# Análisis de regresión

Jhanus Burgos, Bastian Liberona, Bruno Lefiche (grupo 1)

2025-11-14

## Introducción

El interés central de este estudio surge de comprender los factores que influyen en que un individuo tenga una ocupación formal o informal, ya que el estudio de la formalidad e informalidad laboral es fundamental para comprender el funcionamiento del mercado del trabajo en Chile. Para abordar este tema, se seleccionó la base de datos del ENE (Encuesta Nacional de Empleo) elaborada por el INE, la cual proporciona información anual sobre ocupación y desocupación del año 2024. La estructura mensual que ofrece la base de datos permite analizar variaciones a lo largo del año.

En base a lo anterior, se desarrollará una regresión múltiple que busca explicar cómo influyen el sexo, el nivel educativo, la posesión de un contrato y el sector laboral al que pertenece un individuo en la probabilidad de que este se encuentre ocupado formal o informalmente, considerando las variaciones mensuales y los distintos tramos de edad. Cabe destacar que, si bien la variable dependiente (Y) distingue entre ocupación formal e informal, la interpretación de los resultados se centrará en ocupación formal, dado que la dummy tomará el valor de 1 cuando el individuo tenga este tipo de ocupación.

## Pregunta de investigación

¿Cómo influyen el sexo, el sector laboral, el nivel educativo y la posesión de un contrato en la probabilidad de que una persona se encuentre ocupada formal o informalmente, considerando las variaciones mensuales a lo largo del año y los tramos de edad de los individuos?

## Datos Utilizados

El análisis se desarrolla a partir de la Encuesta Nacional de Empleo (ENE) elaborada por el Instituto Nacional de Estadísticas (INE) de Chile. Se utiliza la base anual correspondiente al año 2024, la cual se encuentra estructurada en doce mediciones cada una es de carácter mensual que permiten observar variaciones temporales dentro del mismo año.

La base de datos utilizada contiene 390.367 observaciones y 185 variables, entre las cuales se incluyen características sociodemográficas, educativas y laborales de los individuos, donde el conjunto de datos está formado mayoritariamente por variables categóricas, entre ellas sexo, región, tramo etario, nivel educativo, tipo de contrato,

etc. Estas variables permiten clasificar a la población en distintos grupos relevantes para el estudio y facilitan la comparación entre segmentos específicos dentro del mercado laboral.

Para la elaboración de la regresión múltiple se utilizaron los siguientes datos:

- Ocupación (ocup_form), el cual esta descrita en la base de datos como 1 (formal), 2 (informal), 999 (sin clasificación).
- Sector descrita como sector formal (1), sector informal (2), sector hogares como empleadores (3), sin clasificación (999)
- Sexo el cual esta descrito como hombre (1), mujer (2).
- contrato escrito (b8) descrito como si (1), no (2), 88 (no sabe), 99 (no existe),
- educación (edu2) descrita como nunca asistió (1), educación especial / diferencial (2), sala cuna o nivel medio de jardín infantil (3), pre kínder (4), kínder (5), educación básica (6), educación primaria (sistema antiguo) (7), educación media científico humanista o artística (8), educación media técnico profesional (9), educación humanidades (sistema antiguo) (10), educación técnica comercial, industrial o normalista (sistema antiguo) (11), técnico nivel superior (incluye suboficial ffaa) (12), profesional (incluye oficial ffaa) (13), diplomados o postítulos (de un semestre o más) (14), magíster (15), doctorado (16), no sabe (88), no responde (99).
- Termino nivel descrita como, sí (1), no (2), no sabe (88), no responde (99)
- Meses descrita como enero (1), febrero (2), marzo (3), abril (4), mayo (5), junio (6), julio (7), agosto (8), septiembre (9), octubre (10), noviembre (11), diciembre (12)
- Tramo de edad descrita como, 15 a 19 años (1), 20 a 24 años (2), 25 a 29 años (3), 30 a 34 años (4), 35 a 39 años (5), 40 a 44 años (6), 45 a 49 años (7), 50 a 54 años (8), 55 a 59 años (9), 60 a 64 años (10), 65 a 69 años (11), 70 años o más (12)

## Describir variables

La regresión múltiple tendrá como variable dependiente (Y) una variable dummy que indique si un individuo tiene una ocupación formal o informal (0=informal, 1=formal). El análisis se realizará en base a datos del año 2024 (separados mes a mes) del ENE, estos registros darán la posibilidad de ejecutar un estudio mensual sobre la variable dependiente Y. Por otro lado, las principales variables independientes serán:

| | Mean | SD | Min | Max | N |
|---|---|---|---|---|---|
| Sexo hombre | 0.54 | 0.50 | 0.00 | 1.00 | 123249 |
| Termino de educ | 0.80 | 0.40 | 0.00 | 1.00 | 123249 |
| Educ superior | 0.41 | 0.49 | 0.00 | 1.00 | 123249 |
| Educ postgrado | 0.04 | 0.19 | 0.00 | 1.00 | 123249 |
| Contrato escrito | 0.86 | 0.35 | 0.00 | 1.00 | 123249 |
| Sector de trabajo | 0.93 | 0.26 | 0.00 | 1.00 | 123249 |
| Joven | 0.20 | 0.40 | 0.00 | 1.00 | 123249 |
| Trabajador joven | 0.36 | 0.48 | 0.00 | 1.00 | 123249 |
| Jubilado | 0.13 | 0.34 | 0.00 | 1.00 | 123249 |
| Mes enero | 0.08 | 0.27 | 0.00 | 1.00 | 123249 |
| Mes febrero | 0.09 | 0.28 | 0.00 | 1.00 | 123249 |
| Mes marzo | 0.08 | 0.27 | 0.00 | 1.00 | 123249 |
| Mes abril | 0.09 | 0.28 | 0.00 | 1.00 | 123249 |
| Mes junio | 0.09 | 0.28 | 0.00 | 1.00 | 123249 |
| Mes julio | 0.08 | 0.28 | 0.00 | 1.00 | 123249 |
| Mes agosto | 0.09 | 0.29 | 0.00 | 1.00 | 123249 |
| Mes septiembre | 0.09 | 0.28 | 0.00 | 1.00 | 123249 |
| Mes octubre | 0.06 | 0.24 | 0.00 | 1.00 | 123249 |
| Mes noviembre | 0.08 | 0.27 | 0.00 | 1.00 | 123249 |
| Mes diciembre | 0.08 | 0.27 | 0.00 | 1.00 | 123249 |

- Sex_hombre: variable independiente dummy que toma el valor 1 si el individuo es hombre y 0 si es mujer. Originalmente, en la base de datos, mujer tomaba el valor de 2, pero para transformarlo a una variable dummy se optó por cambiarlo a 0.

- Sector_trabajo: variable independiente dummy que toma el valor 1 si el individuo pertenece al sector formal y 0 si pertenece a informal o al hogar. En un principio existían personas sin clasificación, pero se eliminaron, ya que no sabe a qué grupo pertenecen. Además, aquellos pertenecientes al sector informal y sector hogares como empleadores se encontraban separados en dos categorías, las cuales posteriormente se fusionaron y, para la varible dummy, toman el valor de 0.

- Contrato_escrito: variable independiente dummy que toma el valor de 1 cuando el individuo posee un contrato escrito y 0 en caso contrario. Originalmente, en la base de datos, si el individuo no tenía un contrato escrito tomaba el valor de 2, pero para transformarlo a una variable dummy se optó por cambiarlo a 0.

- Educación: variable categórica que indica el nivel educativo más alto alcanzado por el individuo, con 16 categorías distintas, desde sala cuna a doctorado. Ahora bien, considerando la variable dependiente (Y), para realizar un enfoque más especializado, se filtró la variable, eliminando niveles pre-escolares y los valores perdidos. Después, se agruparon los datos en tres categorías distintas: Educación Obligatoria (incluye a quienes nunca asistieron o alcanzaron hasta el nivel de educación media completa), Superior (Técnico Nivel Superior o Profesional) y Postgrado (Diplomados, Magíster y Doctorado). Por último, se crearon variables dummy que toman el valor de 1 si el nivel educacional corresponde y 0 en caso contrario. Para evitar la multicolinealidad perfecta, se optó por tomar como variable omitida “Educación Obligatoria”, por lo tanto, esta categoría será la base de referencia en torno a la interpretación.

- Termino_educ: variable independiente dummy que toma el valor de 1 cuando el individuo cumple con todos los requisitos para obtener la certificación del nivel educacional declarado como el más alto y 0 en caso contrario. En primera instancia existían aquellos individuos que no sabían o no respondieron, pero se eliminaron ya que no se sabe a qué grupo pertenecen. Además, los que no terminaron su nivel educacional, inicialmente en la base de datos, tomaban el valor de 2, pero para transformarlo a una variable dummy se optó por cambiarlo a 0.

- Factor Mes/es: variable categórica que señala el periodo de recolección de datos, convertida en 11 variables independientes dummy que toman el valor de 1 si el mes corresponde al evaluado y 0 en caso contrario. Como ejemplo: si el mes de estudio es enero, la variable dummy tomará el valor de 1 si el mes corresponde y 0 si es febrero, marzo, diciembre, etc. Se utilizarán 11 variables dummy porque incluir 12 variables independientes de meses generaría multicolinealidad perfecta. Para evitar este problema, se eliminará el mes de Mayo, por lo que será la categoría base al evaluar el factor mes/es. Dicho mes fue escogido porque, al comparar estadísticamente tal elección con respecto a las demás posibles, esta ofrecía una mejor interpretación de los datos.

- Tramo Edad: variable categórica dividida en 12 tramos que señalan la edad al momento de aplicar la encuesta, desde 15 a 19 hasta 70 años o más. A razón de no crear una variable dummy para cada tramo, se agruparon los datos en cuatro categorías: Joven (15 a 29 años), Trabajador Joven (30 a 44 años), Trabajador Adulto (45 a 59 años) y Jubilado (60 o más años). Posteriormente, se transformaron en variables dummy que toman el valor 1 cuando el tramo corresponde y 0 en caso contrario. Ahora bien, para evitar la multicolinealidad perfecta, se eliminó la categoría "Trabajador Adulto", por lo tanto, esta variable será la categoría base al interpretar los tramos de edad.

# Regresión

Considerando la composición de la regresión, todas las variables se especifican como dummies. Por lo tanto, independientemente del coeficiente β que se analice, si este toma un valor mayor que 0, es decir que es positivo, contribuye a la probabilidad de que un individuo se encuentre en una ocupación formal. En cambio, si el coeficiente es menor que 0, valor negativo, incide en la probabilidad de pertenecer a la ocupación informal.

$$
\begin{aligned}
OcupacionFormal_i &= \beta_0 + \beta_1 SexHombre_i + \beta_2 TerminoEduc_i + \beta_3 EducSuperior_i \\
&+ \beta_4 EducPostgrado_i + \beta_5 ContratoEscrito_i + \beta_6 SectorTrabajo_i \\
&+ \beta_7 Joven_i + \beta_8 TrabajadorJoven_i + \beta_9 Jubilado_i + \beta_{10} FactorMes_i \\
&+ u_i
\end{aligned}
$$

# Análisis de resultados

```
## Call:
## lm(formula = ocupacion_formal ~ sex_hombre + termino_educ + educ_super
ior +
##     educ_postgrado + contrato_escrito + sector_trabajo + joven +
##     trabajador_joven + jubilado + mes_ene + mes_feb + mes_mar +
##     mes_abr + mes_jun + mes_jul + mes_ago + mes_sep + mes_oct +
##     mes_nov + mes_dic, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.98654  0.02071  0.03887  0.05450  1.09787
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       0.032433   0.003432   9.450  < 2e-16 ***
## sex_hombre        0.018156   0.001299  13.972  < 2e-16 ***
## termino_educ      0.012781   0.001617   7.904 2.73e-15 ***
## educ_superior    -0.009078   0.001385  -6.553 5.66e-11 ***
```

```
## educ_postgrado   -0.028979   0.003436  -8.435  < 2e-16 ***
## contrato_escrito  0.880935   0.002106 418.348  < 2e-16 ***
## sector_trabajo    0.036883   0.002835  13.009  < 2e-16 ***
## joven            -0.031240   0.001854 -16.849  < 2e-16 ***
## trabajador_joven -0.006229   0.001576  -3.951 7.78e-05 ***
## jubilado         -0.131968   0.002104 -62.722  < 2e-16 ***
## mes_ene           0.005347   0.003027   1.766   0.0774 .
## mes_feb           0.005068   0.002968   1.708   0.0877 .
## mes_mar           0.003133   0.003025   1.036   0.3004
## mes_abr           0.002425   0.002989   0.811   0.4172
## mes_jun           0.005165   0.002987   1.729   0.0838 .
## mes_jul           0.003023   0.003004   1.006   0.3143
## mes_ago           0.004327   0.002950   1.467   0.1425
## mes_sep           0.003218   0.002999   1.073   0.2833
## mes_oct           0.003683   0.003284   1.121   0.2621
## mes_nov           0.001661   0.003052   0.544   0.5863
## mes_dic           0.002695   0.003056   0.882   0.3778
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2219 on 123228 degrees of freedom
## Multiple R-squared:  0.6722, Adjusted R-squared:  0.6722
## F-statistic: 1.264e+04 on 20 and 123228 DF,  p-value: < 2.2e-16
```

Se estimó un modelo poblacional utilizando mínimos cuadrados ordinarios (MCO), para determinar la probabilidad de tener una ocupación formal.

**La significancia estadística del modelo estimado**

Observando el (P-value) y el (T value), podemos determinar la relevancia estadística de los estimadores:

- Significancia global: observando el (P-value), nos indica que el modelo en conjunto es estadísticamente significativo, las variables independientes explican una parte relevante de tener una ocupación formal.
- Significancia individual: observando el (T value), las variables independientes (sexo, tramos de edad, nivel educativos, contrato y sector), son estadísticamente significativas al 1%, rechazando la hipótesis nula. Las variables independientes que no son estadísticamente significativas son las dummies meses, exceptuando los meses enero y febrero, esto sugiere que no hay evidencia estadística suficiente para rechazar la hipótesis nula en dichos meses. Un análisis general sugiere que las variables meses no son estadísticamente significativas.

**Interpretación de lo betas estimados**

**Probabilidad base (Intercepto):**

- Constante: Un individuo con las características base (Mujer, Educación Obligatoria, Sin contrato, Sector Informal, Adulta de 45-59 años, en el mes de Mayo) tiene una probabilidad base del 0,032 de ser formal.

**Variable sexo**

- Sex_hombre: Ser hombre, en comparación con el grupo base mujer, aumenta la probabilidad de ser formal en 1.8 puntos porcentuales, manteniendo todo lo demás constante.

**Variables educación. Respecto a educación obligatoria:**

- Termino_educ: Haber terminado el nivel educativo, en comparación con no haberlo terminado, aumenta la probabilidad de ser formal en 1.3 puntos porcentuales, manteniendo todo lo demás constante.
- Educ_superior: Tener educación superior, comparado con el grupo base (educación obligatoria), reduce la probabilidad de ser formal en 0.9 puntos porcentuales, manteniendo todo lo demás constante.
- Educ_postgrado: Tener un postgrado, comparado con el grupo base (educación obligatoria), reduce la probabilidad de ser formal en 2.9 puntos porcentuales, manteniendo todo lo demás constante.
- Variables laborales:
- Contrato_escrito: Tener un contrato escrito, comparado con no tenerlo, aumenta la probabilidad de ser formal en 88.1 puntos porcentuales, manteniendo todo lo demás constante.
- Sector_trabajo: Estar en el sector formal, comparado con el sector informal/hogares, aumenta la probabilidad en 3.7 puntos porcentuales, manteniendo todo lo demás constante.

**Variable edad. Respecto a Trabajador adulto:**

- Joven: Los jóvenes (15-29), comparados con el grupo base de adultos (45-59), tienen menos probabilidad de ser formales en 3.1 puntos porcentuales, manteniendo todo lo demás constante.
- Trabajador_joven: Los trabajadores jóvenes (30-44), comparados con el grupo base de adultos (45-59), tienen menos probabilidad de ser formales en 0.6 puntos porcentuales, manteniendo todo lo demás constante.
- Jubilado: Los mayores (60+), comparados con el grupo base de adultos (45-59), tienen menos probabilidad de ser formales en 13.2 puntos porcentuales, manteniendo todo lo demás constante.

**Variable temporal (Meses). Respecto al mes base Mayo:**

- Mes_ene: El mes de enero tiene 0.5 puntos porcentuales más de probabilidad de formalidad respecto a mayo.
- Mes_feb: El mes de febrero tiene 0.5 puntos porcentuales más de probabilidad de formalidad respecto a mayo.
- Mes_mar: El mes de marzo tiene 0.3 puntos porcentuales más de probabilidad de formalidad respecto a mayo.
- Mes_abr: El mes de abril tiene 0.2 puntos porcentuales más de probabilidad de formalidad respecto a mayo.
- Mes_jun: El mes de junio tiene 0.5 puntos porcentuales más de probabilidad de formalidad respecto a mayo.
- Mes_jul: El mes de julio tiene 0.3 puntos porcentuales más de probabilidad de formalidad respecto a mayo.
- Mes_ago: El mes de agosto tiene 0.4 puntos porcentuales más de probabilidad de formalidad respecto a mayo.
- Mes_sep: El mes de septiembre tiene 0.3 puntos porcentuales más de probabilidad de formalidad respecto a mayo.
- Mes_oct: El mes de octubre tiene 0.4 puntos porcentuales más de probabilidad de formalidad respecto a mayo.
- Mes_nov: El mes de noviembre tiene 0.2 puntos porcentuales más de probabilidad de formalidad respecto a mayo.
- Mes_dic: El mes de diciembre tiene 0.3 puntos porcentuales más de probabilidad de formalidad respecto a mayo

**Principales observaciones de los estimadores**

Una de las principales observaciones que tenemos con los estimadores son con respecto las variables independientes contrato_escrito y sector_trabajo, las cuales provocan un problema de tautología. Esto ocurre porque ambas variables no actúan como causas externas de la formalidad, sino que son variables que asemejan a la variable dependiente.

Incluir estas dos variables genera una absorción artificial de la varianza del modelo, inflando el R cuadrado a 0.67, quitándole peso a las variables que actúan como causas externas a la formalidad como la educación o la edad.

Respecto a los tramos de edad, los coeficientes calculados respecto a un trabajador adulto representan:

- Entrada al mercado: Los jóvenes (15-29 años) enfrentan un valor negativo de 3.1 puntos porcentuales, esto puede deberse a las dificultades que enfrentan los jóvenes en las primeras experiencias laborales con respecto a la formalidad de los labores que desempeñan.
- Salida del mercado: El valor negativo se intensifica en los mayores de 60 años (-13.2 puntos porcentuales), sugiriendo que los adultos mayores, al jubilarse, tienden a trabajar en empleos informales.

En educación se observa un resultado contraintuitivo, ya que tener educación superior o postgrado presenta coeficientes negativos (-0.9 y -2.9 puntos porcentuales respectivamente) en comparación con la educación obligatoria.

- Dichos coeficientes no implican que educarse reduzca la formalidad de un individuo, sino más bien, que el modelo está representado a aquellos profesionales independientes, que a pesar de tener altos niveles de estudios no poseen un contrato de trabajo como tal, por lo tanto, son clasificados como informales.

Respecto a las variables dummies de meses, ningún mes es estadísticamente significativo, exceptuando enero y febrero, donde no hay evidencia estadística suficiente para rechazar la hipótesis nula. Concluyendo que la probabilidad de tener un empleo formal en el año 2024 no es afectada por los meses del año.}

Por último, respecto a la variable sexo, los coeficientes demuestran que los hombres tienen una ligera tendencia a estar en empleos formales respecto a las mujeres.

# Conclusión

El modelo muestra que la formalidad laboral está influida principalmente por características como edad, sexo y educación, aunque la inclusión de variables como contrato y sector laboral genera un problema de tautología que sobreestima el poder explicativo del modelo. Aun así, se observa que:

- Los jóvenes y adultos mayores presentan menor formalidad respecto a un trabajador adulto.
- Los hombres tienen una ligera ventaja sobre las mujeres con respecto a la formalidad de su labor.
- La educación refleja la presencia de profesionales independientes más no que tener más formación implica menos formalidad.
- Los meses del año no afectan la formalidad.