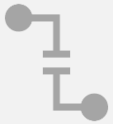
Práctica 5. Análisis de Regresión Lineal

Mínimos cuadrados ordinarios

Introducción

Frecuentemente estamos interesados en explorer los efectos de distintos factores sobre una variable de interés. La herramienta más elemental y usada que ocupamos los economistas para esto es un método estadístico llamado análisis de regresión, el cual se usa para estudiar el efecto de una o mas variable independientes sobre una variable dependiente.



Una variable dependiente es una variable de resultado que queremos explicar usando otras variables. Las variables independientes son las variables que usamos para “explicar” la variación en la variable dependiente (también son llamadas variables explicativas).

Variables dependientes e independientes

- La distinción entre variables dependientes e independientes se basa en una suposición clave del análisis de regresión: el supuesto es que las variables independientes son **exógenas**, lo que significa que no se ven afectados por la variable dependiente, ni hay ninguna variable fuera de ese modelo que afecta tanto a la variable dependiente como a las variables independientes.

Relación lineal simple

- Comenzamos con el caso simple de una relación lineal entre una variable dependiente y una sola variable independiente. Esta relación puede describirse con la siguiente ecuación:

$$y = \beta_0 + \beta_1 x + \varepsilon$$

- Donde y es la variable dependiente, x es la variable independiente, β β_0 es la constante o el intercepto de y, β_1 es la pendiente o el coeficiente de x y ε es el término de error

Relación lineal simple

- El término de error, ε, refleja el hecho de que la relación entre y y x no es exacta, sino que está sujeto a algún error. Tenga en cuenta que β cero y β uno son parámetros que no pueden ser observados directamente, solo podemos estimarlos usando los valores de y y x. Igualmente, el término de error, ε, no se puede observar directamente.

- El valor predicho de y, escrito como $\hat{y}$, se define de la siguiente manera:

$$Predictor\ de\ y = \hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

Donde $\hat{\beta}_i$ es el valor estimado del verdadero parámetro β_i. Como puede ver, el "gorro" indica el estimado del parámetro poblacional basado en los datos de la muestra

Residual

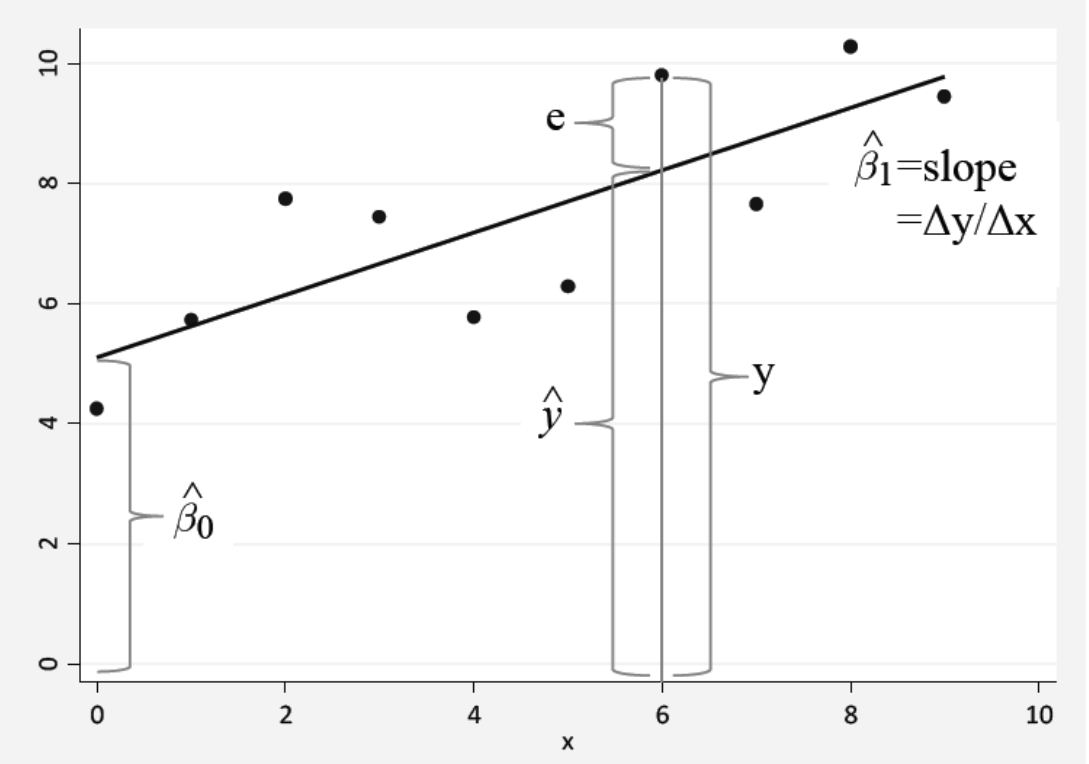
- El residual es la diferencia entre el valor verdadero y el valor predicho:

$$residual = y - \hat{y} = e$$

- No confunda ε y e: ε es el término de error no observado en la "verdadera" relación entre y y x, mientras que e es la diferencia observada entre y y su valor predicho, $\hat{y}$, el último basado en la relación estimada entre y y x.

Interpretación gráfica del análisis de regresión simple

- La relación entre estos conceptos es mostrada de una forma simplificada en la siguiente figura:



Método de Mínimos Cuadrados Ordinarios

- El método de mínimos cuadrados ordinaries minimiza la suma de los residuales al cuadrado para todas las observaciones. El cálculo de los coeficientes estimados y sus estadísticos relaciondos usa algebra matricial. Los lectores interesados pueden encontrar más información en Woolridge (2016), Greene (2018) o en cualquier libro de econometría de su preferencia.

Ejemplo

- La base de datos contiene información sobre una muestra de coches con sus características tales como el precio y su millaje.
- Al hacer la regresión de precio sobre las millas recorridas en cualquier paquete estadístico, debemos obtener los siguientes estadísticos. La presente table de salida corresponde a *stata*:

```
. regress price mpg
```

Source	SS	df	MS		
Model	139449474	1	139449474	Number of obs	= 74
Residual	495615923	72	6883554.48	F(1, 72)	= 20.26
				Prob > F	= 0.0000
				R-squared	= 0.2196
				Adj R-squared	= 0.2087
Total	635065396	73	8699525.97	Root MSE	= 2623.7

price	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
mpg	-238.8943	53.07669	-4.50	0.000	-344.7008	-133.0879
_cons	11253.06	1170.813	9.61	0.000	8919.088	13587.03

Elementos de la tabla de salida

- El estadístico F es una prueba de la hipótesis nula que estipula que todos los coeficientes (excluyendo la constante) son iguales a cero.
- La “Prob>F” da la probabilidad de que el estadístico F pueda ser generado al azar si la hipótesis nula fuera verdadera
- R-squared se refiere a R^2, el coeficiente de determinación de los valores observados de y ($\hat{y}$). En un modelo de regression lineal con una constante, R^2puede también interpretarse como la proporción de la varianza en y que uede ser explicada por el modelo.

Elementos de la tabla de salida

- "Adj R-squared" se refiere a la R^2 ajustada. Una limitación de la R^2 es que, cuando se añade una variable independiente al modelo, R^2 siempre se incrementará, aún si la nueva variable carece de poder predictivo sobre la variable dependiente. La R^2 ajustada es justamente ajustada por el número de variables independientes, de tal forma que sólo incrementará su valor si la nueva variable aumenta el poder explicativo del modelo.

Elementos de la tabla de salida

- Los coeficientes y el error estándar de los coeficientes se explican por si sólos. La t y la $P > |t|$ de los coeficietes es el estadístico de la prueba t de estudent que prueba la hipótesis nula que ese coeficiente es igual a cero y la probabilidad de que el estadístico t pueda ser de ese valor dado que la hipótesis nula es cierta, conocido como *p-value*. La convención es tomar como significativo un coeficiente si su p-value asociado es menor a .05

Ecuación a partir de la tabla de salida

- Mirando la tabla de salida, observamos que los coeficientes son significativos. Por tanto, las variables y coeficientes pueden ser reacomodados de tal forma que formen la ecuación que mejor ajuste los datos de la siguiente manera:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

Precio predicho: 11253+(-238.89)*millaje

- De esta forma, el modelo nos dice que el millaje es una variable que reduce el precio de un coche, lo que es un resultado esperado.

Análisis de regresión Multivariable

- El análisis de regresión simple se refiere al caso en donde sólo hay una variable independiente con una constante. Sin embargo, el análisis de regresión puede aplicarse con múltiples variables independientes además de la constante. E este caso, el modelo asume que los datos siguen el siguiente patron:

$$y = \beta_0 + \sum_{i=1}^{k-1} \beta_i x_i + \varepsilon$$

Donde k es el número de variables independientes, incluyendo la constante, $\beta_i$es uno de los k-ésimos coeficientes, $x_i$es una de las k-1 variables independientes y ε es el término de error.

Realización de la práctica

A partir de la base de datos, lo que usted tiene que hacer es la regresión de la variable dependiente precio (Price), sobre las siguientes variables explicativas:

Millaje (mpg)

Peso (weight)

Largo (length)

Estatús del coche (rep78)

- Tenga en cuenta que la variable rep 78 es una variable categórica con cuatro distintos valores. Tendrá que tratar esta variable antes de hacer la regression.
- Por supuesto, puede usar su paquete estadístico preferido.

La práctica debe incluir:

- Sús líneas de código
- Todos los estadísticos mostrados en esta presentación
- La ecuación formada a partir de los coeficientes estimados y significativos.
- Su ***interpretación*** de estos estadísticos. No se salte esta instrucción, forma el 50% de su calificación de la práctica. No es necesario una explicación técnica, con pocas líneas y sencillas palabras explique qué significan o que le dicen a usted el valor de estos estadísticos.