

# Evaluación de modelos de regresión logística en el tratamiento de la no respuesta en datos de escala likert

Valentina Cardona Saldaña<sup>1\*</sup> y Javier Garzón Hernández<sup>1†</sup>

<sup>1</sup>Departamento de Estadística, Universidad Nacional de Colombia

## Abstract

*Las tasas de participación en los estudios pueden tener un impacto directo en la validez de los datos recopilados, ya que la falta de respuesta siempre conlleva el riesgo de sesgo. Este artículo compara dos modelos de regresión logística como estrategias de imputación para datos faltantes en variables ordinales de escala Likert: regresión logística multinomial y regresión logística ordinal. Para ello, se generaron 90 escenarios de datos faltantes a partir de datos reales del Estudio de Habilidades Sociales y Emocionales (SSES), considerando cada habilidad evaluada, el mecanismo de datos faltantes (MCAR, MAR, NMAR) y dos probabilidades de pérdida de datos (30 % y 50 %). La evaluación y comparación de los métodos se realizó con base en la Exactitud Global (OA), la Curva de Sumas Acumuladas (CS) y el Error Absoluto Medio (MAE). Los resultados sugieren que el modelo multinomial logra mayor exactitud global, mientras que el ordinal es más preciso para errores pequeños. El rendimiento se ve afectado por el mecanismo de datos faltantes. Se recomienda considerar la distribución de las respuestas en escala likert en el momento de seleccionar el mejor método de imputación.*

**Palabras clave:** valores faltantes; métodos de imputación; escala likert; regresión logística

## Abstract

*Participation rates in studies can have a direct impact on the validity of the data collected, as non-response always carries the risk of bias. This article compares two logistic regression models as imputation strategies for missing data in ordinal Likert-scale variables: multinomial logistic regression and ordinal logistic regression. To this end, 90 missing data scenarios were generated using real data from the Study on Social and Emotional Skills (SSES), considering each assessed skill, the missing data mechanism (MCAR, MAR, NMAR), and two probabilities of data loss (30% and 50%). The methods were evaluated and compared based on Overall Accuracy (OA), the Cumulative Sums Curve (CS), and the Mean Absolute Error (MAE). The results suggest that the multinomial model achieves higher overall accuracy, while the ordinal model is more precise for small errors. Performance is affected by the missing data mechanism. It is recommended to consider the distribution of Likert scale responses when selecting the best imputation method.*

**Keywords:** missing values; imputation methods; Likert scale; logistic regression

## 1 Introducción

El formato de pregunta tipo Likert, denominado así en honor a su creador (Likert, 1932), es ampliamente utilizado en las ciencias sociales y del comportamiento para medir actitudes, opiniones, preferencias y percepciones, entre otros aspectos. Las escalas tipo Likert son de naturaleza ordinal y cuentan con múltiples categorías de respuesta, destacándose por su facilidad de aplicación y su adaptabilidad a diversos contextos de evaluación (Muñiz, 2018). Sin embargo, cuando se presentan datos faltantes debido a la no respuesta de las y los participantes, es crucial aplicar estrategias adecuadas de manejo de datos para garantizar estima-

ciones precisas de los parámetros e inferencias válidas sobre la población (Wei Wu & Enders, 2015, p. 484).

La ausencia de datos es común y problemática en estudios que emplean instrumentos de autoinforme (Downey & King, 1998). Un enfoque tradicional para su manejo es la eliminación de casos incompletos (*list-wise deletion*) o de variables (*variable deletion*); no obstante, estas metodologías pueden introducir sesgo, reducir el tamaño de muestra y afectar la validez del constructo (o variable latente) en análisis psicométricos (Raaijmakers, 1999). Para mitigar estos problemas, en muchos casos se recurre a métodos de imputación, que funcionan sustituyendo los datos faltantes con valores de reemplazo, aumentando así la cantidad de datos utilizables (Hu et al., 1998). Sin embargo, gran parte de estas metodologías asumen normalidad en la dis-

\*Correo: vcardonas@unal.edu.co

†Correo: jagarzonhe@unal.edu.co

tribución de los datos, lo que puede no ser adecuado para escalas ordinales como las de tipo Likert. Por ello, es fundamental emplear modelos de imputación que respeten las propiedades de las escalas ordinales y permitan una estimación más precisa y coherente con la naturaleza de los datos.

Una de las estrategias principales para la imputación de datos faltantes en variables ordinales es el uso de modelos diseñados específicamente para datos categóricos, como la regresión logística multinomial. Otra alternativa es la imputación basada en un modelo de variable latente, como el modelo de probabilidades proporcionales, un tipo de regresión logística ordinal que asume que las categorías de la variable ordinal reflejan umbrales en una distribución latente subyacente.

Aquí, el enfoque principal es remuestrear escenarios con diferentes mecanismos de no respuesta para evaluar el desempeño en habilidad predictiva de dos estrategias de imputación. Si bien hay un componente aplicado, debido al uso de datos reales, no se pretende resolver un problema práctico específico, sino analizar el comportamiento de técnicas estadísticas bajo condiciones remuestreadas. Este enfoque se alinea con el método de simulación, ya que contribuye al análisis teórico mediante resultados que permiten conjeturar ventajas de ciertos métodos en contextos específicos.

El objetivo principal de este artículo es comparar dos métodos de regresión para el tratamiento de la no respuesta en datos expresados en escala Likert: la regresión logística ordinal con el *Modelo Logit Acumulativo de Probabilidades Proporcionales* y la regresión multinomial o *Modelo Logit Multicategoría*. Para ello, en primer lugar, se generaron escenarios de falta de datos mediante la eliminación de observaciones en un conjunto de datos con respuestas en escala Likert. Luego, se aplicaron las dos metodologías seleccionadas al conjunto de datos con valores faltantes. Finalmente, se evaluó la habilidad predictiva y las limitaciones de cada enfoque en la imputación de datos faltantes en escalas tipo Likert.

## 2 Marco teórico

### 2.1 Antecedentes

En algunos estudios, la variable de interés puede ser categórica con más de dos categorías, es decir, multinomial. Estas categorías pueden registrarse en una escala nominal u ordinal. En el análisis de resultados clínicos, tratar las variables como ordinales en lugar de dicotomizarlas puede proporcionar mayor precisión y conservar la información contenida en la estructura

ordinal de los datos (Arcese et al., 2018).

Bajo condiciones simuladas, los resultados sugieren que los métodos que utilizan modelos de regresión logística de probabilidades proporcionales y regresión logística multinomial generalmente superan a otros métodos (Wongkamthong & Akande, 2023; Homeister, 2024). Sin embargo, Wei Wu and Enders (2015) desaconsejan el uso de métodos de regresión logística, incluidos los modelos multinomiales y de probabilidades proporcionales, para la imputación de datos dicotómicos o politómicos. A pesar de estar diseñados específicamente para variables categóricas, estos métodos mostraron estimaciones deficientes en los parámetros analizados en su estudio. Se observó una alta sensibilidad al tamaño muestral, a la asimetría en la distribución de los ítems y al número de categorías, lo que implica que una estimación precisa requiere un tamaño muestral considerable. Además, este requisito aumenta a medida que la distribución de los datos se vuelve más asimétrica, el número de categorías se incrementa o se incorporan más variables al modelo de imputación.

En su estudio, Wei Wu and Enders (2015) modelaron 12 variables ordinales dentro del modelo de imputación, lo que pudo haber exacerbado estas limitaciones. No obstante, Allison (2005) encontró que la regresión logística producía estimaciones satisfactorias cuando se contaba con un tamaño muestral grande y se modelaba solo una variable en el proceso de imputación.

Por otra parte, el rendimiento de la imputación está determinado por los mecanismos subyacentes que generan los datos faltantes (Van Buuren, 2018). En particular, depende de si la ausencia de datos ocurre de manera sistemática o aleatoria, lo que da lugar a tres patrones principales: Falta Completamente al Azar (MCAR), Falta al Azar (MAR) y Falta No al Azar (MNAR) (Van Buuren, 2018; Pan & Chen, 2023).

Por último, diferentes estudios (Köhler et al., 2015; Meinck et al., 2017) han evidenciado que la propensión a omitir respuestas en pruebas estandarizadas presenta una componente individual específica; es decir, la probabilidad de omisión presenta una asociación positiva alta con el nivel de competencia de los sujetos. Además, se han identificado diferencias sistemáticas en la frecuencia de omisiones entre personas con antecedentes migratorios y estudiantes de distintos tipos de instituciones educativas, incluso después de controlar por el nivel de competencia. De manera similar, Akinboboye et al. (2024) encontró diferencias estadísticamente significativas en los niveles de habilidad estimados entre grupos, evidenciando variaciones en la probabilidad

de omisión entre ellos. No obstante, el autor enfatiza la importancia de garantizar la comparabilidad de los grupos en términos de habilidad. Estos hallazgos son consistentes con el análisis realizado por Durrant and Schnepf (2017) en el contexto del estudio PISA.

## 2.2 Estrategias de imputación

Los modelos más utilizados para variables de respuesta categórica ordinal son los Modelos Lineales Generalizados (MLG) con funciones de enlace no lineales (Agresti & Tarantola, 2018). Los parámetros de estos modelos se interpretan a través de medidas como los *odds ratios* y los *probits*. Para datos politómicos ( $c > 3$ ), es posible emplear modelos de regresión logística multinomial y regresión logística ordinal. Aunque el primero está diseñado para datos nominales, también puede aplicarse a datos ordinales en ciertos contextos.

### Regresión Logística Multinomial

La regresión logística multinomial o **Modelo Logit Multicategoría** es una técnica estadística utilizada para modelar una variable dependiente categórica con más de dos categorías sin un orden intrínseco. Constituye una extensión de la regresión logística binaria, que se emplea cuando la variable dependiente tiene solo dos categorías (por ejemplo, éxito/fracaso).

En este modelo, la variable endógena puede asumir múltiples categorías posibles, por lo que la distribución de probabilidad adecuada para describir el fenómeno es la distribución multinomial. Es importante distinguir la regresión logística multinomial de la regresión logística condicional y ordinal. En la regresión logística multinomial, se estiman tantos vectores de parámetros como categorías menos una. Por otro lado, la regresión logística ordinal incorpora información sobre el orden de las categorías, lo que impide modelarla con los enfoques mencionados anteriormente (McCullagh & Nelder, 1989).

Este método se aplica cuando el objetivo es estimar la probabilidad de pertenencia a cada categoría en función de una o más variables predictoras. Para ello, se selecciona una categoría de referencia y se modela la razón de probabilidades (*log odds*) de cada categoría en comparación con esta.

$$\log \left[ \frac{P(\mathbf{y} = j)}{P(\mathbf{y} = k)} \right] = \beta_{0j} + \beta_{1j}x_1 + \beta_{2j}x_2 + \dots + \beta_{pj}x_p \quad (1)$$

Donde:

- $P(\mathbf{y} = j)$  es la probabilidad de que la observación pertenezca a la categoría  $j$ .

- $P(\mathbf{y} = j)$  es la probabilidad de que la observación pertenezca a la categoría de referencia.
- $\beta_{0j}, \beta_{1j}, \dots, \beta_{pj}$  son los coeficientes del modelo para la categoría  $j$ .
- $x_1, x_2, \dots, x_p$  son las variables predictoras.

El modelo se basa en varios supuestos. La independencia de alternativas irrelevantes (IIA) establece que la elección entre dos categorías no debe verse afectada por la presencia de otras opciones. La relación entre los predictores y el logaritmo de la razón de probabilidades debe ser lineal. Asimismo, se asume la ausencia de multicolinealidad, es decir, que las variables predictoras no deben estar altamente correlacionadas entre sí. Finalmente, es fundamental contar con un tamaño muestral suficiente para cada categoría.

### Regresión Logística Ordinal

El **Modelo Logit Acumulativo de Probabilidades Proporcional** (*Cumulative Logit Model with Proportional Odds*, en inglés) es un tipo de Regresión Logística Ordinal (*Ordered Logistic Regression*). Este modelo se emplea para analizar la relación entre una variable de respuesta ordinal y un conjunto de covariables, que pueden ser tanto discretas como continuas.

Primero, se abordarán los conceptos fundamentales de la regresión logística ordinal. Sea  $y$  la variable respuesta con  $c$  categorías ordenadas. Entonces, las probabilidades asociadas a cada categoría son  $(\pi_1, \pi_2, \dots, \pi_j)$  y la probabilidad acumulada para la categoría  $j$  es

$$P(\mathbf{y} \leq j) = \pi_1 + \dots + \pi_j, \quad j = 1, \dots, c$$

El logaritmo de *odds* también se conoce como *logit*. Los *logits* de las probabilidades acumuladas, llamados *logits acumulativos*, son

$$\text{logit}[P(\mathbf{y} \leq j)] = \log \left[ \frac{P(\mathbf{y} \leq j)}{1 - P(\mathbf{y} \leq j)} \right] = \log \left( \frac{\pi_1 + \dots + \pi_j}{\pi_{j+1} + \dots + \pi_c} \right)$$

para  $j = 1, \dots, c - 1$ . Los modelos *logits* acumulativos no utilizan  $P(\mathbf{y} \leq c)$  porque este resultado es igual a 1.

Un modelo para el *logit* acumulativo  $j$  es análogo a una regresión logística binaria, donde las categorías 1 a  $j$  se agrupan en una y las categorías  $j + 1$  a  $c$  forman la otra. Para una variable explicativa  $x$ , el modelo

$$\text{logit}[P(\mathbf{y} \leq j)] = \alpha_j + \beta x, \quad j = 1, \dots, c - 1$$

describe el efecto de  $x$  en los *logits* acumulativos. La constancia de  $\beta$  implica que el efecto de  $x$  es idéntico para todos los *logits* acumulativos  $c - 1$  (Wei Wu & Enders, 2015).

Por lo tanto, cuando se consideran múltiples variables explicativas, el Modelo Logit Acumulativo con Probabilidades Proporcionales se expresa como:

$$\text{logit}[P(\mathbf{y} \leq j)] = \alpha_j + \beta_1 \mathbf{x}_1 + \dots + \beta_p \mathbf{x}_p, \quad j = 1, \dots, c-1$$

dado que los interceptos  $\{\alpha_j\}$  varían según categoría, pero los coeficientes  $\beta$  son constantes en todas ellas.

No obstante, resulta realista considerar la respuesta observada como una medición de una variable latente continua (Agresti, 2018), donde las categorías emergen a partir de puntos de corte específicos en dicha variable. En este contexto, se puede probar\* que la variable ordinal observada sigue el modelo:

$$\text{logit}[P(\mathbf{y} \leq j)] = \alpha_j - \beta_1 \mathbf{x}_1 - \dots - \beta_p \mathbf{x}_p, \quad j = 1, \dots, c-1 \quad (2)$$

El resultado es un modelo con menos parámetros, lo que potencialmente aumenta su potencia y facilita su interpretación frente a los modelos multinomiales (Agresti, 2018). Además, al asumir que la variable observada es una versión discretizada de una variable latente, los parámetros de los efectos son *invariantes* a la selección de las categorías de la variable respuesta  $\mathbf{y}$ ; es decir, incluso si se emplean distintas categorizaciones para la misma variable latente, se espera obtener valores similares para  $\{\beta_j\}$ , salvo por el error de muestreo. Esta propiedad del modelo permite comparar estimaciones entre estudios que utilizan diferentes escalas para la variable de respuesta.

## 3 Metodología

### 3.1 Datos reales

El Estudio de Habilidades Sociales y Emocionales (*Social and Emotional Skills Study*, SSES) es una iniciativa de la Organización para la Cooperación y el Desarrollo Económicos (OCDE) que recopila información sobre 15 Habilidades Sociales y Emocionales (HSE) en estudiantes de 10 y 15 años. Para el presente artículo, se utilizaron los datos de la aplicación de 2019, disponibles en acceso abierto a través de la OCDE<sup>†</sup>. En esta aplicación participaron 10 sitios alrededor del mundo, incluyendo Bogotá (Colombia).

La OCDE (2019) emplea el modelo de los cinco grandes factores (*Big Five*) como referente teórico para la evaluación de las HSE. Bajo este enfoque, las 15 HSE

se agrupan en cinco dimensiones principales, consideradas como constructos o variables latentes. Cada HSE fue medida mediante ocho (8) ítems, con respuestas en una escala Likert de cinco puntos (1: Fuertemente en desacuerdo, 2: En desacuerdo, 3: Ni de acuerdo ni en desacuerdo, 4: De acuerdo, 5: Fuertemente de acuerdo). La estructura de la encuesta es la siguiente:

- **Desempeño de tareas:** Autocontrol, Responsabilidad, Persistencia.
- **Regulación emocional:** Resistencia al estrés, Optimismo, Control emocional.
- **Colaboración:** Empatía, Confianza, Cooperación.
- **Apertura mental:** Tolerancia, Curiosidad, Creatividad.
- **Involucrarse con otros:** Sociabilidad, Asertividad, Energía.

La diversidad de habilidades evaluadas es relevante, ya que el proceso de remuestreo de datos faltantes —descrito más adelante— se realizó sobre 15 ítems, cada uno correspondiente a una HSE. Esta diversidad facilita la evaluación del método de imputación en múltiples contextos de medición, fortaleciendo su aplicabilidad y capacidad de generalización.

Finalmente, para llevar a cabo el remuestreo y el análisis posterior, primero se invirtió la escala de 38 ítems con puntuación inversa. Tras el procesamiento de los datos, se obtuvo un conjunto final de 120 ítems y 54.050 estudiantes.

### 3.2 Remuestreo de datos faltantes

En primer lugar, los datos se dividieron en conjuntos de entrenamiento y prueba, asegurando la reproducibilidad mediante la definición de una semilla. Se seleccionaron aleatoriamente 37.835 estudiantes (70%) para el conjunto de *entrenamiento* y 16.215 estudiantes (30%) para el conjunto de *prueba*. Para conocer la distribución de la respuesta en cada HSE para los datos de entrenamiento y prueba, el lector puede dirigirse al Apéndice A (Sección 5).

Posteriormente, se eliminaron datos del conjunto de prueba para generar conjuntos de datos incompletos de manera artificial, los cuales fueron utilizados en el proceso de imputación. Mientras tanto, el conjunto de entrenamiento, que contiene datos completos, se empleó para ajustar los modelos. La eliminación de datos se realizó considerando tres factores (ver Tabla 1): 1) Habilidad Social y Emocional (HSE), 2) Tipo de dato faltante, 3) Probabilidad de ausencia de datos.

El primer factor, como se mencionó anteriormente, consistió en la eliminación de datos de un ítem por

\* Para más detalles, véase Agresti, 2012

<sup>†</sup> Los datos pueden descargarse en: <https://www.oecd.org/en/data/datasets/SSES-Round-1-Database.html>

**Table 1:** Resumen de escenarios remuestreados

Factor	Categorías	Niveles
HSE	15 HSE	15
Tipo de datos faltantes	MCAR, MAR, NMAR	3
Probabilidad de valores faltantes	30%, 50%	2
	<b>Total de escenarios</b>	90

cada HSE. En cuanto al tipo de dato faltante, existen tres formas principales mediante las cuales pueden presentarse ausencias en un conjunto de datos (Scheffer, 2002):

- **Completamente al azar (MCAR):** los valores perdidos son independientes de cualquier variable observada en el conjunto de datos. Para simular este mecanismo, se eliminaron valores de manera completamente aleatoria en los conjuntos de datos.
- **Al azar (MAR):** los datos faltantes dependen de variables observadas en el conjunto de datos, pero no de los valores faltantes en sí. Para simular este mecanismo, se utilizó el índice global estimado por la OCDE para cada HSE, estableciendo una probabilidad de ausencia en el ítem correspondiente. De acuerdo con la literatura, la relación fue inversa: a menor puntaje en la HSE, mayor probabilidad de datos faltantes.
- **No al azar (NMAR):** los datos faltantes dependen de medidas no observadas (ej. los valores faltantes en sí), pero es independiente de los datos observados. En este caso, los valores faltantes se generaron en función del ítem a imputar, eliminando datos aleatoriamente en las categorías más bajas (1 y 2).

Finalmente, se establecieron dos probabilidades de pérdida de valores para cada escenario: 30% y 50%. Debido a las restricciones establecidas en los escenarios MAR y NMAR, se seleccionaron probabilidades relativamente grandes para revelar la diferencia entre las estrategias de imputación (Wei Wu & Enders, 2015).

### 3.3 Imputación

En primer lugar, se ajustaron los modelos 1 y 2 con el conjunto de datos de entrenamiento para cada HSE, donde la variable respuesta y correspondía al ítem que se iba a imputar y el resto de los ítems de la misma HSE se utilizaron como variables predictoras  $x$ . Para esto, se utilizaron los paquetes `multinom` y `polr` de las librerías `nnet` y `MASS` respectivamente, en R versión 4.4.3 con RStudio 2024.12.1, macOS Mozilla/5.0

**Table 2:** Resumen de escenarios obtenidos

Tipo	Probabilidad	Promedio*
MCAR	30%	29.95%
MCAR	50%	49.85%
MAR	30%	13.59%
MAR	50%	22.67%
NMAR	30%	5.96%
NMAR	50%	9.88%

\*Promedio de la proporción de valores perdidos en las 15 HSE después del remuestreo.

(Macintosh; Intel Mac OS X 10\_15\_7). El código está disponible en GitHub (<https://github.com/vcardonas/imputacion-likert>).

Una vez estimados los parámetros de los modelos, se utilizaron para predecir los valores faltantes en cada escenario de remuestreo. Es decir, el Modelo Logit Acumulativo de Probabilidades Proporcionales y el Modelo Logit Multicategoría fueron aplicados a cada conjunto de datos incompletos generados en el paso anterior. Como resultado, se obtuvieron múltiples conjuntos de datos imputados.

### 3.4 Estrategia de comparación de modelos

Dado que el objetivo es evaluar exclusivamente los dos métodos de imputación mencionados en las secciones 2.2 y 2.2 bajo los diferentes escenarios seleccionados, el análisis se realizó únicamente sobre los datos con valores faltantes generados mediante remuestreo. Es decir, no se consideró el conjunto completo de prueba para la evaluación.

En esta etapa, cada conjunto de datos imputados se comparó con el conjunto de datos original para medir el rendimiento de la imputación. A continuación, se presentan los criterios de evaluación, basados en Jonsson and Wohlin (2004) and Oviedo Mozo (2024), utilizados para medir el desempeño de cada método:

- **Exactitud (Overall Accuracy):** Mide la proporción de valores correctamente imputados respecto al total de valores imputados. Se define como:

$$OA = \frac{A}{N}$$

donde  $A$  es el número de valores correctamente imputados y  $N$  el total de valores imputados. Se esperan valores cercanos a 1.

- **Curva Sumas Acumuladas (CS):** Ordena primero las predicciones exactas y luego incorpora progresivamente aquellas con diferencias de hasta 4

niveles en la escala Likert. La curva se representa con la distancia en el eje  $x$  y el porcentaje acumulado en el eje  $y$ .

- **Error Absoluto Medio (MAE):** Mide la distancia promedio entre los valores reales y los imputados incorrectamente. Aunque las categorías ordinales no poseen todas las propiedades matemáticas de una variable continua, el MAE es una métrica comúnmente reportada en la literatura (Niu et al., 2016). Se define como:

$$MAE = \frac{1}{N} \sum_{i=1}^N |y_i - \hat{y}_i|$$

donde:

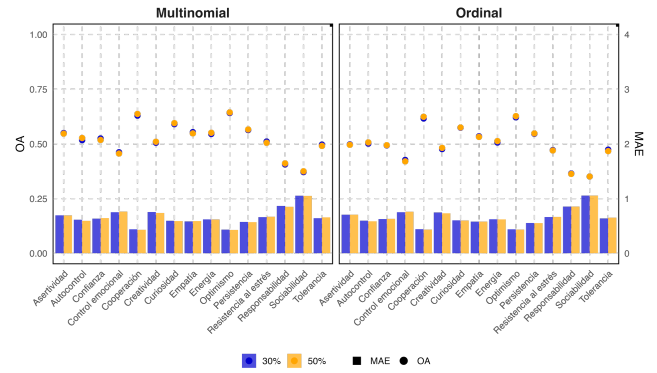
- $N$  es el número total de valores imputados.
- $y_i$  es el valor real del ítem para el individuo  $i$ .
- $\hat{y}_i$  es la categoría asignada por el modelo para el individuo  $i$ .

## 4 Resultados

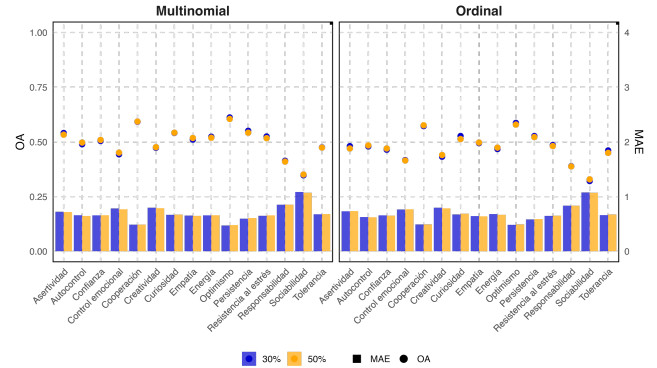
En esta sección se presentan los principales resultados obtenidos por los dos modelos de regresión logística bajo los diferentes escenarios de datos faltantes.

La Figura 1 compara el rendimiento de los modelos de regresión multinomial y ordinal en **Exactitud Global (OA)** y **Error Absoluto Medio (MAE)** bajo **MCAR**, con probabilidades de pérdida del 30% y 50%. El OA oscila entre 0 y 1, representando la proporción de valores correctamente imputados, mientras que el MAE, como promedio de distancias, varía entre 0 y 4. Ambos modelos presentan valores similares de OA, con variaciones entre HSE pero sin diferencias significativas según la probabilidad de pérdida. El modelo multinomial muestra mejor desempeño general, alcanzando su mayor OA (0.64) en optimismo. El MAE se mantiene por debajo de 1 en ambos modelos, indicando imputaciones cercanas a los valores reales, patrón que se repite en **MAR** (Figura 2).

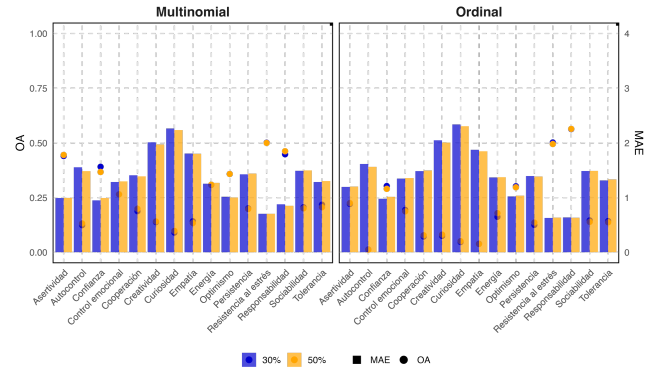
En contraste, bajo **NMAR** (Figura 3), el rendimiento se deteriora: el OA disminuye en HSE con buen desempeño previo y aumenta en las de menor exactitud, mientras que el MAE supera 1 en la mayoría de los casos.



**Figure 1:** Rendimiento de cada modelo en OA y MAE dadas probabilidades de pérdida en **MCAR**



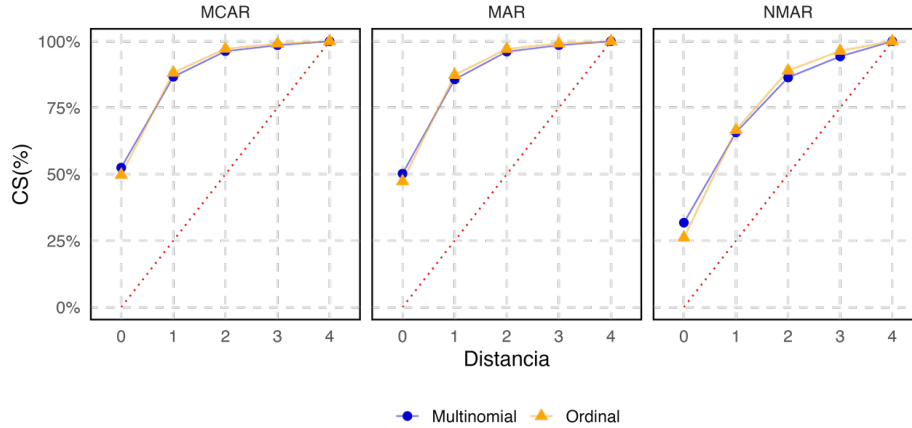
**Figure 2:** Rendimiento de cada modelo en OA y MAE dadas probabilidades de pérdida en **MAR**



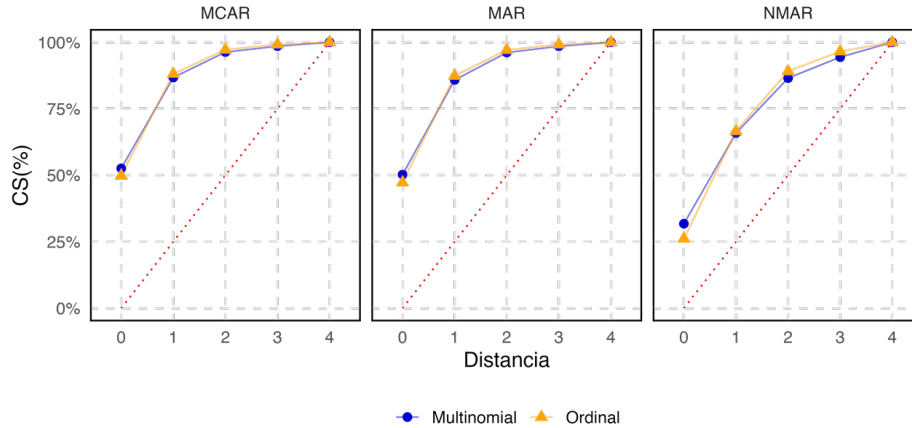
**Figure 3:** Rendimiento de cada modelo en OA y MAE dadas probabilidades de pérdida en **NMAR**

La Figura 4 presenta las curvas CS correspondientes a cada uno de los modelos seleccionados, desagregadas por el tipo de datos faltantes generados con una probabilidad de pérdida del 30%. De igual forma, la Figura 5 muestra los resultados para una probabilidad de pérdida del 50%. El comportamiento de ambos modelos no varía entre estas dos probabilidades.

Como se mencionó anteriormente, aunque el modelo multinomial muestra una mayor exactitud global, a medida que la distancia aumenta a 1, 2 o 3 unidades, el modelo ordinal presenta un rendimiento ligeramente superior en comparación con el multinomial. Por otro lado, NMAR sigue mostrando el peor desempeño entre los tipos de datos faltantes.



**Figure 4:** Sumas acumuladas para los modelos y tipos de datos faltantes con 30% de probabilidad de pérdida



**Figure 5:** Sumas acumuladas para los modelos y tipos de datos faltantes con 50% de probabilidad de pérdida

## 5 Discusión y Conclusiones

En este artículo se compara el desempeño de dos estrategias de imputación para datos faltantes en escalas Likert: la regresión logística multinomial y la regresión logística ordinal. El análisis se desarrolla bajo distintos escenarios de remuestreo, considerando el constructo evaluado, el mecanismo de generación de datos faltantes y la probabilidad de pérdida.

Desde un punto de vista teórico, ambos modelos presentan ventajas y limitaciones. La regresión logís-

tica ordinal incorpora el orden de las categorías de respuesta y es más parsimonioso, ya que requiere estimar un menor número de parámetros en comparación con la regresión logística multinomial. Sin embargo, este modelo asume el supuesto de probabilidades proporcionales, lo que implica que los coeficientes son constantes a lo largo de todas las categorías de la variable dependiente. En contraste, la regresión logística multinomial no impone esta restricción, permitiendo que los coeficientes varíen entre categorías, lo que la hace más flexible cuando el supuesto de probabilidades

proporcionales no se cumple.

Los resultados indican que ambos modelos como métodos de imputación tienen un desempeño similar, con ligeras diferencias. Aunque ambos predicen valores cercanos a los reales, la regresión multinomial muestra mayor precisión, mientras que la regresión ordinal presenta menor dispersión. Es decir, a medida que la distancia entre el valor imputado y el real aumenta, el modelo ordinal mejora su rendimiento en comparación con el multinomial. Por otro lado, la probabilidad de pérdida no parece afectar sustancialmente la precisión de la imputación.

El rendimiento de los modelos es más robusto bajo los mecanismos MCAR y MAR, mientras que se deteriora en NMAR. Esto sugiere inicialmente que el mecanismo de datos faltantes influye significativamente en la calidad de la imputación, destacando la importancia de considerarlo, en línea con la literatura existente. No obstante, al analizar las distribuciones de las respuestas en cada constructo evaluado, se observa que, bajo NMAR, los modelos presentan un mejor desempeño cuando las respuestas en la escala Likert siguen una distribución simétrica en comparación con MCAR y MAR. Esto indica que, en presencia de una distribución asimétrica, incluso con un tamaño de muestra grande ( $N = 37.835$ ), la calidad de la imputación bajo NMAR sigue viéndose afectada.

Se recomienda elegir el modelo más adecuado según el objetivo del análisis. Si el objetivo es minimizar errores pequeños y permitir imputaciones cercanas, aunque no exactas, el modelo ordinal es una mejor opción, especialmente cuando los análisis se realizan a nivel de puntuaciones totales en lugar de ítems individuales. Por otro lado, si se busca maximizar la imputación exacta para analizar cada ítem de manera independiente, el modelo multinomial es más adecuado.

Además, es fundamental considerar la distribución teórica subyacente esperada en cada constructo y el mecanismo de datos faltantes, ya que el desempeño de los modelos varía en función de estos factores. En particular, los modelos tienden a mostrar un mejor rendimiento en distribuciones asimétricas bajo mecanismos MCAR y MAR, mientras que en distribuciones simétricas el desempeño mejora bajo NMAR.

Como limitaciones, se destaca que, en estudios aplicados, los modelos de imputación pueden ser más complejos que en estudios simulados (Wei Wu & Enders, 2015). Aunque los enfoques basados en modelos de regresión logística son teóricamente sólidos, su aplicabilidad en contextos reales puede estar significativamente restringida por la necesidad de un tamaño

muestral elevado para obtener estimaciones precisas. Si bien en el presente estudio se cumplen ambas condiciones—un tamaño muestral grande y la modelación de una sola variable—estos resultados no pueden generalizarse a otros contextos con estructuras de datos más complejas.

Por lo tanto, para trabajos futuros, se recomienda explorar estrategias de imputación más robustas en contextos aplicados, donde las condiciones pueden diferir de los estudios simulados. Asimismo, es importante profundizar en el estudio de las particularidades del mecanismo NMAR, ya que sigue representando un desafío en la literatura (Jonsson & Wohlin, 2004). No obstante, en el presente estudio se han identificado algunas características importantes en el tratamiento de datos faltantes bajo este mecanismo, lo que sugiere la necesidad de seguir investigando enfoques que puedan mejorar su manejo y aplicabilidad en distintos contextos.

## References

- Likert, R. (1932). A technique for the measurement of attitudes. *Archives of psychology*.
- Muñiz, J. (2018). *Introducción a la psicometría*. Ediciones Pirámide.
- Wei Wu, F. J., & Enders, C. (2015). A comparison of imputation strategies for ordinal missing data on likert scale variables [PMID: 26610248]. *Multivariate Behavioral Research*, 50(5), 484–503. <https://doi.org/10.1080/00273171.2015.1022644>
- Downey, R. G., & King, C. (1998). Missing data in likert ratings: A comparison of replacement methods. *The Journal of General Psychology*, 125(2), 175–191. <https://doi.org/10.1080/00221309809595542>
- Raaijmakers, Q. A. W. (1999). Effectiveness of different missing data treatments in surveys with likert-type data: Introducing the relative mean substitution approach. *Educational and Psychological Measurement*, 59(5), 725–748. <https://doi.org/10.1177/0013164499595001>
- Hu, M., Salvucci, S. M., & Cohen, M. P. (1998). Evaluation of some popular imputation algorithms. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 308–313.
- Arcese, G., Lucchetti, M. C., Massa, I., & Valente, C. (2018). State of the art in s-lca: Integrating literature review and automatic text analysis. *The International Journal of Life Cycle Assessment*, 23,



- 394–405. <https://doi.org/10.1007/s11367-016-1135-4>
- Wongkamthong, C., & Akande, O. (2023). A comparative study of imputation methods for multivariate ordinal data. *Journal of Survey Statistics and Methodology*, 11(1), 189–212. <https://doi.org/10.1093/jssam/smac013>
- Homeister, L. (2024). *Application and comparison of missing data methods for factor analysis and multinomial logistic regression* [Master's thesis, University of Washington].
- Allison, P. D. (2005). Imputation of categorical variables with proc mi [Retrieved from <http://www2.sas.com/proceedings/sugi30/113-30.pdf>]. *SUGI 30 Proceedings*, 113–30, 1–14. <http://www2.sas.com/proceedings/sugi30/113-30.pdf>
- Van Buuren, S. (2018). *Flexible imputation of missing data* (2nd). CRC Press.
- Pan, S., & Chen, S. (2023). Empirical comparison of imputation methods for multivariate missing data in public health. *International Journal of Environmental Research and Public Health*, 20(2), 1524. <https://doi.org/10.3390/ijerph20021524>
- Köhler, C., Pohl, S., & Carstensen, C. H. (2015). Investigating mechanisms for missing responses in competence tests. *Psychological Test and Assessment Modeling*, 57(4), 499.
- Meinck, S., Cortes, D., & Tieck, S. (2017). Evaluating the risk of nonresponse bias in educational large-scale assessments with school nonresponse questionnaires: A theoretical study. *Large-scale Assessments in Education*, 5(3). <https://doi.org/10.1186/s40536-017-0038-6>
- Akinboboye, J., Ayanwale, M., Adewuni, D., & Vincent, Y. (2024). Impact of unanswered questions on examinees' latent traits: An item response theory perspective. *African Journal of Psychological Assessment*, 6(0), 9. <https://doi.org/10.4102/ajopa.v6i0.161>
- Durrant, G. B., & Schnepf, S. V. (2017). Which schools and pupils respond to educational achievement surveys?: A focus on the english programme for international student assessment sample. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 181(4), 1057–1075. <https://doi.org/10.1111/rssa.12337>
- Agresti, A., & Tarantola, C. (2018). Siple ways to interpret effects in modeling ordinal categorical data. *Statistica Neerlandica*, 72(3), 210–223. <https://doi.org/10.1111/stan.12130>
- McCullagh, P., & Nelder, J. A. (1989). Binary data. In *Generalized linear models* (pp. 98–148). Springer US.
- Agresti, A. (2018). *An introduction to categorical data analysis*, 3e. Wiley.
- Agresti, A. (2012). *Categorical data analysis*, 3e. Wiley.
- OCDE. (2019). *Assessment framework of the oecd study on social and emotional skills* (tech. rep.). OECD.
- Scheffer, J. (2002). Dealing with missing data. *Research Letters in the Information and Mathematical Sciences*, 3, 153–160.
- Jonsson, P., & Wohlin, C. (2004). An evaluation of k-nearest neighbour imputation using likert data. *10th International Symposium on Software Metrics, 2004. Proceedings.*, 108–118. <https://doi.org/10.1109/METRIC.2004.1357895>
- Oviedo Mozo, J. S. (2024). Estimación de índice de pobreza multidimensional (ipm) en bogotá d.c. y algunas ciudades cercanas usando imágenes satelitales. <https://repositorio.unal.edu.co/handle/unal/86359>
- Niu, Z., Zhou, M., Wang, L., Gao, X., & Hua, G. (2016). Ordinal regression with multiple output cnn for age estimation. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 4920–4928. <https://doi.org/10.1109/CVPR.2016.532>

Apéndice A

Distribución de respuestas en escala likert

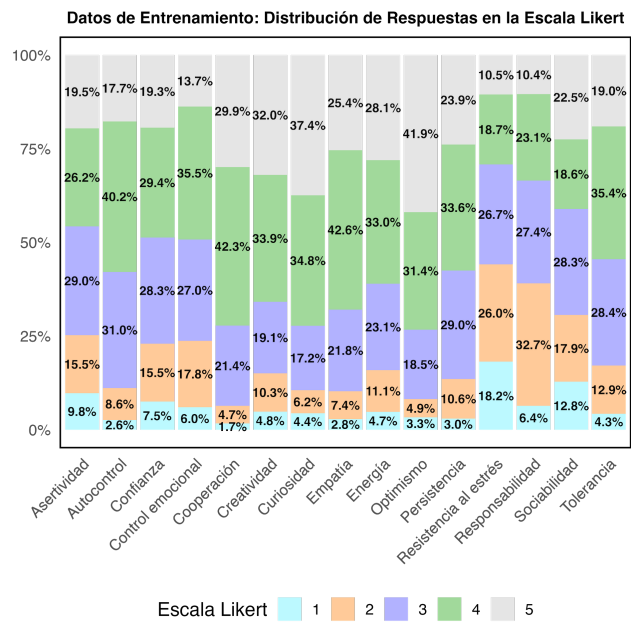


Figure 6: Distribución de los datos de entrenamiento

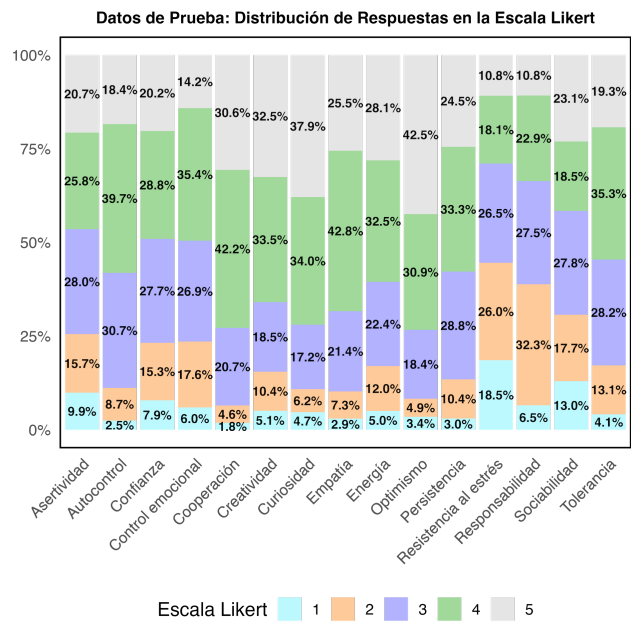


Figure 7: Distribución de los datos de prueba