

Pauta Auxiliar 3 - Semestre Otoño 2016

29 de Marzo, 2016

Problema 1: Multicolinealidad

Considere los siguientes modelos estimados en Stata con la base de datos "auto"¹. En ambos modelos se estima el precio de un automóvil según características del auto: cantidad de millaje, tamaño del motor y si es extranjero o no:

$$\text{Precio}_i = 11253,06 - 238,9 \cdot \text{Millaje}_i + \varepsilon_i$$

$$\text{Precio}_i = 10033,08 - 261,9 \cdot \text{Millaje}_i + 83,65 \cdot \text{Motor}_i + 1887,5 \cdot \text{Extranjero}_i + \nu_i$$

Usando la fórmula para la varianza del estimador β_{MCO} , donde SST_j es la variabilidad de la variable x_j , calculada como $SST_j = \sum_i (x_{ij} - \bar{x}_j)^2$ y R_j^2 es la bondad de ajuste de la regresión de x_j sobre el resto de las demás variables explicativas:

$$\text{Var}(\hat{\beta}_{MCO}) = \frac{\sigma^2}{SST_j(1 - R_j^2)}$$

Considere σ^2 constante y conocido; y además la tabla de valores obtenida del mismo modelo:

	R_j^2	SST_j
Millaje_i	0.378	0.494
Motor_i	0.358	7.6
Extranjero_i	0.18	0.09

- Calcule las varianzas del estimador $\hat{\beta}_{\text{Millaje}}$ para ambos modelos.
- Compare y concluya qué sucede cuando se agregan regresores a un modelo y por qué hay que tener precaución.

¹Disponible al usar el comando "sysuse auto"

Respuesta

- a) En realidad es un cálculo simple, pues solamente se debe utilizar la fórmula propuesta y los datos en la tabla. Lo importante es notar que para el modelo que tiene más de una variable, se deben utilizar los valores R_j^2 :

$$\begin{aligned} Var(\hat{\beta}_{Milla je}^1) &= \frac{\sigma^2}{SST_{Milla je}} = \frac{\sigma^2}{0,494} \\ Var(\hat{\beta}_{Milla je}^2) &= \frac{\sigma^2}{SST_{Milla je}(1 - R_{Milla je}^2)} = \frac{\sigma^2}{0,494 \cdot (1 - 0,378)} = \frac{\sigma^2}{0,307} \end{aligned}$$

- b) Notar que la varianza del estimador del segundo modelo es mayor que la varianza del estimador del primer modelo. Aun cuando en ambos casos es el estimador MCO, cuando se estiman modelos con más variables independientes, la varianza aumenta. Esto es producto de la multicolinealidad. Pensar que SIEMPRE existirá algún nivel de correlación entre las variables independientes, pues es realmente difícil que sean exactamente independientes entre sí. Esto significa que siempre que se agreguen variables a un modelo, se afectará el estimador producto de la multicolinealidad. Mientras estos niveles de multicolinealidad se mantengan bajos, no hay problema. Para testear los niveles de multicolinealidad se pueden usar varios métodos, entre ellos el del Factor de Inflación de Varianza VIF.

Problema 2: Coeficientes Dummy

Se tiene el siguiente modelo:

$$PASE = \beta_0 + \beta_1 CHI + \beta_2 ARG + u$$

Considerando que el estudio se hará sólo para jugadores provenientes de Chile, Argentina y Brazil.

- a) ¿Cuánto es la diferencia esperada del valor del pase de un jugador argentino y uno chileno? ¿Y de ambos con uno brasileño?

Respuesta

Se deben ver los valores esperados condicionados a cada uno de las nacionalidades:

$$\mathbb{E}(PASE \mid CHI = 0, ARG = 1) = \beta_0 + \beta_2$$

$$\mathbb{E}(PASE \mid CHI = 1, ARG = 0) = \beta_0 + \beta_1$$

Ahora teniendo ambos valores, se ve la diferencia $\beta_0 + \beta_2 - \beta_0 - \beta_1 = \beta_2 - \beta_1$

Considerando que la variable base de las categorías dummies es la nacionalidad brasilera, se debe tener en cuenta que el valor esperado del pase de un jugador brasileño es

$$\mathbb{E}(PASE \mid CHI = 0, ARG = 0) = \beta_0$$

Corresponde por tanto al intercepto o media, que en este caso se considera la referencia. Por lo tanto las diferencias:

Entre un jugador argentino y uno brasileño: $\beta_0 + \beta_2 - \beta_0 = \beta_2$

Entre un jugador chileno y uno brasileño: $\beta_0 + \beta_1 - \beta_0 = \beta_1$

Suponiendo ahora que el modelo a estudiar considera el número de goles que ha marcado cada jugador en su historia:

$$PASE = \beta_0 + \beta_1 CHI + \beta_2 ARG + \beta_3 G + \beta_4 CHI \cdot G + \beta_5 ARG \cdot G + u$$

- b) ¿Cómo afecta la cantidad de goles a cada nacionalidad de jugador?

Respuesta

Estudiamos cómo afectan los goles condicionando en la nacionalidad del jugador. Queremos saber cuál es el retorno.^{en} el pase del jugador por cada gol convertido.

$$\frac{\partial \mathbb{E}(PASE \mid CHI = 1, ARG = 0)}{\partial G} = \beta_4 + \beta_3$$

$$\frac{\partial \mathbb{E}(PASE \mid CHI = 0, ARG = 1)}{\partial G} = \beta_5 + \beta_3$$

$$\frac{\partial \mathbb{E}(PASE \mid CHI = 0, ARG = 0)}{\partial G} = \beta_3$$

Notar que a simple vista se puede confundir la interpretación del coeficiente G en el modelo.

- c) ¿Qué representan entonces los coeficientes β_4 y β_5 ?

Respuesta

Como se vio anteriormente, los coeficientes pueden ser representados por diferencias de retornos esperados:

La diferencia de retornos de los goles para jugadores chilenos y brasileños:

$$\beta_4 = \frac{\partial \mathbb{E}(PASE \mid CHI = 1, ARG = 0)}{\partial G} - \frac{\partial \mathbb{E}(PASE \mid CHI = 0, ARG = 0)}{\partial G}$$

La diferencia de retornos de los goles para jugadores argentinos y brasileños:

$$\beta_5 = \frac{\partial \mathbb{E}(PASE \mid CHI = 0, ARG = 1)}{\partial G} - \frac{\partial \mathbb{E}(PASE \mid CHI = 0, ARG = 0)}{\partial G}$$

- d) Comente sobre la inclusión de la variable BRA en el modelo.

Respuesta

Si se añadiera la variable BRA en el modelo, quedaría como referencia un jugador que no fuera ni chileno, ni argentino ni brasileño, cosa que en el estudio no fue medido. Esto llevaría a una mala interpretación de los resultados.

Por otra parte, se podría modificar el modelo hacia uno que no incluyera un intercepto o valor medio, pero eso dificulta la correcta implementación en software sobre regresiones lineales.

- e) Explique con sus palabras que una variable sea estadísticamente significativa.

Respuesta

Cuando conocemos (o asumimos) la distribución de probabilidades de una variable aleatoria, podemos conocer sobre su comportamiento y caracterizarla. Como explica la matemática detrás de la regresión MCO, se asume que los estimadores lineales se comportan

como una distribución normal a medida que tenemos una muestra suficientemente grande. Lo que nos dicen los test de hipótesis, es que **existe un patrón o relación que se escapa del comportamiento aleatorio de la distribución** que estamos estudiando, es decir, que los datos muestran que no es azar que exista la relación que estamos analizando.

Por poner un ejemplo banal y simple, imaginemos que tenemos dos dados (dos variables aleatorias que toman los valores 1 a 6 con la misma probabilidad). Si lanzamos los dados al mismo tiempo 100 veces, uno esperaría, por azar (probabilidades conocidas) que salieran un sexto de las veces cada número independientemente de cada dado. Si analizáramos los datos, y vieramos que en muchísimas ocasiones salieron juntos ambos números 6, sospecharíamos de los dados, y diríamos que hay los dados están significativamente cargados; o bien, que existe una correlación que es significativa, por que se escapa del comportamiento normal que esperamos por la distribución de probabilidades. Lo mismo ocurre con una distribución normal, en una variable aleatoria muestral que analizamos con los test de hipótesis.