

**Quinta guía de Ayudantía Estadística III**

**Profesor:** Catalina Canals, Eduardo Toro **Ayudante:** Gabriel Sotomayor

Contenido:

1. Regresión Lineal simple y múltiple
2. Estimación del modelo
3. Comprobación de supuestos

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

# Regresión lineal simple y múltiple

La regresión lineal es una técnica de análisis estadístico que nos permite estimar los efectos de ciertas variables (independientes o predictores) en una variable **cuantitativa** (dependiente o predicha).

Dependiendo de la cantidad de variables independientes utilizadas, esta técnica se distingue en regresión lineal simple (una variable independiente) o regresión lineal múltiple (dos o más variables independientes).

Este modelo solo sirve para relaciones entre variables que sean lineales, es decir, relaciones que se comporten según la formula $Y=α+βX+ε$ para el caso de la regresión lineal simple, siendo $α$ el intercepto, es decir el valor que toma la varialbe dependiente cuando la variable independiente es 0, $β$ la pendiente es decir, el efecto en Y de un aumento de 1 en X y $ε$ el error, a saber, el efecto de variables no observadas que influyen en Y. Para regresiones múltiples la fórmula es similar, con la diferencia de que incluye varias variables independientes $Y=α+β\_{1}X\_{1}…+β\_{K}X\_{K}+ε$, siendo k el número de variables independientes incluidas en el modelo.

# Estimación del modelo

Para estimar un modelo de regresión lineal debemos ir a analizar 🡪 regresión 🡪 lineal. Se abrirá una pestaña en la que tendremos que introducir las variables dependientes e independientes.



|  |
| --- |
| **Resumen del modelob** |
| Modelo | R | R cuadrado | R cuadrado ajustado | Error estándar de la estimación | Estadísticas de cambios |
| Cambio de cuadrado de R | Cambio en F | df1 | df2 | Sig. Cambio en F |
| 1 | ,397a | ,158 | ,153 | 3,03961 | ,158 | 31,808 | 4 | 679 | ,000 |
| a. Predictores: (Constante), Nota al gobierno: Disminuir la delincuencia , Sexo (Control Hombre), Posición política (1 izquierda - 10 derecha), Percepción de efectividad de las medidas punitivas |
| b. Variable dependiente: Legitimidad de las detenciones ciudadanas |

 En este caso utilizaremos como variable dependiente un índice de legitimidad de las detenciones ciudadanas[[1]](#footnote-1), y como variables independientes un índice de percepción de efectividad de las medidas punitivas contra la delincuencia[[2]](#footnote-2), el sexo[[3]](#footnote-3), la posición política y la nota al gobierno respecto de su capacidad de disminuir la delincuencia. Para obtener estadísticas referentes al ajuste del modelo debemos ir a la pestaña estadísticos y marcar “cambio en R cuadrado”. Una vez apretemos aceptar SPSS nos entregará dos tablas relevantes a interpretar, el resumen del modelo y los coeficientes.

El valor de R cuadrado nos indica el ajuste global de la regresión, es decir, la proporción de la varianza de la variable dependiente que es explicada por las independientes, en este caso un 15,8%, mientras que el R cuadrado ajustado, es una modificación del estadístico anterior que penaliza por cantidad de variables independientes (ya que añadir variables siempre aumenta el ajuste del modelo). Por otro lado, el valor P asociado al estadístico F (columna Sig. Cambio en F) nos permite poner a prueba la hipótesis nula de que los coeficientes beta asociados a todos los predictores son iguales a 0. Al obtener valores p menores a 0,05 (95% de confianza) o 0,01 (99% de confianza) rechazaríamos la hipótesis nula, y por ende, concluiríamos que al menos uno de los coeficientes beta, en la población, es distinto de 0. Esto quiere decir que podemos afirmar que en la población, al menos uno de los predictores del modelo efectivamente influye en la variable dependiente. Hay que tener presente que la significación del estadístico F nunca es exactamente 0 (SPSS lo aproxima), por lo cual a la hora de presentar resultados se debe señalar que se obtuvo un p<0,01.

|  |
| --- |
| **Coeficientesa** |
| Modelo | Coeficientes no estandarizados | Coeficientes estandarizados | t | Sig. | Estadísticas de colinealidad |
| B | Error estándar | Beta | Tolerancia | VIF |
| 1 | (Constante) | 4,046 | ,654 |  | 6,188 | ,000 |  |  |
| Percepción de efectividad de las medidas punitivas | ,422 | ,042 | ,357 | 9,939 | ,000 | ,960 | 1,042 |
| Sexo (ref. Hombre) | -,550 | ,233 | -,083 | -2,357 | ,019 | ,994 | 1,006 |
| Posición política (1 izquierda - 10 derecha) | ,102 | ,063 | ,058 | 1,616 | ,107 | ,972 | 1,028 |
| Nota al gobierno: Disminuir la delincuencia | -,139 | ,079 | -,064 | -1,769 | ,077 | ,961 | 1,040 |
| a. Variable dependiente: Legitimidad de las detenciones ciudadanas |

Esta tabla nos entrega los coeficientes que nos permiten interpretar el efecto de cada una de las variables independientes sobre la dependiente y escribir la ecuación de la regresión. La constante corresponde a $ α$ y nos indica el valor de la variable dependiente en caso de que el valor de todas las variables independientes sea 0. Los B (no confundir con Beta estandarizados) nos indican el efecto de cada una de la VI en la VD, controlando por las demás VI, y corresponden a los $β$ de la ecuación. Es importante tener presente que estos coeficientes son estimaciones, por lo cual tienen asociado un error estándar a partir del cual es posible construir intervalos de confianza para los efectos (beta) poblacionales, lo cual en SPSS puede encontrarse en la pestaña estadísticos dentro de la función regresión lineal.

Podemos señalar por tanto que la ecuación del modelo es:

$Legitmidad de las DC=4,046+(0,422\*efectividad de la medida punitivas)+(-0,550\*sexo)+(0,102\*posicion politica)+(-0,139\*evaluacion del gobierno en delincuencia)$

Por último, el valor P (Columna Sig.) asociado al test T (equivalente al test de Wald) nos permite poner a prueba la hipótesis nula de que el efecto de cada variable independiente es igual a cero (es decir, que la variable independiente no influye en la variable dependiente para la población). Al ser p<0,05 (con un 95% de confianza) rechazamos H0 concluyendo que la VI influye en la VD.

Respecto de la interpretación de las variables podemos señalar que:

1. Por cada punto adicional en el índice de percepción de efectividad de las medidas punitivas contra la delincuencia, la legitimidad de las DC predicha por el modelo aumenta en 0,422 puntos, controlando por sexo, posición política y evaluación del gobierno en disminución de la delincuencia.
2. Las mujeres legitiman 0,55 puntos menos que los hombres las detenciones ciudadanas, controlando por percepción de efectividad de la medidas punitivas, posición política y evaluación del gobierno en disminución de la delincuencia.
3. Por cada punto adicional en la escala de posición política (es decir, una posición más a la derecha) la legitimidad de la DC predicha por el modelo aumenta en 0,102 puntos, controlando por sexo, percepción de efectividad de la medidas punitivas y evaluación del gobierno en disminución de la delincuencia.
4. Por cada punto adicional de evaluación de la capacidad del gobierno de disminuir la delincuencia, disminuye en 0,139 la legitimación de las DC predicha por el modelo, controlando por sexo, percepción de efectividad de la medidas punitivas y posición política.

# Comprobación de supuestos

Para una correcta aplicación de un modelo de regresión lineal múltiple debemos comprobar que los datos cumplen con 7 supuestos que de no cumplirse tendrán diferentes consecuencias negativas en el modelo.

## Relación lineal entre las variables

En primer lugar, debemos comprobar que la relación entre las variables sea lineal para que esta pueda ser descrita por una RL, en caso contrario el modelo tendrá un ajuste muy bajo.

Para esto podemos utilizar una evaluación gráfica de la relación entre las variables, mediante un gráfico de dispersión, los cuales podemos encontrar en Gráficos 🡪 Cuadro de diálogos antiguos 🡪 Dispersión/Puntos



En este punto, el programa nos ofrece 5 tipos de gráficos de dispersión, para dar cuenta de una relación lineal bivariada debemos utilizar un gráfico de dispersión simple. El problema de esta forma de comprobar el supuesto es que solo sirve para n pequeños, ya que para n mayores el gráfico se hace in interpretable.

Otra forma de observar relación lineal entre las variables independientes y la dependiente es utilizar correlaciones bivariadas, las cuales han sido revisadas en guías anteriores.

## Ausencia de multicolinealidad entre las variables independientes

Para poder utilizar esta técnica las variables independientes deben ser independientes entre sí, es decir, no estar correlacionadas. De no cumplirse esta condición no es posible distinguir el efecto de cada una de las variables correlacionadas, lo cual además producirá altos errores estándar y falta de precisión en los coeficientes calculados. En un caso extremo de colinealidad no se podrá calcular el modelo.

Para verificar la ausencia de multicolinealidad entre las variables es necesario calcular una matriz de correlaciones bivariadas, lo cual podemos realizar desde la pestaña de regresión lineal, donde debemos ir a la pestaña estadísticos y marcar “matriz de covarianzas”.



Una vez hayamos marcado esta casilla, al correr el modelo nos aparecerá una tabla extra con una matriz de correlaciones, en la cual debemos comprobar que no existan correlaciones altas. Si dos variables independientes tienen una correlación mayor a 0,8 debe evaluarse si eliminar a una de las dos del modelo (probar con cual se obtienen mejores resultados) o crear un índice a partir de ambas.

|  |
| --- |
| **Correlaciones de coeficientea** |
| Modelo | Nota al gobierno: Disminuir la delincuencia | Sexo (Control Hombre) | Posición política (1 izquierda - 10 derecha) | Percepción de efectividad de las medidas punitivas |
| 1 | Correlaciones | Nota al gobierno: Disminuir la delincuencia | 1,000 | ,015 | ,108 | ,153 |
| Sexo (Control Hombre) | ,015 | 1,000 | -,052 | ,063 |
| Posición política (1 izquierda - 10 derecha) | ,108 | -,052 | 1,000 | -,099 |
| Percepción de efectividad de las medidas punitivas | ,153 | ,063 | -,099 | 1,000 |
| a. Variable dependiente: Legitimidad de las detenciones ciudadanas |

En este caso todas nuestras correlaciones tienen valores muy bajos, menores a 0,2, lo cual da cuenta de ausencia de correlación.

La multicolinealidad también puede evaluarse mediante el factor de inflación de la varianza (VIF), para esto, al momento de correr el modelo de regresión lineal debemos ir a estadísticos y solicitar diagnósticos de colinealidad.



El VIF nos indica cuanto aumenta el error estándar debido a problemas de colinealidad. La raíz cuadrada de VIF corresponde a el efecto de la multicolinealidad en el error estándar, es decir, si obtenemos un VIF de 4, el error estándar será el doble de lo que sería sin problema de multicolinealidad. Un VIF de hasta 3 resulta aceptable. En este todos los VIF están bajo ese valor.

|  |
| --- |
| **Coeficientesa** |
| Modelo | Coeficientes no estandarizados | Coeficientes estandarizados | t | Sig. | Estadísticas de colinealidad |
| B | Error estándar | Beta | Tolerancia | VIF |
| 1 | (Constante) | 4,046 | ,654 |  | 6,188 | ,000 |  |  |
| Percepción de efectividad de las medidas punitivas | ,422 | ,042 | ,357 | 9,939 | ,000 | ,960 | 1,042 |
| Sexo (Control Hombre) | -,550 | ,233 | -,083 | -2,357 | ,019 | ,994 | 1,006 |
| Posición política (1 izquierda - 10 derecha) | ,102 | ,063 | ,058 | 1,616 | ,107 | ,972 | 1,028 |
| Nota al gobierno: Disminuir la delincuencia | -,139 | ,079 | -,064 | -1,769 | ,077 | ,961 | 1,040 |
| a. Variable dependiente: Legitimidad de las detenciones ciudadanas |

## Errores independientes

Para utilizar una regresión lineal el error asociado a cada caso debe ser independiente del de los demás. En la mayoría de los datos que usamos como sociólogos este supuesto se cumple porque los casos de las muestras son elegidos de manera independiente. Este supuesto no necesariamente se cumple en las series temporales ni en las muestras donde de se seleccionan diadas (por ejemplo, parejas en una encuesta de vida sexual), donde puede ser puesto a prueba mediante un gráfico de tiempo/ubicación, aunque por lo general en esos casos es preferible optar por otras técnicas como modelos de autocorrelación.

De no cumplirse este supuesto habrá una incorrecta de los errores estándar.

## Errores que distribuyen de forma normal

Los errores de cada uno de los valores predichos de Y deben distribuirse de forma normal. Para determinar esto podemos utilizar un gráfico de distribución de residuos y un gráfico q-q de residuos, para lo cual en SPSS debemos, una vez en la pantalla de regresión lineal, ir a la pestaña gráficos y seleccionar la opción histograma y gráfico de prob. normal.





En el caso de nuestro ejemplo, podemos observar en ambos gráficos que los residuos se aproximan a una distribución normal. De no cumplirse este supuesto, habrá una inadecuada estimación del test de Wald.

## Homocedasticidad de los errores

Los residuos (diferencia entre el valor estimado y el real) deben tener una varianza constante a lo largo de los distintos valores predichos de Y. Para evaluar el cumplimiento de este supuesto debemos utilizar gráficos de residuos versus valores predichos.

Para obtener este gráfico en SPSS debemos ir, dentro de la pestaña de regresión lineal a la pestaña de gráficos y pedir un gráfico que incluya los residuos estandarizados en Y (ZRESID) y los valores pronosticados estandarizados en X (ZPRED).





En este caso no observamos que no existe una clara homocedasticidad de los errores ya que la distribución se “abre” hacia el final, mostrando que en los valores predichos más altos existen mayores residuos.

De no cumplirse este supuesto habrá una inadecuada estimación de los errores estándar de los coeficientes del modelo y del test de Wald.

## Ausencia de casos atípicos

Un modelo de RL se ve afectado por los casos atípicos ya que estos no siguen el patrón de relación con los demás casos, por lo cual el modelo los predice de peor manera, disminuyendo el ajuste del modelo, e influenciando sus resultados, especialmente para muestras pequeñas.

Para detectar casos atípicos, una vez dentro de la pestaña de regresión lineal debemos ir a guardar, y seleccionar residuos estandarizados (o tipificados), lo cual creara una nueva variable llamada “ZRE\_1” que nos indica los residuos estandarizados para cada caso. Se considerarán atípicos los casos con residuos estandarizados menores a -2 y mayores a 2.



Una vez hecho esto se pude volver a realizar la RL, excluyendo los casos atípicos de la base de datos. Para esto debemos usar el comando seleccionar casos, revisado en ayudantías anteriores y utilizar la opción “si cumple la condición” donde debemos introducir la siguiente formula:

ZRE\_1 >-2& ZRE\_1 < 2

La cual implica que solo se trabajará con los casos que tengan valores entre 2 y -2 en la variable ZRE\_1 de residuos estandarizados. Una vez hecho esto volvemos a ejecutar la regresión lineal. Para nuestro ejemplo esto implica una gran mejoría de la varianza explicada desde un 15,8% a un 23%, eliminando solo 23 casos.

|  |
| --- |
| **Resumen del modelo** |
| Modelo | R | R cuadrado | R cuadrado ajustado | Error estándar de la estimación |
| 1 | ,480a | ,230 | ,226 | 2,80701 |
| a. Predictores: (Constante), Posición política (1 izquierda - 10 derecha), Sexo (Control Hombre), Nota al gobierno: Disminuir la delincuencia , Percepción de efectividad de las medidas punitivas |

Es importante tener presente que, en caso de eliminar los casos atípicos, todas las estimaciones cambian, por lo cual es importante realizar esta prueba al principio, de modo de evaluar si se eliminarán o no los atípicos. De no eliminarse los atípicos, debe desactivarse la selección de casos (en lo que sigue de la guía se seguirá trabajando con casos atípicos).

## Independencia de X y $ε$

El error no se debe encontrar relacionado con las variables independientes. Este supuesto no se puede evaluar utilizando el residuo (estimador del error), ya que el método de mínimos cuadrados ordinarios asume que el residuo es independiente a X.

De no cumplirse este supuesto, si la correlación de x y el error es positiva se sobreestima los coeficientes beta y, si la correlación es negativa se subestima los coeficientes beta.

1. Este índice fue construido mediante la suma de las variables P46A, P46B y P46C de la encueta UDP 2015 (<http://encuesta.udp.cl/banco-de-datos/>), esto a fin de obtener una variable intervalar que nos permita usar la técnica, ya que las variables originales son ordinales. [↑](#footnote-ref-1)
2. Este índice fue construido mediante la suma de las variables P50A, P50B y P50C. [↑](#footnote-ref-2)
3. La variable sexo ha sido recodificada desde la variable P54, recodificando “hombres” como 0 y “mujeres” como 1. [↑](#footnote-ref-3)