Tarea 3: Econometría

Antonino Ávila [1], Cristóbal Donoso[2], Sebastián Rivera [3], Esteban Puentes [4], Camila Carrasco [5], y Santiago Garcıa [6]

1-3 Estudiantes
4 Profesor
5-6 Ayudantes

June 23, 2025

Objetivo general

Estimar el efecto de los síntomas depresivos sobre los resultados laborales utilizando modelos econométricos aplicados a datos longitudinales de la Encuesta ELSOC. El documento incluye la implementación manual de modelos Probit, 2SLS y de datos de panel.

Análisis Descriptivo de los Datos

El análisis se basa en datos del panel de la Encuesta Longitudinal Social de Chile (ELSOC) para el período 2016-2021. Antes de proceder con la construcción de variables y la estimación, es crucial realizar una inspección de la calidad de los datos brutos. La Figura presenta un el patrón de datos faltantes (missing values) en el conjunto de datos original.

El tratamiento de datos ausentes (missing values) es un paso crucial en la preparación del análisis econométrico, ya que las decisiones tomadas en esta etapa pueden tener un impacto significativo en los resultados. La estrategia óptima, depende fundamentalmente de la naturaleza de la variable afectada.

Se distinguen dos escenarios principales:

1. **Ausencia en la Variable Dependiente** (y_i): Cuando una observación carece de valor en la variable dependiente, resulta inutilizable para el propósito de la estimación. La función objetivo del estimador, ya sea la Suma de Cuadrados de los Residuos en Mínimos

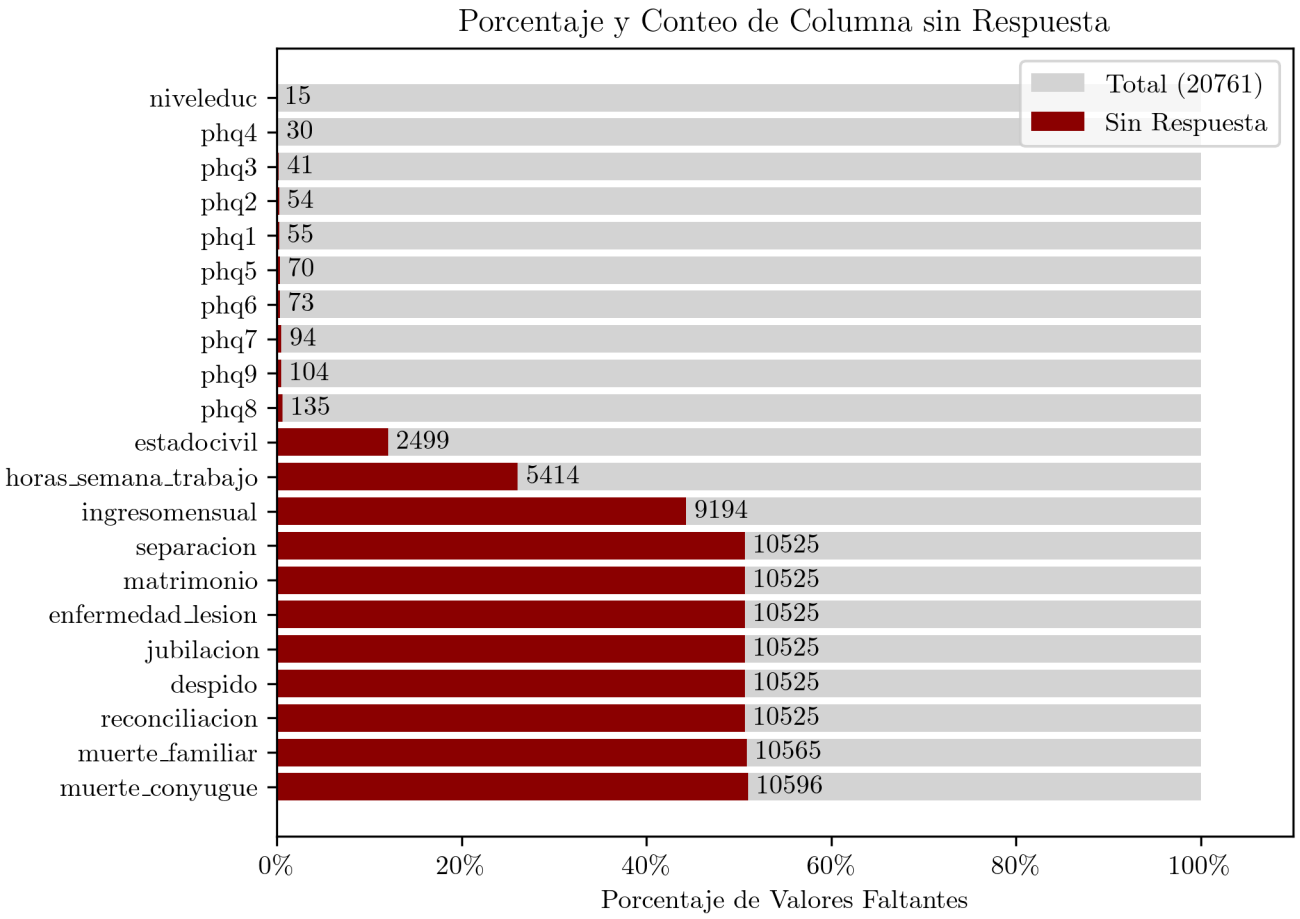

Figure 1: Distribución de datos faltantes en la muestra bruta de la encuesta ELSOC (2016-2021). Las celdas en rojo representan observaciones faltantes (NaN; del ingles Non a Number).

Cuadrados Ordinarios (MCO) o la función de Log-Verosimilitud en Estimación por Máxima Verosimilitud (MLE), no puede ser evaluada para dicha observación. Por lo tanto, el único curso de acción viable es la exclusión de estas observaciones de la muestra de estimación.

2. **Ausencia en las Variables Independientes** (X_i): Si el valor ausente se encuentra en una o más de las variables explicativas, existen varias alternativas. La más simple es la eliminación por lista (listwise deletion), que consiste en descartar la observación completa. Este método es insesgado bajo el supuesto de que los datos faltan de forma completamente aleatoria (MCAR, Missing Completely at Random). Otras técnicas, como la imputación (e.g., sustitución por la media) o la creación de variables indicadoras (dummy) para la ausencia, permiten retener un mayor número de observaciones. Sin embargo, estas últimas pueden introducir sesgos si el proceso de ausencia no es aleatorio (MAR o MNAR), potencialmente llevando a conclusiones erróneas sobre la relación entre las variables.

Dado el contexto de esta tarea y para garantizar la consistencia interna entre los distintos ejercicios, se optó por el enfoque de eliminación por lista para todas las variables clave. Este procedimiento, aunque reduce el tamaño muestral, asegura que los coeficientes estimados para los distintos modelos se basen en un conjunto de información idéntico y completo. No obstante, se debe mencionar que también se exploraron estrategias alternativas en algunas especificaciones como parte del análisis de robustez.

Construcción Indice de Depresión PHQ-9

La variable explicativa de interés en este estudio es una medida de la intensidad de los síntomas depresivos, construida a partir del **Cuestionario de Salud del Paciente-9 (PHQ-9)**. Este es un instrumento psicométrico, validado internacionalmente [2], que consiste en 9 ítems que evalúan la frecuencia de síntomas de depresión durante las últimas dos semanas. Cada ítem se puntúa en una escala de 0 a 3, correspondiendo a las respuestas "Nunca", "Algunos días", "Más de la mitad de los días" y "Casi todos los días".

La construcción de la variable final siguió los siguientes pasos:

1. Armonización de Respuestas. Se identificó que la encuesta contenía una quinta categoría de respuesta ("Todos los días"), la cual no pertenece a la escala estándar del PHQ-9. Para asegurar

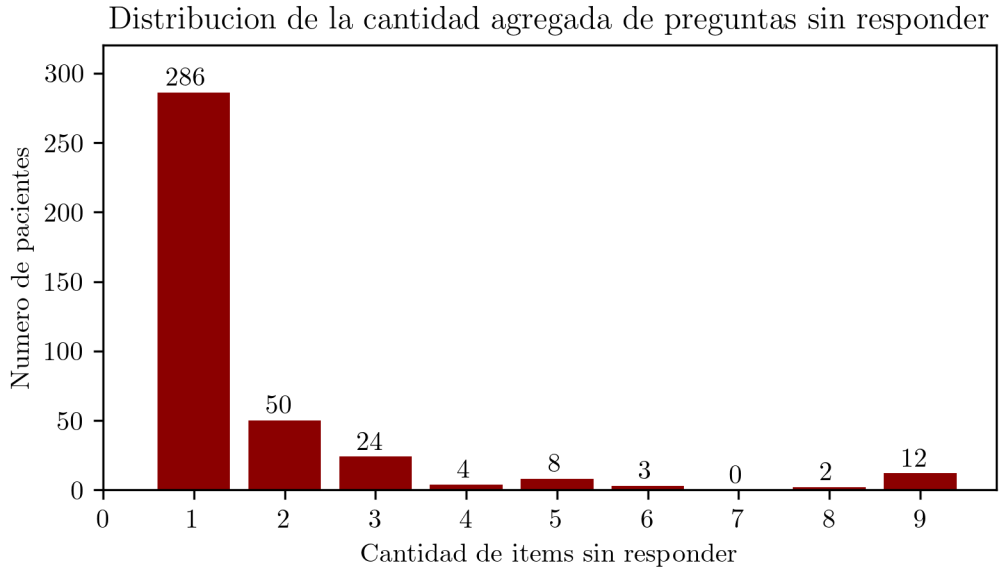

Figure 2: Distribución de cantidad de items sin responder en la encuesta PHQ-9 por paciante

la consistencia metodológica, se optó por fusionar esta categoría con la puntuación más alta y teóricamente más cercana ("Casi todos los días", con valor 3).

2. Cálculo del Puntaje. El puntaje total para el individuo i en el tiempo t, denotado como $\texttt{phq9\_score}_{it}$, se calcula como la suma de las puntuaciones de los 9 ítems:

$$\texttt{phq9\_score}_{it} = \sum_{j=1}^{9} \texttt{item\_j}_{it} \tag{1}$$

3. Manejo de Datos Ausentes. Para las observaciones con valores ausentes (*missing values*) en alguno de los 9 ítems, se consideró la alternativa de imputar un valor (por ejemplo, usando el promedio de las respuestas disponibles). Como se ve en la Figura 2, la mayoría de los pacientes omitió solo una de las 9 preguntas. Asi una imputacion podria ser una buena alternativa para no perder esas 286 personas.

Sin embargo utilizar un variable ficticio podría introducir un error de medición sistemático si la propensión a no responder está correlacionada con el nivel de depresión del individuo —un caso de ausencia no ignorable. Para evitar este potencial sesgo, se aplicó un criterio de *eliminación por lista*, excluyendo de la muestra final a toda observación que no tuviera los 9 ítems completos.

Nivel de Severidad de la Depresión	Puntaje PHQ-9
Mínima	1–4
Leve	5–9
Moderada	10–14
Moderadamente severa	15–19
Severa	20–27

Table 1: Niveles de Depresión según Puntaje PHQ-9

4. Descripción de la Variable Final. El resultado es una variable discreta que varía entre 0 y 27. La Tabla 1 detalla la clasificación clínica estándar de estos puntajes según el nivel de severidad de la depresión. Adicionalmente, la Figura 3 presenta la distribución empírica del puntaje PHQ-9 en la muestra de análisis, donde se puede observar la variabilidad de los síntomas en la población estudiada luego de eliminar los missing values.

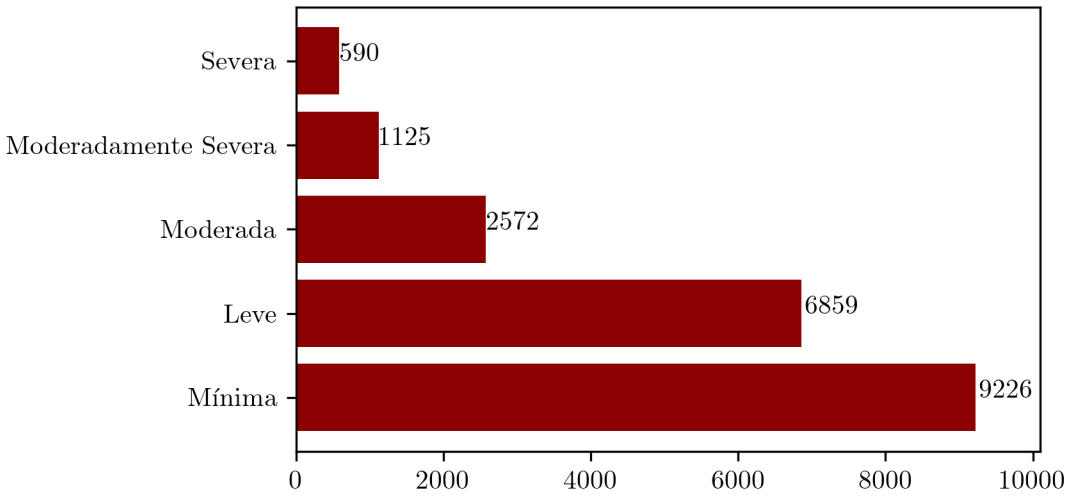

Figure 3: Distribución del Puntaje PHQ9 en la Muestra de Análisis

Variable Dependiente para el Modelo Probit: Participación Laboral

La variable dependiente para la estimación del modelo Probit es un indicador binario de participación en el mercado laboral, denotado como `participacion_laboral`. Esta variable no se encontraba directamente en los datos, por lo que fue construida a partir de la información sobre las horas trabajadas semanalmente (`horas_semana_trabajo`).

Se asignó un valor de 1 si el individuo reporta un número de horas trabajadas superior a cero, y

un valor de 0 en caso contrario (cero horas o dato no informado). La definición formal de esta transformación se presenta en la Ecuación 2:

$$\texttt{participacion\_laboral}_{it} = \begin{cases} 1 & \text{si } \texttt{horas\_semana\_trabajo}_{it} > 0 \\ 0 & \text{si } \texttt{horas\_semana\_trabajo}_{it} \leq 0 \text{ o es nulo} \end{cases} \tag{2}$$

La Figura 4 ilustra la distribución de la muestra según esta nueva variable. Como se puede observar, una proporción significativa de las observaciones originales cae en la categoría de "No Responde", que corresponde a los casos donde las horas trabajadas no fueron informadas (valores nulos).

Dado que la participación laboral es la variable dependiente del modelo, estas observaciones con valores ausentes deben ser excluidas del análisis. Este procedimiento de limpieza resulta en una muestra final para la estimación del modelo Probit de **15,093 observaciones**, lo que constituye aproximadamente el 74% del total de la muestra inicial.

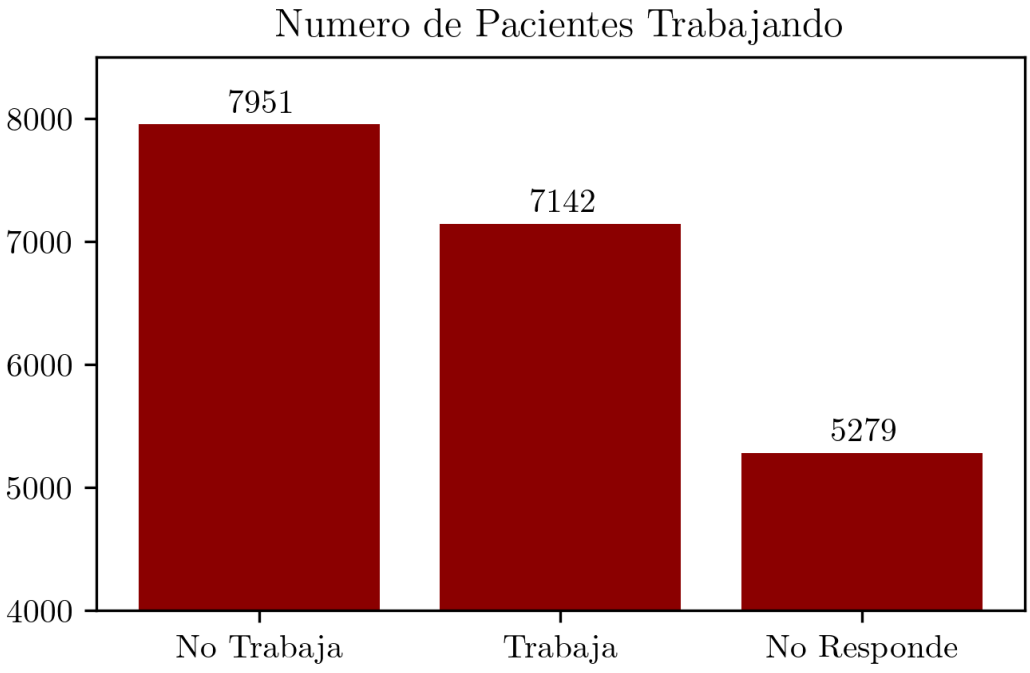

Figure 4: Distribución de la Participación Laboral, construida a partir de las horas trabajadas semanalmente.

Metodología de Preparación de Datos

Para transformar los datos brutos de la encuesta en un formato apto para la estimación econométrica, se implementó un flujo de trabajo sistemático y configurable. El objetivo principal de este proceso es construir una matriz de variables explicativas (X) y un vector de variable dependiente (y) que sean numéricamente coherentes y metodológicamente sólidos.

Definición de la Muestra de Análisis

El primer paso consistió en definir la muestra para cada modelo a estimar. Se aplicó un criterio de exclusión estricto basado en la **variable dependiente**: cualquier observación (individuo-ola) que presentara un valor ausente o no válido para la variable de resultado de interés fue eliminada del conjunto de datos. Este paso es fundamental, ya que dichas observaciones no aportan información a la función de verosimilitud o a los momentos del modelo.

Construcción de la Matriz de Regresores y Tratamiento de Datos Faltantes

Una vez definida la muestra en función de la variable dependiente, se procedió a la construcción de la matriz de regresores X. Para el manejo de datos faltantes en las propias variables explicativas, el proceso permite dos estrategias alternativas:

- **Estrategia 1: Análisis de Casos Completos (Eliminación por Lista).** Bajo este enfoque, se realiza una eliminación completa por lista. Cualquier observación que contenga uno o más valores ausentes en *cualquiera* de las variables seleccionadas (tanto dependientes como explicativas) es descartada. Este método, aunque potencialmente reduce de manera significativa el tamaño muestral, garantiza que el análisis se realiza sobre un subconjunto de datos sin ninguna imputación.
- **Estrategia 2: Retención de Observaciones con Imputación e Indicadores.** Este enfoque busca preservar el mayor número de observaciones posible. El tratamiento de los datos faltantes depende de la naturaleza de la variable:
  - **Variables Categóricas:** Si una variable categórica (e.g., "estado civil") contenía un valor ausente, este no se eliminaba. En su lugar, era tratado como una categoría

adicional y explícita (e.g., "Sin Respuesta"). Al momento de crear las variables indicadoras (*dummies*), esta categoría se trata como cualquier otra, permitiendo al modelo estimar el efecto asociado a la falta de respuesta.

- **Variables Numéricas:** Para las variables numéricas con valores ausentes (e.g., "edad"), se aplicó una técnica de **imputación por la media con una variable indicadora**. Este proceso implica dos acciones: 1) Se crea una nueva columna binaria que actúa como `FLAG`, tomando el valor 1 si el dato original estaba ausente y 0 en caso contrario. 2). En la columna original, el valor ausente es reemplazado por la media muestral de todas las observaciones no ausentes. Esta técnica permite retener la observación y, a la vez, controlar explícitamente por el potencial efecto de la ausencia del dato.

Ejemplo Ilustrativo de la Matriz Final

Para ilustrar el resultado de la **Estrategia 2**, la siguiente estructura representa la forma final de la matriz de regresores X para un conjunto hipotético de observaciones. Supongamos que la media de la variable `edad` en la muestra es 38.5 años.

	`intercept`	`edad`	`edad_missing`	`educ_Superior`
Individuo 1 (Completo)	1	45	0	1
Individuo 2 (Edad Faltante)	1	**38.5**	**1**	0
Individuo 3 (Completo)	1	25	0	0

(3)

En la Ecuación 3, se observa que para el Individuo 2, cuya edad era originalmente un dato faltante, el valor ha sido reemplazado por la media muestral (38.5) y se ha activado la variable indicadora `edad_missing` a 1.

Estandarización de las Variables

Como paso final opcional, la matriz de regresores X resultante puede ser **estandarizada**. En este procedimiento, cada columna (a excepción del intercepto) es transformada para tener una media de cero y una desviación estándar de uno. Esta transformación no altera la interpretación económica del modelo, pero a menudo mejora la estabilidad numérica y la velocidad de los algoritmos de estimación.

Parte 1: Modelo Probit

- Variable dependiente: indicador de participación laboral (1 si trabaja, 0 si no)
- Variable de interés: *score* PHQ-9 (debe resumir la información disponible en un único indicador de síntomas depresivos)
- Instrucciones:
  1. Construir la función de verosimilitud usando la CDF de la normal estándar
  2. Maximizar la verosimilitud en MATLAB (`fminunc` o similar)
  3. Calcular errores estándar usando la matriz de información observada
  4. Reportar efectos marginales para PHQ-9

Resultados

La Tabla 2 presenta los coeficientes estimados del modelo Probit que analiza la relación entre los síntomas depresivos PHQ-9 y la participación laboral:

Table 2: Resultados de la estimación Probit

Predictor	Coeficiente	Error Estándar	Efecto Marginal
Intercepto	0.2111	0.0208	0.0834
Puntaje PHQ-9	-0.0185	0.0025	-0.0073

Note que el coeficiente PHQ-9 es estadísticamente significativo, pues mediante el test t:

$$t : \frac{-0.0185}{0.0025} = -7.4,$$

que en valor absoluto es mayor a 1.96 (significativo al 5%) y el signo negativo indica que mayores síntomas depresivos se asocian con menor probabilidad de participación laboral. Además, los datos establecen que la probabilidad de participar en el mercado laboral disminuye en aproximadamente 0.73 puntos porcentuales (efecto marginal = -0.0073). Este efecto, que pareciera ser poco en términos unitarios, se vuelve más relevante cuando se consideran cambios clínicamente significativos en el PHQ-9. Por ejemplo, un aumento de 10 puntos (equivalente a pasar de una depresión mínima a moderada) reduciría la probabilidad de empleo en alrededor de 7.3 puntos porcentuales.

De igual manera, el intercepto también estadísticamente significativo:

$$t : \frac{0.2111}{0.0208} \approx 10.15.$$

Luego, los resultados en relación a él sugieren que en ausencia de síntomas depresivos, es decir, PHQ-9 $= 0$, la probabilidad base de participación laboral es del 8.34%. En suma, la significancia estadística de ambos coeficientes confirma la solidez de las estimaciones. Estos resultados son consistentes con la literatura previa ([1],[2]) que vincula la salud mental con menores tasas de empleo, aunque el diseño del estudio no permite inferir causalidad directa ya que podría existir endogeneidad; actores no observados (enfermedades crónicas) que afecten tanto la depresión como la empleabilidad.

Por otra parte, desde los demás resultados, se puede observar los errores estándar y los efectos marginales para PHQ-9 en la Figura 5. En el caso de los coeficientes, se puede ver que tanto para el intercepto como para PHQ-9 su error estándar tiene un valor menor a 0.005. Por otro lado, el error estándar de los efectos marginales está cercano a los 0.01 para el intercepto, mientras que el de PHQ-9 ronda los 0.05. Siguiendo con estos últimos, la figura de la derecha también indica como el valor del intercepto es cercano 0.08 y como el de PHQ-9 es cercano a -0.01.

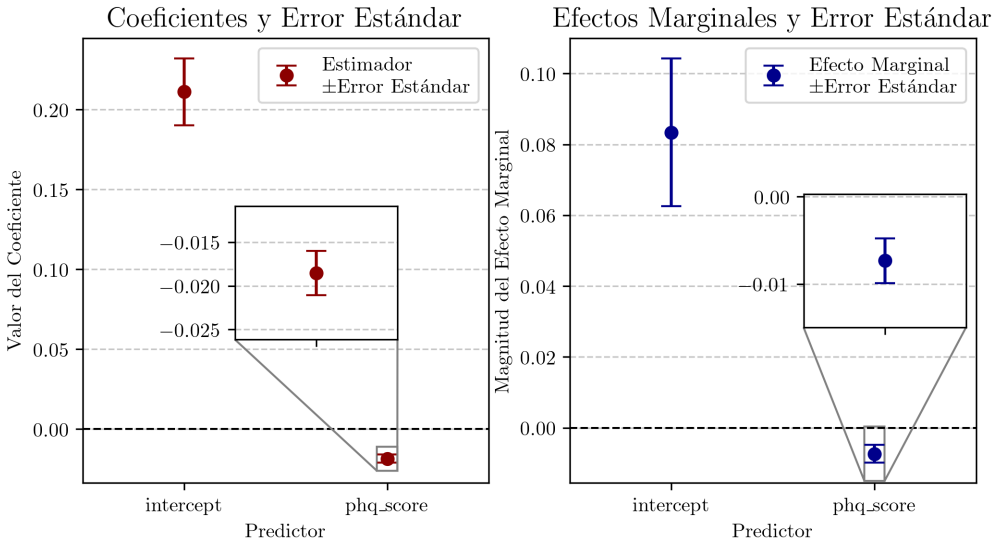

Figure 5: A la izquierda, la figura con el valor de los coeficientes estimados con su error estándar. La leyenda roja incluye estos dos. A la derecha, la figura con los efectos marginales de PHQ-9 con su error estándar. De forma análoga, la leyenda azul representa estos dos.

Teniendo ya la interpretación de los coeficientes, solo queda analizar el efecto marginal sobre la variable explicativa. El valor de -0.01 indica que un aumento de una unidad en PHQ-9 está asociado con una disminución de un punto porcentual en la probabilidad de participación laboral. En un caso hipotético esto se puede ver como que si el mismo individuo tuviese un puntaje PHQ-9 de 5 en lugar de 4, su probabilidad de participación laboral podría disminuir de 11% a 10%, con las demás variables constantes.

Con respecto a su poder estadístico, aproximando su error estándar a 0.03, el valor absoluto de su valor crítico queda cercano a 3.3, lo que nos indica que también es significativo al 5%. Finalmente, podrán encontrar la derivación de los efectos marginales en la sección de Apéndice.

En último lugar, es relevante mencionar que una limitación importante es que el modelo no captura posibles efectos no lineales, por ejemplo, si el impacto del PHQ-9 es más severo en niveles altos de depresión, ni diferencias por subgrupos (género, edad). Además, la eliminación de observaciones con datos faltantes, aunque necesaria, podría introducir sesgos si la ausencia de respuestas está correlacionada con la salud mental.

Parte 2: Modelo 2SLS

- Variable dependiente: log(ingresomensual)
- Variable endógena: *score* PHQ-9
- Controles: edad, sexo, educación, estado civil
- Instrumentos candidatos: *shocks* personales
- Instrucciones:
  1. Estimar el modelo usando MCO (coeficientes y errores estándar)
  2. Estimar el modelo 2SLS usando todos los instrumentos (coeficientes y errores estándar)
  3. Comparar resultados de OLS y 2SLS
  4. Calcule la primera etapa del modelo de 2SLS, indique si hay presencia de instrumentos débiles

Resultados

Los datos obtenidos están dispuestos en la Tabla 6.

Al comparar las estimaciones de MCO y 2SLS, se observa que el coeficiente del PHQ-9 disminuye considerablemente en magnitud al pasar de -0.0276 (MCO) a -0.0951 (2SLS) (una ponderación de $\lambda \in (3, 4)$), lo que sugiere que el estimador de MCO podría estar subestimando el verdadero efecto negativo de la depresión sobre los ingresos laborales. Sin embargo, el error estándar en 2SLS (0.0984) es casi cuatro veces mayor que en MCO (0.0258), lo que hace que el coeficiente no sea estadísticamente significativo al nivel convencional del 5%:

$$\begin{aligned} t_{\text{MCO}} &: \frac{-0.0276}{0.0257} \approx -1.0739 \\ t_{\text{2SLS}} &: \frac{-0.0951}{0.0984} \approx -0.9665. \end{aligned}$$

Además, el gran incremento en los errores estándar puede ser debido al uso de instrumentos débiles, lo que significa que las variables instrumentales utilizadas; shocks personales, podrían no estar suficientemente correlacionadas con la variable endógena PHQ-9.

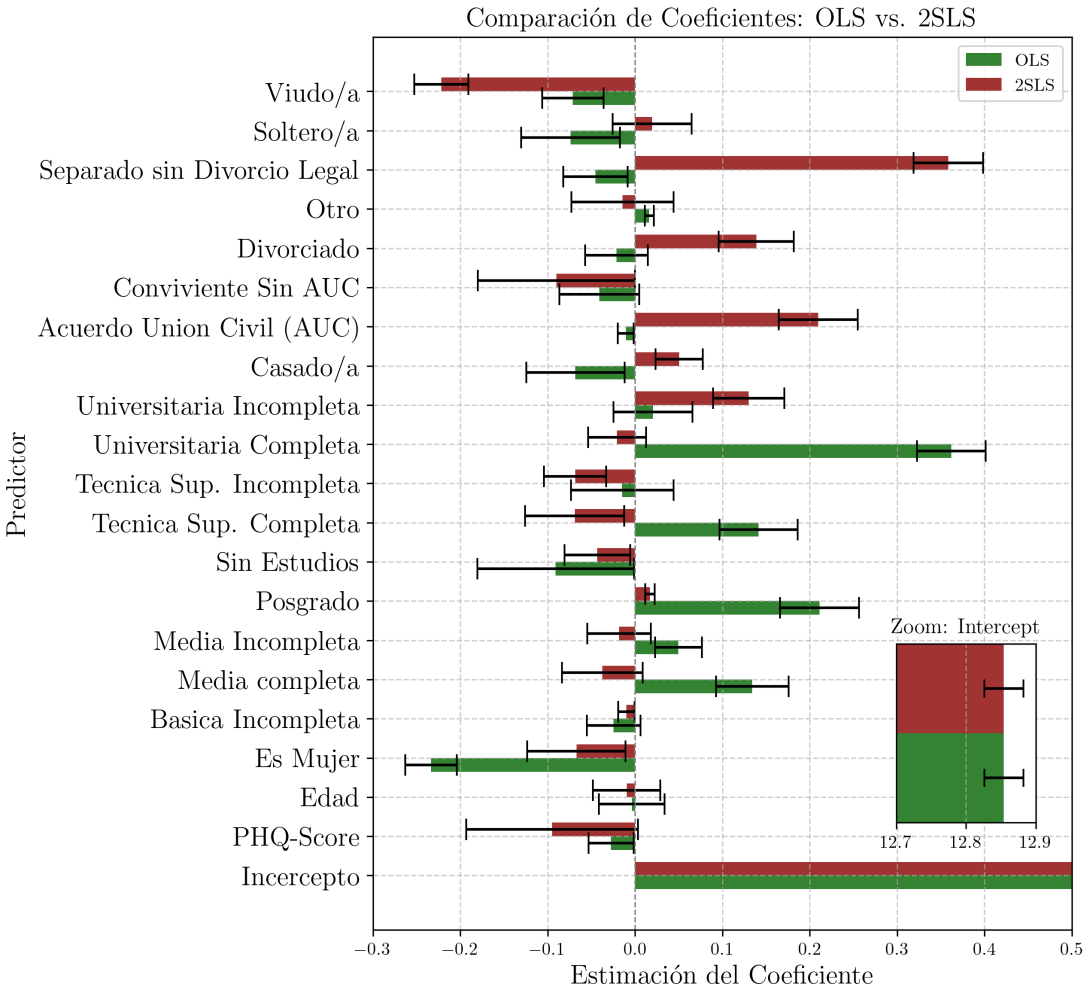

Figure 6: Impacto de la Endogeneidad en las Estimaciones de Coeficientes: OLS vs. 2SLS. El gráfico ilustra las diferencias en magnitud y significancia estadística de los predictores al ser estimados con un modelo OLS y un modelo 2SLS, este último diseñado para corregir por posible endogeneidad.

Particularmente en variables demográficas clave como el sexo, se observa que el efecto negativo de ser mujer sobre los ingresos se atenúa al pasar de MCO (-0.2336) a 2SLS (-0.0673), aunque en ambos casos los errores estándar relativamente altos hacen que estas estimaciones no sean estadísticamente significativas al 5%. Similarmente, el efecto de la edad, aunque pequeño en ambas especificaciones, se intensifica ligeramente en valor absoluto en 2SLS (-0.0097 versus -0.0036 en MCO).

Finalmente, estos resultados deben interpretarse considerando varias limitaciones metodológicas. Primero, como se mencionó, la posible debilidad de los instrumentos cuestiona la validez de las inferencias. Segundo, la muestra efectiva para 2SLS podría ser considerablemente menor que para MCO, dependiendo de la disponibilidad de datos para las variables instrumentales.

Parte 3: Modelo de datos de panel (Within)

- Variable dependiente: log(ingresomensual)
- Variable explicativa: PHQ-9 + controles
- Instrucciones:
    1. Estimar el modelo de datos de panel Within para estimar los parámetros del modelo
    2. Calcular errores estándar robustos
    3. Comparar con MCO y comentar posibles sesgos
    4. Comparar con MCO, 2SLS y el estimador Within y comentar posibles sesgos en cada caso

Resultados

Para estimar los determinantes del ingreso mensual, se utilizó un modelo de panel con efectos fijos, también conocido como estimador *within*. Este método es particularmente útil para controlar por toda la heterogeneidad inobservable que es constante en el tiempo para cada individuo. La Figura 7 presenta los resultados de tres especificaciones distintas de este modelo, permitiendo evaluar la sensibilidad de los coeficientes ante diferentes decisiones de modelado. Los resultados se desglosan en un panel general y en tres subgrupos temáticos para facilitar su inspección.

El hallazgo más notorio, visible en la Figura 7, es el coeficiente de gran magnitud y signo positivo para la variable **Edad**. En el contexto de un modelo de efectos fijos, este resultado es esperable e intuitivo: la edad no solo captura el paso del tiempo, sino que funciona como un proxy de la acumulación de experiencia laboral, la progresión de carrera y otros factores asociados al ciclo de vida que son los principales motores del crecimiento del ingreso para un mismo individuo.

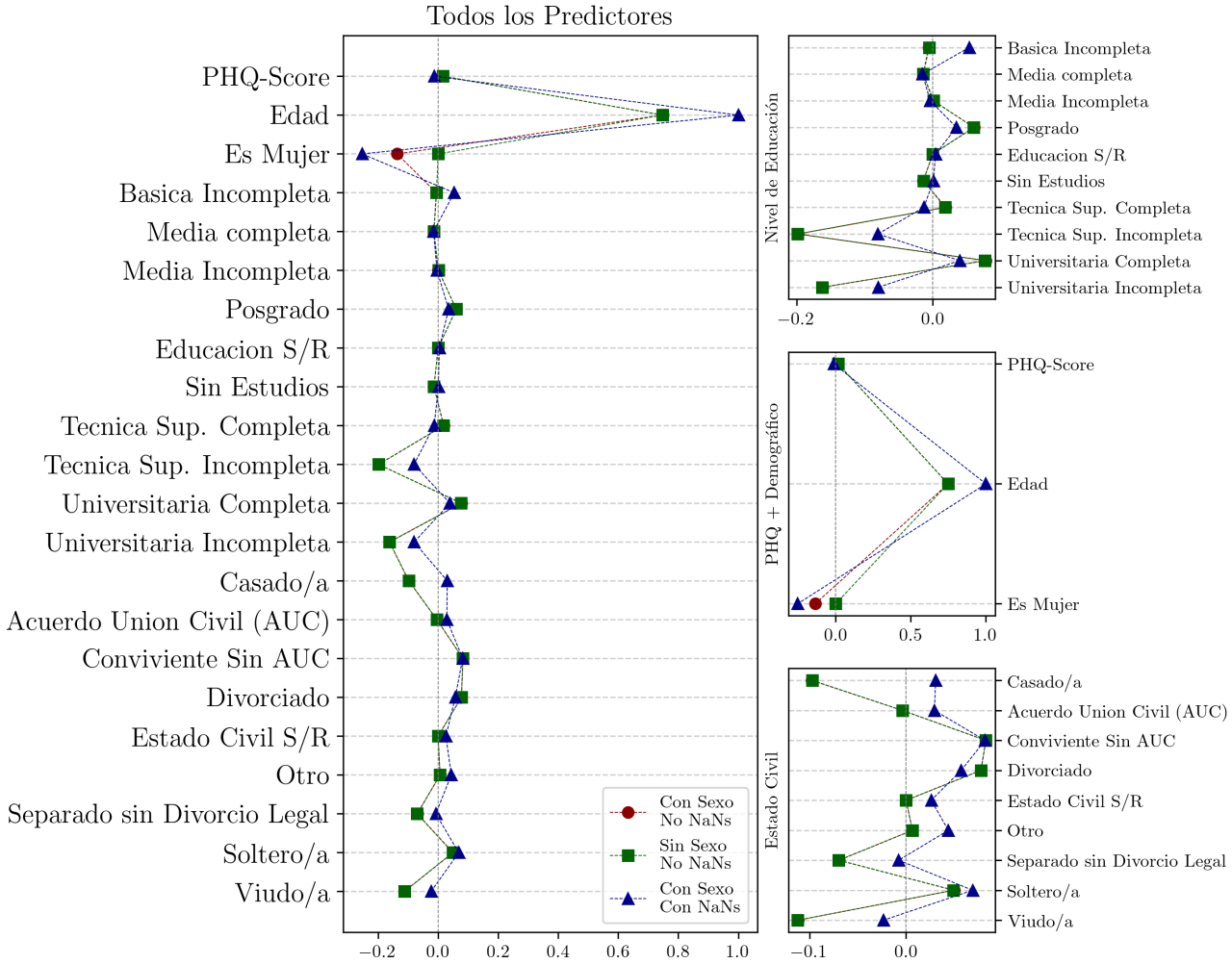

Figure 7: Comparación de coeficientes de un modelo de panel con estimador *within*, bajo tres especificaciones para la predicción del ingreso mensual. El panel izquierdo muestra el conjunto completo de predictores, mientras que los de la derecha detallan los resultados por subgrupos temáticos. Las especificaciones del modelo, detalladas en la leyenda, varían según la inclusión de la variable 'Sexo' y el tratamiento de datos faltantes (NaNs) mediante imputación (ver sección de Datos).

Para comprender por qué la edad domina los resultados y por qué los coeficientes de otras variables son cercanos a cero, es fundamental analizar la variabilidad de los predictores, la cual se muestra en la Figura 8. Este gráfico de dispersión revela dos puntos clave:

1. **Dispersión vs. Efecto:** Variables como el Nivel de Educación o el Estado Civil muestran una considerable dispersión en la muestra general (Figura 8), lo que indica que existen amplias diferencias *entre* individuos. Sin embargo, sus coeficientes en el modelo de efectos fijos (Figura 7) son marginales. Esto se debe a que el estimador *within* ignora la variación entre sujetos y se concentra únicamente en los cambios *dentro* del mismo sujeto a lo largo del tiempo. Los coeficientes pequeños sugieren que los eventos de cambio (e.g., casarse, obtener un título) fueron infrecuentes en la muestra o tuvieron un impacto limitado en la trayectoria de ingresos una vez controlados los efectos fijos.

2. **El Artefacto del Coeficiente de Género:** El caso de la variable **Es Mujer** ilustra un punto técnico crucial. La Figura 8 confirma visualmente que esta variable no presenta prácticamente ninguna variación (el boxplot es casi una línea plana). El coeficiente no nulo que se observa en la Figura 7 es, por lo tanto, un artefacto estadístico, ya que se identifica únicamente a partir del reducido subgrupo de 33 individuos para quienes esta variable reportó un cambio en el tiempo[1]. Consecuentemente, este coeficiente no debe ser interpretado como un efecto causal significativo.

En conclusión, el análisis conjunto de ambos gráficos demuestra la naturaleza específica del modelo de panel con efectos fijos. La robustez de los resultados a través de las tres especificaciones sugiere que, una vez que se controla por las características fijas e inobservables de los individuos, la trayectoria de vida asociada a la edad y la experiencia es el determinante predominante del crecimiento del ingreso, mientras que los cambios en características mayormente estáticas tienen un impacto considerablemente menor.

[1]El *término artefacto estadístico* se utiliza aquí en un sentido puramente técnico. El coeficiente se estima utilizando únicamente la variación de una submuestra muy pequeña ($N = 33$), lo que impide generalizar el resultado y lo hace altamente sensible a las características particulares de dicho grupo.

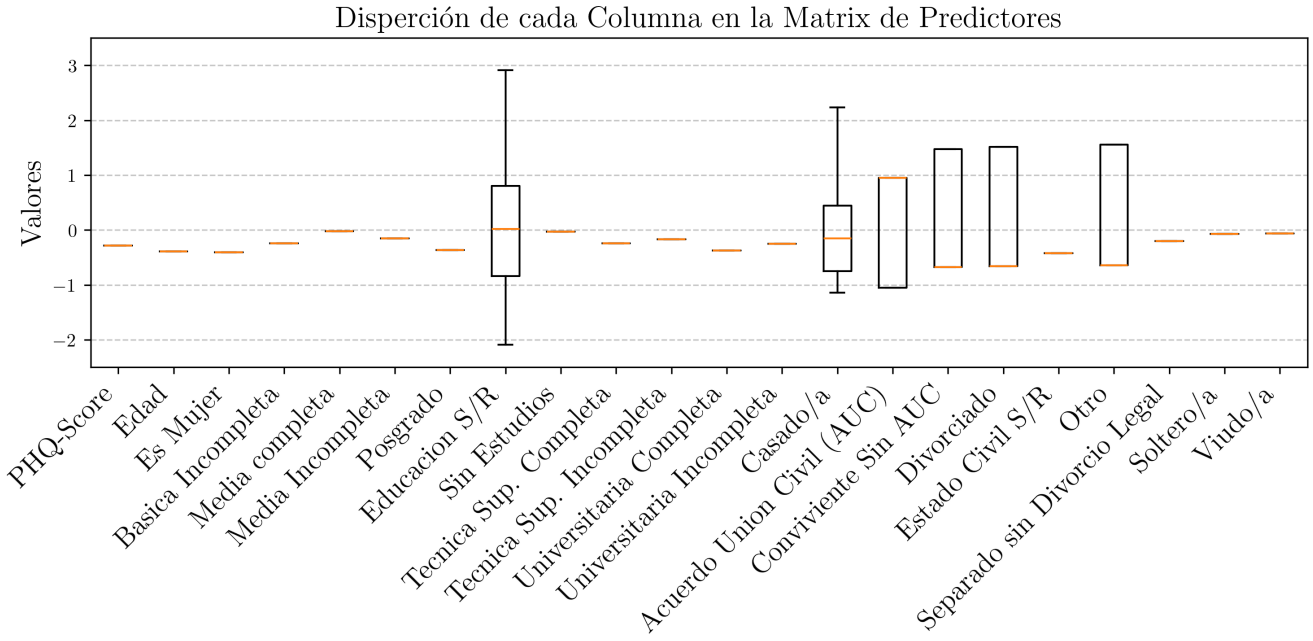

Figure 8: Caption

References

[1] S Bouwhuis, P C Koopmans, J W Groothoff, J J L van der Klink, and U Bültmann. Depression and work participation: A cross-sectional study in a large population-based sample. *Journal of Affective Disorders*, 256:550–557, 2019.

[2] Kurt Kroenke, Robert L Spitzer, and Janet BW Williams. The phq-9: validity of a brief depression severity measure. *Journal of general internal medicine*, 16(9):606–613, 2001.

Apéndice

Cálculo de efectos marginales

Sea $\Phi(\cdot)$ la función de distribución acumulada de la normal estándar con $X\beta = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k$. Siendo X_j la variable explicativa, el efecto marginal vendría siendo:

$$\frac{\partial P(Y = 1 \mid X)}{\partial X_j}$$

Desarrollando queda:

$$\frac{\partial P(Y = 1 \mid X)}{\partial X_j} = \frac{\partial \Phi(X\beta)}{\partial X_j} = \phi(X\beta) \cdot \frac{\partial (X\beta)}{\partial X_j} = \phi(X\beta) \cdot \beta_j$$

donde $\phi(\cdot)$ es la función de densidad de la normal estándar.