

Imputación Múltiple

Departamento de Metodologías e Innovación Estadística
Subdepartamento de Investigación Estadística
Instituto Nacional de Estadística

Felipe Molina Jaque

Jefe Subdepartamento: José Bustos

Contents

1 Simulación imputación encuesta CASEN	2
1.1 Distribución de los ingresos	2
1.2 Proceso de Imputación con <code>mice</code>	4
1.3 Imputación múltiple	5
1.4 Resultados imputacion	6
1.5 Resultados de la variable original	6
1.6 Estimaciones en el diseño de muestreo complejo	7
1.7 Imputaciones transformacion ingreso	7
1.8 Estimaciones regionales y Nacional	9

1 Simulación imputación encuesta CASEN

Utilizando la base de datos de la Encuesta de Caracterización Socioeconómica Nacional (CASEN) del año 2017, se indujo pérdida de datos de forma aleatoria en hogares donde el jefe de familia posee menos de 7 años de escolaridad, enfocándose en la variable `ingcorte`, que representa el ingreso per cápita de los hogares. Esta base fue integrada a la base de datos `xencuesta`, la cual incluye variables relacionadas con la composición y características del hogar, como la proporción de miembros ocupados, desocupados, inactivos, presencia de al menos un miembro del hogar en ciertas ramas de actividad, y la existencia de miembros indígenas, entre otros.

Para la variable `ingcorte` y otras derivadas de la pérdida de datos, se aplicó una transformación logarítmica del tipo log-shift. En este contexto, el coeficiente α en la fórmula $\log(x + \alpha)$ fue determinado basándose en una serie de valores donde el coeficiente de asimetría de Fisher se aproxima a cero.

```
hogar <- casen %>% filter(pco1 == "Jefe(a) de hogar",
                           ingcorte != 0) %>%
  mutate(ingcorte_mar = ifelse(anoest < 7, NA, ingcorte),
        log_ingcorte = log(ingcorte + 8600),
        log_ingcorte_mar = log(ingcorte_mar + 8600)) %>%
  left_join(xencuesta, by = "id_hogar")
```

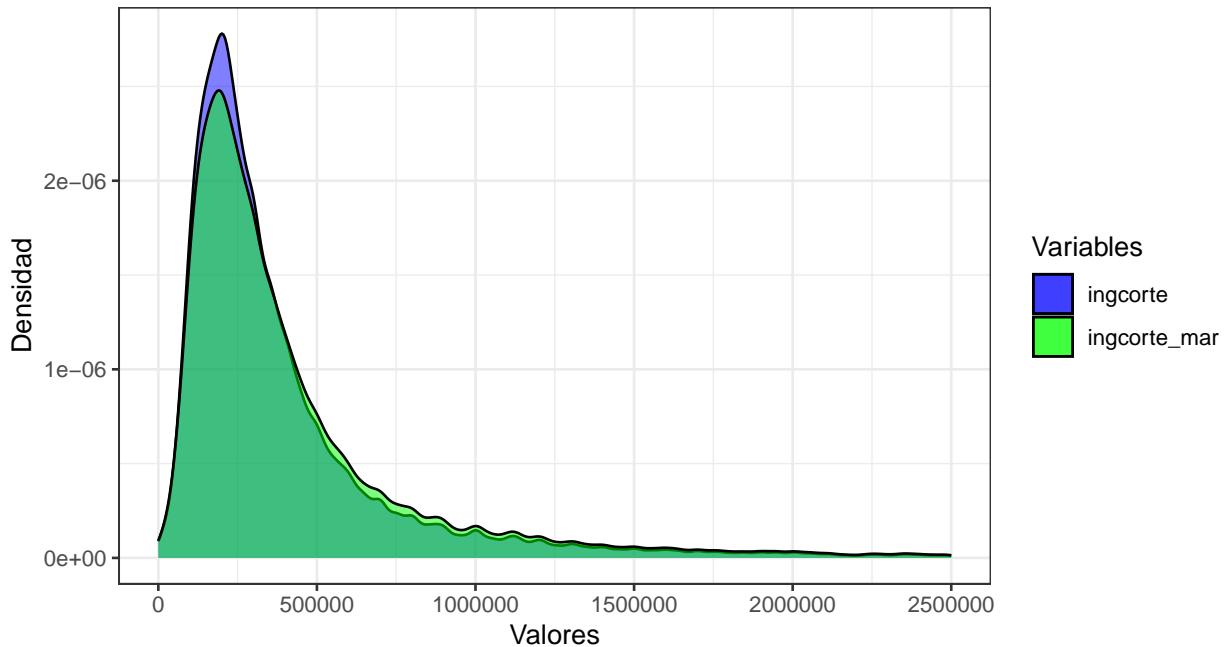
1.1 Distribución de los ingresos

El gráfico visualiza la distribución del ingreso per cápita de hogares de la base de datos CASEN 2017: la variable original y la variable después de introducir pérdida de datos aleatoria. La curva azul representa la distribución de la variable original de ingresos per cápita (`ingcorte`), mientras que la curva verde muestra la misma variable con valores faltantes perdidos (`ingcorte_mar`).

La distribución de `ingcorte`, la variable sin pérdida de datos, se caracteriza por un pico más pronunciado y una menor dispersión, indicando una concentración de hogares alrededor de un rango específico de ingreso per cápita. En contraste, `ingcorte_mar` muestra una dispersión mayor y un pico menos definido, lo cual sugiere que la introducción de la pérdida de datos ha tenido un efecto en la distribución general de los ingresos, posiblemente reduciendo la precisión de la estimación de la densidad y cambiando la centralidad y variabilidad de la distribución.

Estas diferencias son esenciales para entender cómo los valores perdidos pueden influir en la interpretación de la distribución de ingresos en la encuesta. La comparación directa de las curvas permite evaluar el impacto de la pérdida y la necesidad de considerar estos cambios al realizar análisis estadísticos y al interpretar los resultados en estudios socioeconómicos.

Histograma de Densidad para los ingresos



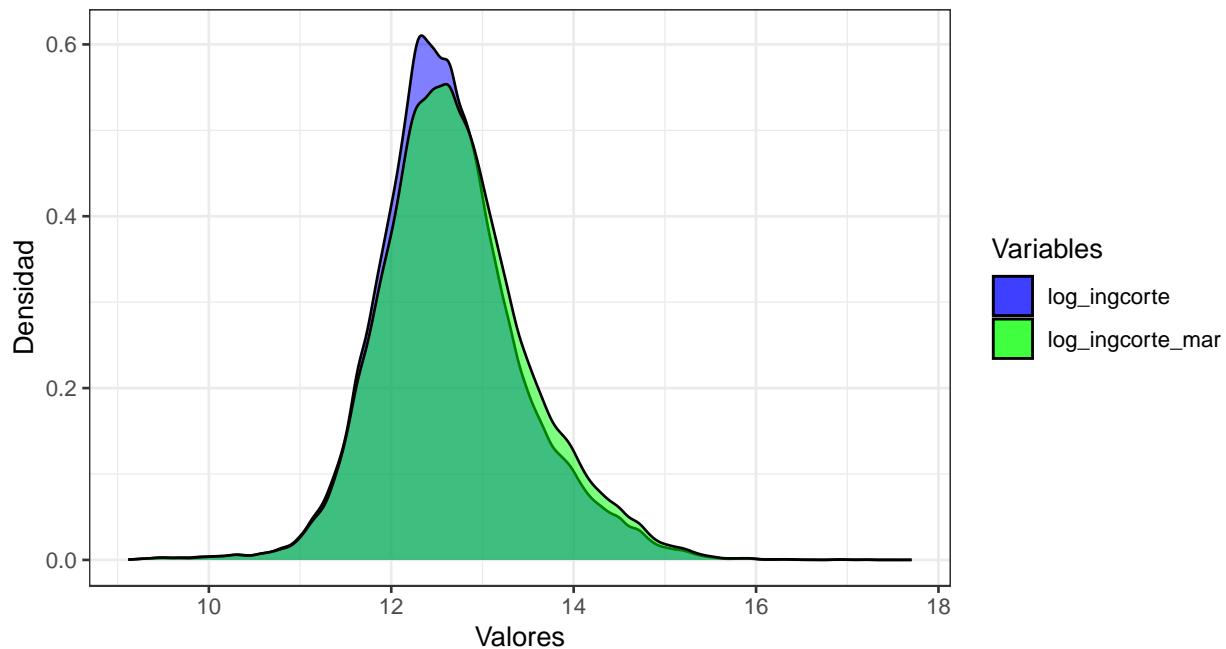
Por otra parte, la imagen siguiente muestra un histograma de densidad para la distribución del logaritmo de los ingresos, donde se comparan dos versiones de la variable de ingreso per cápita transformada logarítmicamente. La curva azul, etiquetada como `log_ingcorte`, representa la variable original transformada sin pérdida de datos. La curva verde, `log_ingcorte_mar`, representa la misma variable después de inducir aleatoriamente la pérdida de datos, sin ninguna imputación realizada para corregir o estimar los valores faltantes.

Observamos que la curva azul (`log_ingcorte`) presenta un pico más alto y estrecho alrededor de los valores medios, indicando una concentración más fuerte de hogares en un rango estrecho de ingreso per cápita transformado. Esto es consistente con una distribución que no ha sido alterada por la pérdida de datos.

Por otro lado, la curva verde (`log_ingcorte_mar`) es más plana y ancha, lo que indica una distribución más uniforme de los ingresos per cápita transformados y sugiere una variabilidad mayor después de la pérdida de datos. La reducción en la altura del pico y el ensanchamiento de la distribución en `log_ingcorte_mar` refleja la ausencia de una imputación que ajuste o estime los valores faltantes, lo que puede resultar en una representación menos precisa de la verdadera distribución de ingresos en la población estudiada.

La comparación de estas dos curvas es crucial, ya que destaca la importancia de la imputación en la preservación de la estructura original de la distribución de ingresos per cápita, y proporciona una visualización directa del efecto de la pérdida de datos en el análisis estadístico de encuestas de hogares.

Histograma de Densidad para logaritmo de los ingresos



1.2 Proceso de Imputación con `mice`

La imputación de valores ausentes se realizará mediante la librería `mice` en R. Iniciamos por seleccionar las variables pertinentes del conjunto de datos `hogar` que incluyen identificadores de hogar, el logaritmo transformado de los ingresos con datos faltantes (`log_ingcorte_mar`), años de estudio, proporción de ocupados, y otras variables estratégicas y de unidad.

Para `log_ingcorte_mar`, que representa el logaritmo de los ingresos per cápita con datos perdidos, optaremos por un modelo lineal bayesiano, conocido por su capacidad para manejar incertidumbres y aportar estimaciones robustas.

```
# Asignación de un modelo bayesiano lineal para 'log_ingcorte_mar'
meth["log_ingcorte_mar"] <- "norm.nob"

# Construcción de la matriz predictiva
predM <- make.predictorMatrix(datos)

# Especificación del modelado predictivo para la imputación
predM["log_ingcorte_mar", c("anoest", "prop_ocupados")] <- -2
```

La elección del método `norm.nob` se debe a su enfoque en la imputación de variables continuas bajo un esquema normal sin valores fuera de límites, lo que es adecuado para variables log-transformadas como `log_ingcorte_mar`. La configuración de la matriz predictiva determina las relaciones entre las variables, lo que refleja una decisión basada en conocimiento previo o estrategias de modelado específicas.

1.3 Imputación múltiple

Se ha establecido la generación de $m = 5$ conjuntos de imputaciones para abordar los valores perdidos en nuestro conjunto de datos. Este número de imputaciones ha sido elegido para proporcionar una base sólida para estimaciones estadísticas fiables y para reflejar la variabilidad inherente en los datos faltantes. Tras completar las imputaciones, iniciamos la fase de agrupamiento, también conocida como la fase de pooling.

Empleando la librería `mice` y el modelo definido, se lleva a cabo la imputación múltiple. El proceso se detalla a continuación:

```
# Ejecución de la imputación múltiple
imp <- mice(datos, method = meth,
             predictorMatrix = predM, m = 5)

# Inicialización de una matriz para almacenar las imputaciones de 'log_ingcorte_mar'
log_imp <- matrix(nrow = nrow(datos), ncol = 5)

# Se considera la primera imputación para almacenar los datos utilizados
completed_data <- complete(imp, 1)

# Bucle para extraer cada conjunto de imputaciones y almacenarlas en la matriz
for(i in 1:5){
  log_imp[, i] <- complete(imp, i)$log_ingcorte_mar
}

# Cálculo del promedio de las imputaciones para 'log_ingcorte_mar'
completed_data$log_ingcorte_mar <- rowMeans(log_imp, na.rm = TRUE)
```

Este bucle for recorre cada conjunto de imputaciones, extrayendo la variable `log_ingcorte_mar` y almacenándola en una matriz. Posteriormente, calculamos la media de estas imputaciones para obtener una única estimación imputada para cada observación.

Una vez completada la fase de agrupamiento, los datos promediados son retransformados al espacio original para mantener la consistencia con los datos no imputados y facilitar la interpretación analítica.

```
# Retransformación de la variable 'log_ingcorte_mar' al espacio original
completed_data$imp_ingcorte <-
  exp(completed_data$log_ingcorte_mar) - 8600
```

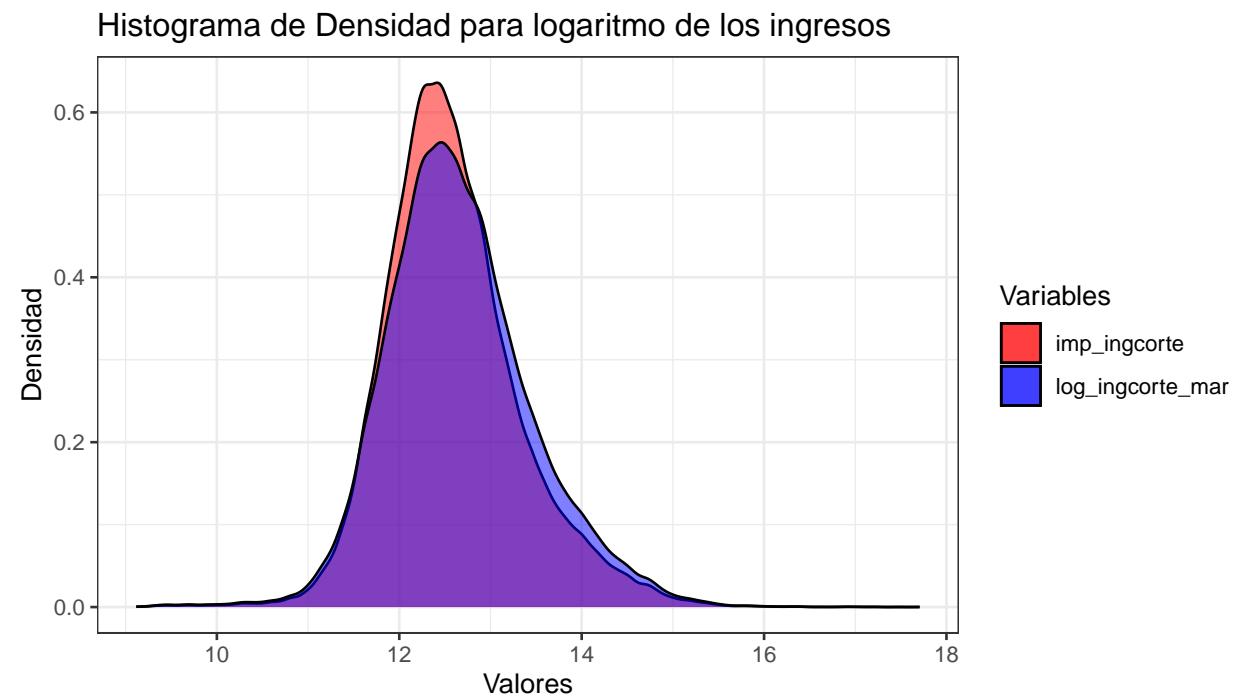
En esta etapa, aplicamos la función exponencial a los valores imputados promediados para revertir la transformación logarítmica y luego restamos 8600, que es el valor utilizado en la transformación log-shift inicial, para alinear los datos con su escala original de ingresos. Este paso final asegura que los datos imputados sean comparables con los datos originales, lo que es esencial para cualquier análisis subsiguiente y para la toma de decisiones basadas en datos.

1.4 Resultados imputacion

El siguiente gráfico presenta un histograma de densidad que compara la distribución del logaritmo de los ingresos antes y después del proceso de imputación.

La curva azul muestra la forma de la distribución original con datos perdidos, mientras que la curva roja refleja la distribución después de que los valores perdidos han sido imputados y los datos han sido retransformados al espacio original.

Al analizar las dos curvas, se observa que la distribución post-imputación (roja) tiende a seguir la forma general de la distribución pre-imputación (azul), pero con una densidad ligeramente más alta alrededor del pico, lo cual sugiere que la imputación ha contribuido a una concentración ligeramente mayor de valores en torno al promedio de los ingresos transformados. Esto puede indicar que la imputación ha sido efectiva en mantener la estructura general de la distribución de los ingresos, mientras se ocupa de los valores perdidos de manera coherente con la variabilidad observada en los datos originales.



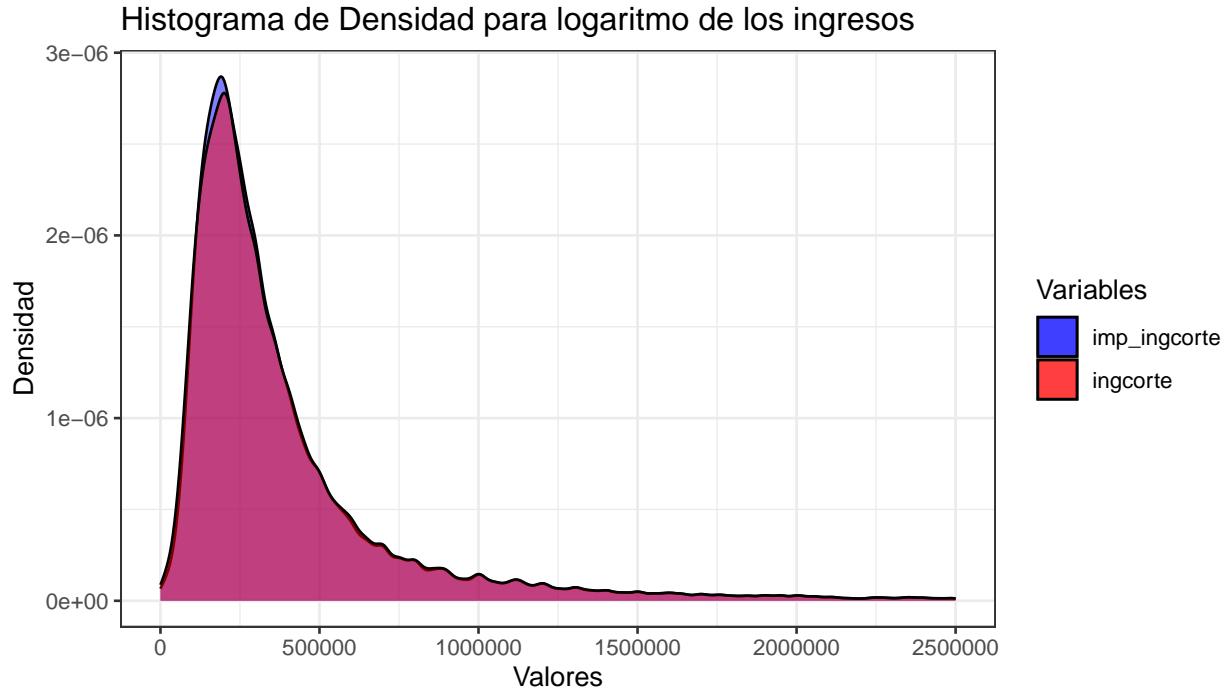
1.5 Resultados de la variable original

La curva roja en el gráfico representa la distribución original de la variable ingcorte sin datos perdidos. Esta distribución muestra un pico muy marcado cerca del origen, indicando una alta concentración de hogares con ingresos per cápita bajos. La curva azul representa la misma variable después de la imputación de los datos perdidos.

Al observar las curvas, se nota que la distribución de la variable imputada (azul) sigue de cerca la forma de la distribución original (roja), lo que indica que el proceso de imputación ha podido mantener la forma general de la distribución de los ingresos. Sin embargo, hay una diferencia notable en la altura de los picos, con la curva de imputación siendo más suave y menos pronunciada, lo que sugiere que la imputación ha distribuido los valores perdidos de

una manera que reduce la concentración extrema de hogares con ingresos muy bajos.

En este caso, la imputación no logra de manera precisa llevar los datos perdidos a la distribución original de los datos, lo cuál puede producir un sesgo considerable en la estimación de los indicadores correspondientes.



1.6 Estimaciones en el diseño de muestreo complejo

Para observar el efecto en los errores del muestreo, a los resultados de la imputación indexamos los valores previos y definimos el diseño de muestreo complejo

```
contrastes <- completed_data %>%
  left_join(hogar %>%
    transmute(id_hogar,
              log_ingcorte_mar_NA = log_ingcorte_mar,
              ingcorte, ingcorte_mar,
              region))

## Joining with 'by = join_by(id_hogar)'

diseno <- contrastes %>%
  as_survey_design(ids = varunit, strat = varstrat,
                   weights = expr, nest = TRUE)
options(survey.lonely.psu = "adjust")
```

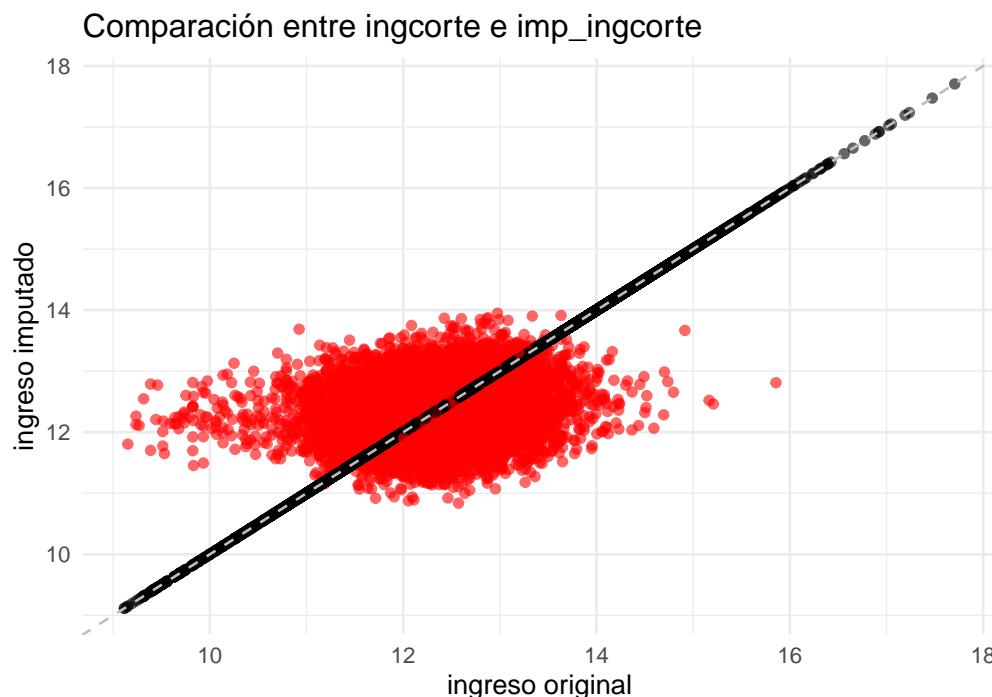
1.7 Imputaciones transformacion ingreso

La imagen muestra un gráfico de dispersión que compara los valores de ingreso original con los valores de ingreso imputados. Los puntos rojos representan los valores imputados, mientras

que los puntos negros indican los valores originales no imputados. Una línea negra, la línea de identidad donde el ingreso imputado es igual al ingreso original dado que no representa valores perdidos, atraviesa el gráfico.

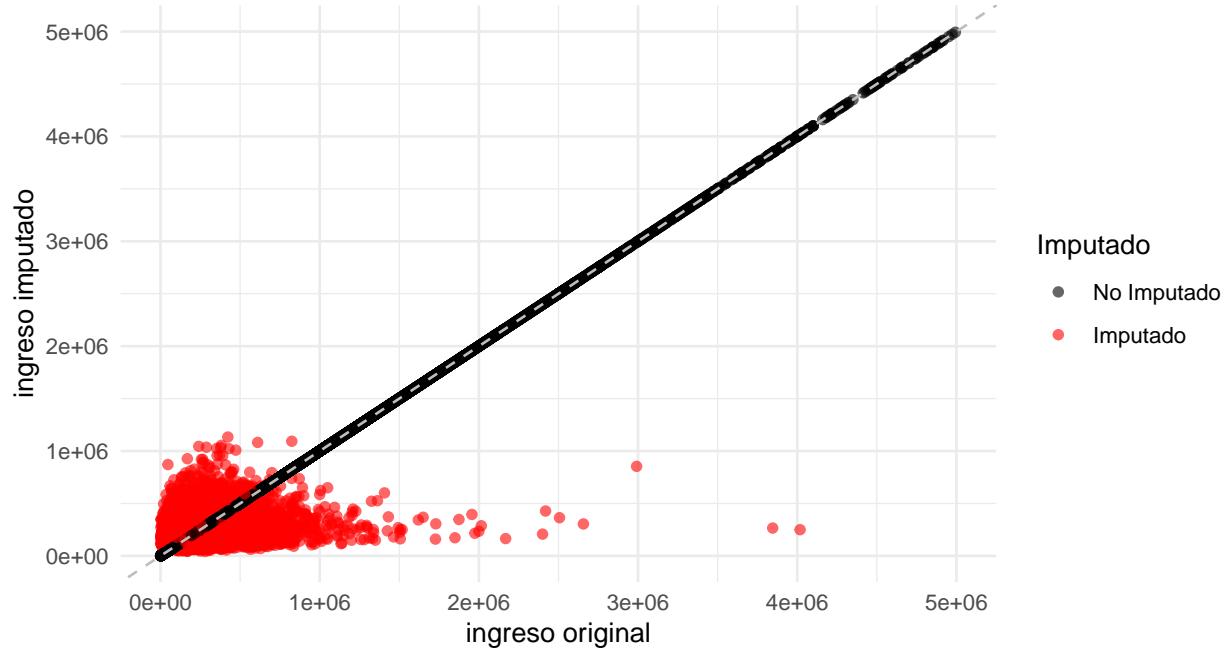
La densidad de puntos rojos alrededor de esta línea sugiere que los ingresos imputados tienden a estar en concordancia con los ingresos originales, lo cual indica que la imputación ha sido, en gran medida, consistente con la variabilidad inherente a la variable original, salvo al considerar ingresos bajos los cuales tienden a ser sobreestimados por el modelo lo que refleja la incertidumbre y la variabilidad introducidas por el proceso de imputación.

Mientras que algunos puntos imputados se alinean estrechamente con la línea, indicando una imputación cercana al valor real, otros están más alejados, lo que refleja diferencias entre los valores imputados y los originales. Esta variabilidad es esperada en la imputación múltiple y es una representación gráfica del rango de posibles valores que podrían tomar los datos faltantes.



Misma dispersión se observa al comparar la variable transformada, donde los ingresos menores fueron los mayormente afectados debido al proceso de perdida aleatoria generada, generando mayor incertidumbre y variabilidad en dichos ingresos imputados.

Comparación entre ingcorte e imp_ingcorte



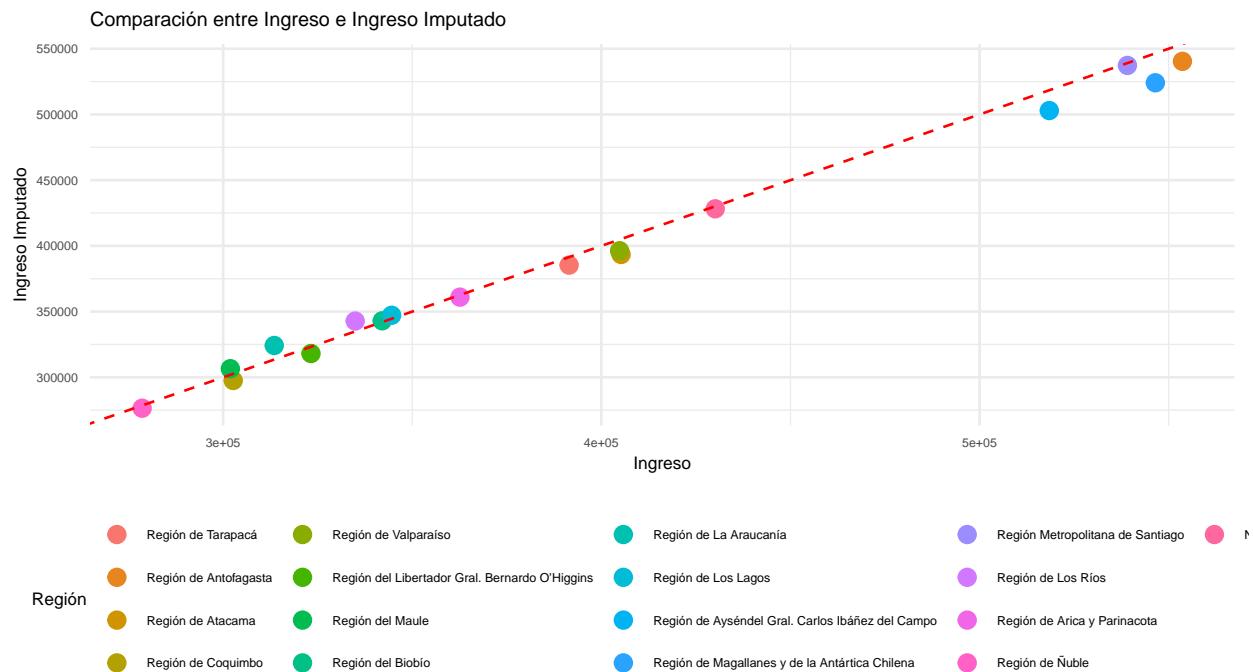
1.8 Estimaciones regionales y Nacional

A partir del análisis de los datos regionales y considerando el diseño de muestreo, se calcula la media del ingreso para cada región, tanto con los valores originales como con los valores imputados. La sobreestimación y subestimación de los ingresos imputados en ciertas regiones sugiere que el método de imputación no ha captado de buena forma el ingreso en dichas regiones. Esto puede deberse a varias razones, como la presencia de valores atípicos en los datos imputados o una distribución desigual de los datos faltantes que no refleja con precisión la distribución subyacente de los ingresos en algunas regiones.

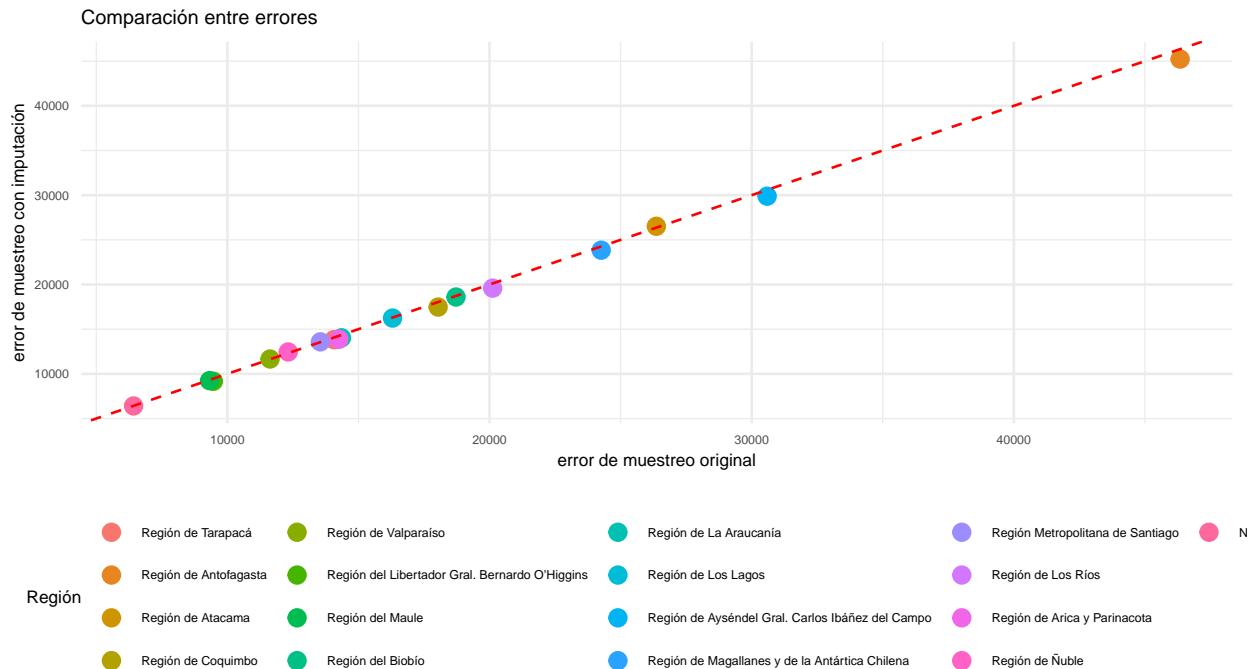
```
## # A tibble: 17 x 7
##   region ingreso ingreso_se ingreso_NA ingreso_NA_se ingreso_imp ingreso_imp_se
##   <fct>    <dbl>     <dbl>      <dbl>      <dbl>      <dbl>      <dbl>
## 1 Regió~  391478.   14065.    403339.    15070.    385358.    13826.
## 2 Regió~  553641.   46343.    573736.    49644.    540431.    45233.
## 3 Regió~  405213.   26363.    431458.    31020.    393462.    26516.
## 4 Regió~  302665.   18039.    323271.    22672.    297575.    17473.
## 5 Regió~  404828.   11626.    435749.    13647.    396312.    11661.
## 6 Regió~  323226.   9464.     348277.    11902.    318103.    9169.
## 7 Regió~  301908.   9320.     334657.    12498.    306528.    9240.
## 8 Regió~  342057.   18721.    374149.    23031.    342985.    18599.
## 9 Regió~  313509.   14351.    358013.    19128.    324229.    14033.
## 10 Regió~ 344551.   16300.    388764.    22016.    347208.    16244.
## 11 Regió~ 518436.   30590.    578209.    36107.    502952.    29883.
## 12 Regió~ 546481.   24260.    578835.    26104.    524109.    23837.
## 13 Regió~ 539096.   13547.    583108.    15421.    537312.    13580.
## 14 Regió~ 334911.   20120.    372031.    25902.    342823.    19591.
```

## 15 Regió~	362612.	14249.	376517.	16041.	360997.	13867.
## 16 Regió~	278578.	12321.	307169.	17153.	276486.	12450.
## 17 Nacio~	430112.	6421.	473386.	7724.	428218.	6407.

Este resultado es importante al interpretar las estimaciones de ingreso medio del hogar, ya que indica que la imputación puede haber influido en la representación de la capacidad económica regional.



Por otra parte, el gráfico que muestra muestra la comparación entre los errores de muestreo del ingreso original y los errores de muestreo del ingreso después de la imputación, diferenciados por región sugiere que la imputación puede subestimar los errores de muestreo en comparación con los valores originales en determinadas regiones. Esto implica que la imputación puede no reflejar adecuadamente la variabilidad y la incertidumbre inherente a los datos faltantes en dichas regiones. A pesar de ello, no se observa un gran cambio en las estimaciones de varianza de los indicadores.



Del procedimiento anterior y considerando la sección de consideraciones de imputar en encuestas complejas, el ejercicio anterior se puede tener en cuenta los siguientes puntos.

- Consistencia en la Subestimación de Errores: Si la subestimación es sistemática a través de las regiones, esto podría indicar una característica inherente del método de imputación empleado que necesita ajuste.
- Impacto en la Inferencia Estadística: La subestimación del error podría llevar a un exceso de confianza en las estimaciones estadísticas y, potencialmente, a conclusiones erróneas sobre significancia estadística o la magnitud de los efectos.
- Revisión de Métodos de Imputación: Sería prudente revisar los métodos de imputación utilizados para asegurarse de que estén capturando adecuadamente la variabilidad de los datos. Esto podría incluir la exploración de métodos alternativos o la incorporación de ajustes en el modelo para reflejar mejor la incertidumbre.
- Importancia del Diseño de Muestreo: La imputación en encuestas complejas debe tener en cuenta el diseño de muestreo. La estratificación, el agrupamiento y los pesos deben ser considerados para evitar sesgos y mantener la representatividad.
- Validación de Resultados Imputados: Sería beneficioso realizar una validación de las imputaciones comparándolas con datos auxiliares o utilizando técnicas como la validación cruzada para evaluar la calidad de las imputaciones.

Es importante ser cauteloso al generalizar los resultados de datos imputados, especialmente cuando se utilizan para informar políticas o decisiones que afectan a las regiones evaluadas, en ese sentido, es crucial documentar y comunicar claramente el proceso de imputación, incluyendo las técnicas empleadas y las implicaciones de los errores de muestreo subestimados, para que los usuarios de los datos puedan tomar decisiones informadas.