encuesta de ingresos y gastos.
- Se debe identificar qué covariables están relacionadas con las variables a imputar.
- Primero se imputan todas las covariables y se reemplazan los valores faltantes.
- Se consideran variables como
    1. Tamaño del hogar.
    2. Número de hombres y mujeres dentro del hogar.
    3. Número de niños y adultos en el hogar.
    4. Edad del jefe de hogar.
    5. Estado de ocupación del jefe de hogar.
    6. Grado educativo más alto del jefe de hogar.
    7. Número de personas empleadas en el hogar.

Proceso de imputación

- Primero se imputan los ingresos, que son la principal covariable del gasto y el consumo en el ejemplo.
- Luego se imputan los filtros, que son preguntas sobre bienes o servicios adquiridos por el hogar.
- El tercer paso es la imputación de los valores de gasto anualizados en cada unidad.
- Este proceso metodológico es seguipor por Hayes y Watson (2009) y Sun (2010) en el *Australian Bureau of Statistics* para imputación en la encuesta *Household, Income and Labour Dynamics in Australia (HILDA)*

Imputación del ingreso

- Las fuentes de ingreso en un hogar incluyen el trabajo, la propiedad de activos, la producción de servicios para consumo propio y las transferencias gubernamentales.
- Los ingresos son un predictor importante de los gastos según la teoría y la evidencia empírica.
- La técnica de imputación del ingreso puede basarse en un modelo de vecino más cercano con regresión, donde se define un modelo lineal y se estiman los coeficientes de regresión para pronosticar los ingresos faltantes.

Imputación del ingreso

La imputación se realiza mediante la siguiente ecuación:

$$\tilde{y}_k = \mathbf{x}_k \hat{\beta}_i$$

Donde:

- $\tilde{y}_k$ es el ingreso imputado para la unidad k.
- $\mathbf{x}_k$ son las covariables asociadas a la unidad k.
- $\hat{\beta}_i$ son los coeficientes de regresión estimados para el modelo lineal, dados por

$$\hat{\beta}_i = \left(\sum_{r_i} a_k \mathbf{x}_k \mathbf{x}_k' \right)^{-1} \sum_{r_i} a_k \mathbf{x}_k y_k$$

Imputación del ingreso

- El vector de coeficientes de regresión $\widehat{\beta}_i$ se obtiene mediante un ajuste de regresión múltiple utilizando los datos $(y_k, \mathbf{x}_k)$ para cada unidad $k \in r_i$ con pesos a_k adecuadamente especificados.
- Las covariables en el vector $\mathbf{x}_k$ deben incluir información como:
  - *Composición del hogar*: número de adultos, número de niños, número de hombres, número de mujeres, edad adulta media, edad media de los niños, etc.
  - *Ocupación y fuerza de trabajo*: situación laboral del jefe de hogar, número de personas empleadas, número de desempleados en el hogar, etc.
  - *Calidad de la vivienda*: índice de hacinamiento, material de las paredes, fuente de agua potable, etc.
  - *Ubicación del hogar*: municipalidad y provincia.

Imputación del ingreso

- Se asume que valores similares de las predicciones del modelo lineal $\tilde{y}$ producirán valores similares en las observaciones del ingreso y.
- Esto permite "pedir prestado" un valor real de ingreso y para imputar el valor faltante con la información de un vecino que tenga valores similares en las predicciones $\tilde{y}$ del modelo lineal.
- El valor imputado para la unidad k está dado por: $\widehat{y}_k = y_{l(k)}$. Donde $l(k)$ es el "elemento donante".

Imputación del ingreso

El donante es determinado mediante la minimización de una medida simple de distancia entre todos los posibles donantes l y la unidad k. Esta distancia se calcula como:

$$D_{lk} = |\tilde{y}_k - y_l|$$

El donante l del elemento k será aquel hogar en el conjunto r_i con la menor distancia D_{lk}.

Imputación del ingreso

- Se utiliza el valor real de ingreso de un vecino similar para imputar el valor faltante, basado en la predicción del modelo lineal. Esto se expresa con la ecuación $\widehat{y}_k = y_{l(k)}$.
- La imputación se realiza minimizando la distancia entre las predicciones del modelo y los valores reales de ingreso de los posibles donantes. La distancia se calcula como $D_{lk} = |\widetilde{y}_k - y_l|$.
- Se selecciona como donante al hogar con la menor distancia, siempre y cuando esté ubicado en la misma provincia que la unidad faltante.

Resultados de la imputación

La figura muestra un diagrama de caja junto con el histograma de los ingresos (antes de la imputación), así como la relación lineal entre los valores pronosticados derivados del modelo y los valores imputados tomados de los donantes.

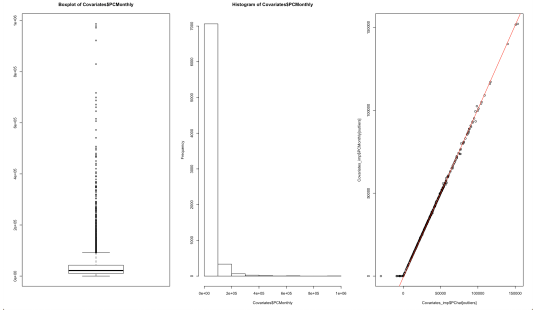

Figura 17: Distribución de los ingresos (izquierda y centro) y Relación entre los valores predichos e imputados para los hogares con datos de ingresos faltantes (derecha).

Imputación del filtro

- Se utiliza la información imputada de las covariables relacionadas con el gasto para imputar el consumo o adquisición de bienes o servicios, conocido como el filtro en la encuesta.
- Los filtros en las encuestas de ingresos y gastos preguntan si el hogar consumió o adquirió un bien o servicio específico, siendo la respuesta dicotómica (sí o no). Un modelo de regresión logística es apropiado para modelar la probabilidad de consumo.
- La probabilidad de consumo $p_k = Pr(Filtro_i = 1)$ puede ser estimada mediante un modelo de regresión logística:

$$\widetilde{p}_k = \text{logit}^{-1}(\mathbf{x}_k\widehat{\beta}_i) = \frac{\exp(\mathbf{x}_k\widehat{\beta}_i)}{1+\exp(\mathbf{x}_k\widehat{\beta}_i)}$$

Imputación del filtro

- Las covariables incluidas en la matriz $\mathbf{x}$ pueden ser las mismas utilizadas para la imputación de ingresos.
- Para imputar los valores faltantes en el filtro, se asume que valores similares de $\tilde{p}$ producirán valores de filtro similares.
- Se utiliza la técnica de vecino más cercano para imputar el filtro faltante, tomando el valor del vecino con un valor similar de $\tilde{p}$.
- La ecuación para imputar el filtro es $Filtro_k = Filtro_{l(k)}$, donde $l(k)$ es el elemento donante determinado por la minimización de la distancia $D_{lk} = |\tilde{p}_k - p_l|$.

Imputación del filtro

- Por regla general, los donantes utilizados para la imputación deben estar en la misma provincia que la unidad con el valor faltante en el filtro de compra.
- En el ejemplo del artículo arroz, se observa que la mayoría de hogares probablemente hayan comprado este artículo, lo que se refleja en la distribución sesgada de las probabilidades estimadas de compra hacia el valor uno en la regresión logística.
- La distribución de los valores imputados para el filtro de compra de arroz también estará cargada hacia el uno, reflejando la realidad de la compra masiva de este artículo esencial.
- Para artículos de bajo consumo, la distribución de las probabilidades estimadas de compra estará más sesgada hacia el valor cero, lo que se refleja en la distribución de los valores imputados para estos artículos.

Distribución de las probabilidades

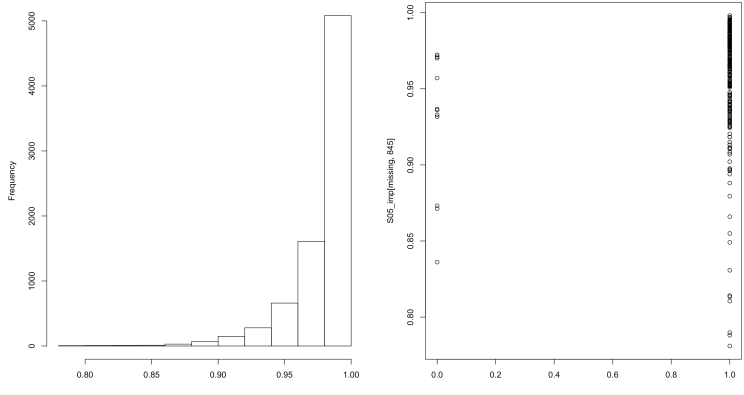

Figura 18: Distribución de las probabilidades estimadas de compra de arroz (izquierda) y valores imputados para los hogares con valores faltantes en el filtro (derecha).

Distribución de las probabilidades

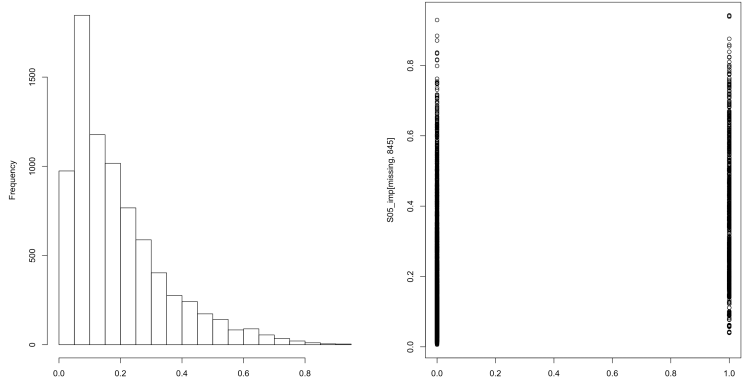

Figura 19: Distribución de la probabilidad estimada de compra de un artículo de bajo consumo (izquierda) y sus valores imputados para los hogares que no respondieron el filtro (derecha).

Imputación del gasto

- Los hogares con valor imputado de filtro cero tendrán automáticamente un cero imputado como la cantidad de dinero gastado en ese artículo, ya que no compraron ni produjeron el artículo en el periodo de referencia.
- Para unidades con valor de filtro diferente de cero, se identificará un donante para la imputación del gasto. El grupo de donantes está limitado a aquellos con un valor de gasto distinto de cero en el artículo específico.

Imputación del gasto

La técnica del vecino más cercano con el método de regresión se puede considerar para la imputación del gasto, utilizando un modelo lineal que incluya covariables como la composición del hogar, el estado de ocupación y fuerza de trabajo, la calidad de la vivienda, la ubicación del hogar y los ingresos.

Distribución de los gastos

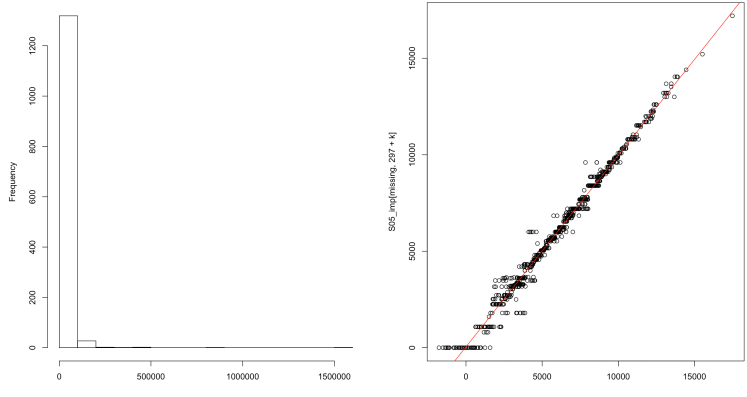

Figura 20: Distribución de los gastos imputados sobre el salmón (izquierda) y relación entre los valores predichos e imputados para los hogares con valores faltantes en el gasto (derecha).

Consideraciones sobre la imputación múltiple

Consideraciones sobre la imputación múltiple

- Antes de elegir un método de imputación, es crucial considerar cómo afectará a las propiedades estadísticas de los estimadores en las encuestas de hogares.
- La imputación múltiple puede modificar las propiedades estadísticas de los estimadores. Subestimar la variación de las estimaciones puede ser un error grave, ya que afecta la cobertura nominal de los intervalos de confianza, las pruebas de hipótesis y el cálculo de los valores p.

Consideraciones sobre la imputación múltiple

- Por ejemplo, en un modelo de regresión simple con una muestra aleatoria s y dos variables X y Y, los errores ε_k tienen distribución normal con $E(\varepsilon_k) = 0$ y $Var(\varepsilon_k) = \sigma^2$ para todo $k \in s$.
- Supongamos que la variable dependiente de interés solo está disponible para n_1 individuos, mientras que no hay datos disponibles para los n_0 individuos restantes, con $n_1 + n_0 = n$. Además, se asume que la covariable X está disponible para todos los individuos en la muestra.

Imputación Múltiple (Rubin 1987)

- La imputación múltiple (Rubin 1987) es valiosa principalmente para estimar los errores estándar.
- Ignorar la naturaleza estocástica de los valores imputados puede resultar en estimaciones de varianza subestimadas.
- El enfoque implica generar $M > 1$ conjuntos de valores para los registros faltantes, lo que permite capturar la incertidumbre asociada con la imputación y obtener estimaciones más precisas de la varianza.

Imputación Múltiple (Rubin 1987)

- El valor *imputado* corresponderá al promedio de esos M valores.
- El modelo final de imputación (para los valores faltantes) toma la siguiente forma:

$$\dot{y}_i = \dot{\beta} x_{i_{(missing)}} + \dot{\varepsilon}_i$$

Dos maneras de realizar la imputación

- **Ingenua**: en este escenario, el valor imputado para el registro faltante toma la siguiente forma:

$$\dot{y}_i = \hat{\beta} x_{i_{(missing)}}$$

  Esta clase de imputación carece de aleatoriedad y por tanto, la varianza de β será subestimada.

- **Múltiple**: en este caso, se tiene en cuenta el término de error en la generación de los valores imputados, tales que

$$\dot{y}_i = \dot{\beta} x_{i_{(missing)}} + \dot{\varepsilon}_i$$

  Es posible realizar la imputación múltiple de forma frecuentista o bayesiana.

Imputación Múltiple Frecuentista

- Seleccionar M muestras *bootstrap*.
- Para cada una se estiman los parámetros β y σ
- Generar $\dot{y}_i$.
- Promediar los M valores y se imputa el valor faltante.

Imputación Múltiple Bayesiana.

- En el enfoque bayesiano, se definen distribuciones posteriores para los parámetros β y σ.
- Se generan M valores de estos parámetros a partir de sus distribuciones posteriores, lo que lleva a la generación de M valores de $\dot{y}_i$ para los registros faltantes.
- Finalmente, se calcula el promedio de los M valores generados para imputar el valor faltante.

Imputación Múltiple para la Estimación del Parámetro

- La esperanza estimada de un parámetro β utilizando imputación múltiple se calcula como

$$E(\hat{\beta}|Y_{obs}) = E(E(\hat{\beta}|Y_{obs}, Y_{mis})|Y_{obs})$$

- Esta expresión es estimada por:

$$\bar{\hat{\beta}} = \frac{1}{M} \sum_{m=1}^{M} \hat{\beta}_m$$

Imputación Múltiple para la Estimación de la Varianza

- La varianza estimada de β utilizando imputación múltiple se calcula como

$$V(\hat{\beta}|Y_{obs}) = E(V(\hat{\beta}|Y_{obs}, Y_{mis})|Y_{obs}) + V(E(\hat{\beta}|Y_{obs}, Y_{mis})|Y_{obs})$$

- La primera parte de la anterior expresión se estima como

$$\bar{U} = \frac{1}{M}\sum_{m=1}^{M} Var(\hat{\beta})$$

- El segundo término se estima como

$$B = \frac{1}{M-1}\sum_{m=1}^{M}(\hat{\beta}_m - \bar{\hat{\beta}})^2$$

Imputación Múltiple para la Estimación de la Varianza

Es necesario tener en cuenta un factor de corrección (puesto que M es finito). Por tanto, la estimación del segundo término viene dada por la siguiente expresión:

$$\left(1 + \frac{1}{M}\right) B$$

Por tanto, la varianza estimada es igual a:

$$\hat{V}(\hat{\beta}|Y_{obs}) = \bar{U} + \left(1 + \frac{1}{M}\right) B$$

Medición de la Pobreza Usando Imputación Múltiple

- Para medir la pobreza usando imputación múltiple, primero se establece un modelo sobre los ingresos y se generan Q posibles valores para cada individuo que no respondió.
- Se estima la proporción acumulada de ingresos por debajo de un umbral α para cada conjunto de datos generado.

$$\hat{F}_{\alpha}^{q} = \frac{1}{N} \sum_{k \in s} w_k \left(\frac{l - y_k}{l} \right)^{\alpha} I(y_k < l) \qquad q = 1, \dots, Q.$$

- El estimador final de la proporción acumulada

$$\tilde{F}_{\alpha} = \frac{1}{Q} \sum_{q=1}^{Q} \hat{F}_{\alpha}^{q}$$

Medición de la Pobreza Usando Imputación Múltiple

- La varianza del estimador final tiene dos componentes:

$$\hat{V}(\tilde{F}_\alpha) = \frac{1}{Q}\sum_{q=1}^{Q}\hat{V}(\hat{F}_\alpha^q) + \left(1+\frac{1}{Q}\right)\frac{1}{Q-1}\sum_{q=1}^{Q}(\hat{F}_\alpha^q - \tilde{F}_\alpha)^2$$

- Se puede estimar $\hat{V}(\hat{F}_\alpha^q)$ de cada conjunto de datos utilizando técnicas de Jackknife y conglomerados.
- La imputación múltiple permite aproximar con precisión los valores faltantes utilizando información auxiliar, manteniendo un sesgo nulo o despreciable en las estimaciones poblacionales y la confiabilidad de la estrategia de muestreo.

Simulación empírica

Simulación empírica

Se un conjunto de $n = 100$ datos con una pendiente $\beta = 100$ y con una dispersión de $\sigma = 2$. A su vez, el conjunto de datos tendrá $n_0 = 40$ valores faltantes

Tabla 10: Ejemplo de un conjunto de datos con valores faltantes.

x	y	faltantes	y (no imputado)
11	991	Si	NA
12	1282	Si	NA
12	1164	No	1164
12	1217	No	1217
13	1325	No	1325
11	1086	No	1086
12	1210	Si	NA
13	1272	Si	NA
15	1459	Si	NA
11	1182	No	1182

Relación de la variable de interés para los datos completos

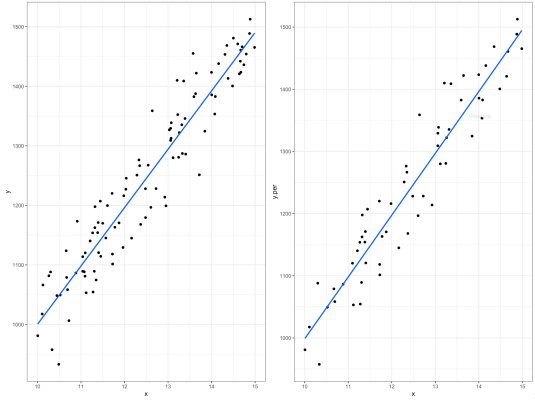

Figura 21: Relación de la variable de interés para los datos completos (izquierda) y con ausencia de valores (derecha).

Imputación Ingenua

Tabla 11: Ejemplo de un conjunto de datos con valores imputados ingenuamente.

x	y (original)	faltantes	y (imputado)
11	991	Si	1047
12	1282	Si	1221
12	1164	No	1164
12	1217	No	1217
13	1325	No	1325
11	1086	No	1086
12	1210	Si	1221
13	1272	Si	1290
15	1459	Si	1485
11	1182	No	1182

Imputación Ingenua

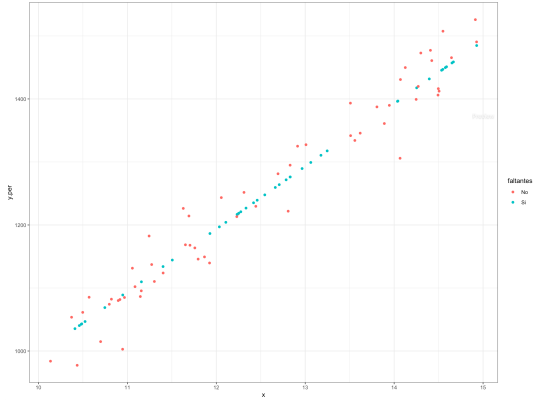

Figura 22: Relación de la variable de interés con la covariable auxiliar para el enfoque de imputación ingenua.

Imputación múltiple con el enfoque *bootstrap*.

Tabla 12: Ejemplo de un conjunto de datos con múltiples (3) valores imputados.

x	y (original)	faltantes	y1 (imputado)	y2 (imputado)	y3 (imputado)
11	991	Si	1047	950	1040
12	1282	Si	1221	1254	1198
12	1164	No	1164	1164	1164
12	1217	No	1217	1217	1217
13	1325	No	1325	1325	1325
11	1086	No	1086	1086	1086
12	1210	Si	1252	1199	1198
13	1272	Si	1304	1302	1292
15	1459	Si	1485	1493	1478
11	1182	No	1182	1182	1182

Imputación múltiple con el enfoque *bootstrap*.

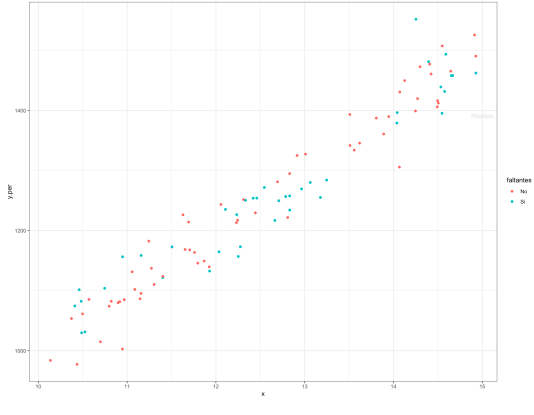

Figura 23: Relación de la variable de interés con la covariable auxiliar para el enfoque de imputación múltiple Bootstrap

Imputación Múltiple Bayesiana con Distribuciones Previas no Informativas

- La distribución posterior de σ^2 es:

$$\sigma^2|y,x \sim \frac{\sum_{i=1}^{n_1}(y_i - \hat{\beta}x_i)^2}{\chi^2_{n_1-1}}$$

con $\hat{\beta} = \frac{\sum_{i=1}^{n_1} x_i y_i}{\sum_{i=1}^{n_1} x_i^2}$.

- La distribución posterior de β es:

$$\beta|\sigma^2,y,x \sim Normal\left(\hat{\beta}, \frac{\sigma^2}{\sum_{i=1}^{n_1} x_i^2}\right)$$

- Asumiendo el anterior enfoque bayesiano de imputación múltiple, se llega a resultados similares.

Tabla comparativa de los metodos

Propiedades	Ingenuo	Bootstrap	Bayesiano
Esperanza	100.00	100.01	100.01
Error estándar	0.24	0.41	0.42
Amplitud	0.96	1.60	1.66
Cobertura	0.83	0.97	0.95

Conclusiones

- La imputación determinista puede subestimar la dispersión de la variable de interés.
- Los tres métodos de imputación mostraron estimaciones puntuales insesgadas.
- El método ingenuo subestima gravemente el error estándar y la amplitud de los intervalos de confianza.
- Los métodos Bootstrap y Bayesiano ofrecen una mejor cobertura del intervalo de confianza al 95%.

¡Gracias!

Email: andres.gutierrez@cepal.org

Referencias

Bethlehem, Jelke, Fannie Cobben, y Barry Schouten. 2009. «Indicators for the Representativeness of Survey Response». En *Statistics Canada's International Symposium*, 10.

Fuller, W. A. 2009. *Sampling Statistics*. Wiley.

Hayes, Clinton, y Nicole Watson. 2009. «HILDA Imputation methods». *Working paper*.

Little, R., y D. B. Rubin. 2002. *Statistical Analysis with Missing Data*. 2nd ed. Wiley.

Lumley, Thomas. 2010. *Complex surveys: a guide to analysis using R*. Wiley series en survey methodology. Wiley.

Rubin, Donald B. 1987. *Multiple Imputation for nonresponse in surveys*. Wiley series en probability y mathematical statistics Applied probability y statistics. Wiley.

Sarndal, Carl-Erik. 2011. «Three Factors to Signal Non-Response Bias With Applications to Categorical Auxiliary Variables». *International Statistical Review / Revue Internationale de Statistique* 79 (2).

Särndal, C. 2011. «The 2010 Morris Hansen lecture: Dealing with survey nonresponse in data collection». *Journal of Official Statistics* 27: 1-21.

Särndal, Carl-Erik, y Sixten Lundström. 2006. *Estimation in surveys with nonresponse*. Repr. Wiley series en survey methodology. Wiley.

Shlomo, Natalie, Chris Skinner, y Barry Schouten. 2012. «Estimation of an indicator of the representativeness of survey response». *Journal of Statistical Planning and Inference* 142 (1): 201-11. https://doi.org/10.1016/j.jspi.2011.07.008.