

## Pauta Auxiliar 3.5 - Semestre Otoño 2016

5 de Abril, 2016

### Problema 1: Multicolinealidad

Considere los siguientes modelos estimados en Stata con la base de datos `auto`<sup>1</sup>. 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  $\beta_{MCO}$ , donde  $SST_j$  es la variabilidad de la variable  $x_j$ , calculada como  $SST_j = \sum_i (x_{ij} - \hat{x}_j)^2$  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  $\sigma^2$  constante y conocido; y además la tabla de valores obtenida del mismo modelo:

	$R_j^2$	$SST_j$
$\text{Millaje}_i$	0.378	0.494
$\text{Motor}_i$	0.358	7.6
$\text{Extranjero}_i$	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.

---

<sup>1</sup>Disponible 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  $R_j^2$ :

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

- 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: Test de Hipótesis

En base a la estimación del estudio de la natalidad de ciertas zonas geográficas de los EE.UU en el censo de 2003, responda las siguientes preguntas.

natalidad	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
edad_prom	-109.0957	13.52452	-8.07	0.000	-136.3526	-81.83886
edad_prom2	1.635208	.2290536	7.14			
zona						
2	15.00284	4.252068			6.433365	23.57233
3	7.366435	3.953336	1.86		-1.6009898	15.33386
4	21.39679	4.650602	4.60	0.000	12.02412	30.76946

## Individual

- a) Explique con sus palabras qué significa una variable significativa.
- b) Con los datos de la tabla, concluya el test de significancia individual para cada variable.

## Múltiples restricciones lineales

c) Plantee la hipótesis nula para las siguientes afirmaciones:

1. ¿Es el efecto de estar en la zona 2 equivalente a estar en la zona geográfica 3?
2. ¿Es el efecto de estar en la zona 1 un 25 % menor que estar en la zona 2?

## Global

En base a los resultados de la tabla de la estimación del modelo

$$natalidad_i = \beta_0 + \sum_{j=2}^4 \beta_j zona_j + \varepsilon_i$$

Source	SS	df	MS	Number of obs =	50
Model	19545.4359	3	6515.14531		
Residual	22651.3841	46	492.421393	R-squared =	0.4632
				Adj R-squared =	0.4282
Total	42196.82	49	861.159592	Root MSE =	22.191

Responda:

d) ¿Son las zonas geográficas conjuntamente significativas?

## Bonus - Otros tests

De los siguientes test realizados en Stata, identifique la hipótesis nula, la distribución del estadístico y concluya el resultado del test.

```
Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
Ho: Constant variance
Variables: fitted values of natalidad

chi2(1)      =    18.78
Prob > chi2  =    0.0000
```

```
Ramsey RESET test using powers of the fitted values of natalidad
Ho: model has no omitted variables

F(3, 41) =    0.49
Prob > F =    0.6934
```

```

      mean(diff) = mean(edadpro - edadpro2)          t = -60.9186
Ho: mean(diff) = 0                                degrees of freedom = 49

Ha: mean(diff) < 0          Ha: mean(diff) != 0          Ha: mean(diff) > 0
Pr(T < t) = 0.0000          Pr(|T| > |t|) = 0.0000          Pr(T > t) = 1.0000

```

```
. estat dwatson
```

```
Durbin-Watson d-statistic( 6, 50) = 2.002082
```

## Respuesta

a) Una variable significativa indica que los patrones de los datos, la evidencia, sugiere que el comportamiento de la variable se escapa del que sería su comportamiento al azar, de acuerdo a su distribución. En otras palabras, es tan poco probable que la variable se comporte así, que puede ser adjudicado a la evidencia de los datos. Tiene un comportamiento anormal.

b) Al notar que hay valores que están tapados, se deben utilizar los otros indicadores.

1 Edad\_prom se puede ver que, tanto el estadístico  $t$  está por debajo de -1,96 (o de -2 al ojo), el p-valor es menor a 0,05; y además el intervalo NO contiene al cero. Por tanto se puede rechazar la hipótesis nula de  $\hat{\beta}_1 = 0$ .

2 edad\_prom2 se puede ver que tiene el estadístico en un nivel mayor al calor crítico de esa distribución al 95 %, por lo tanto se llega a la misma conclusión.

3 zona.2 se puede ver que su intervalo de confianza NO contiene al cero, por lo tanto es una variable significativa.

4 zona.3 al contrario, el valor del estadístico  $t$  es más bajo que el valor crítico, por lo que no se puede rechazar la hipótesis nula. Esto se puede ver también en que el intervalo de confianza sí contiene al cero.

5 zona.4 de la misma maner, se rechaza la hipótesis nula y sería una variable significativa.

6 falta información para la constante.

c) Las hipótesis quedarían planteadas de la siguiente forma para poder ser testeadas:

1  $\beta_{zona,2} - \beta_{zona,3} = 0$

2  $\beta_{zona,1} - 0,75 \cdot \beta_{zona,2} = 0$

Lo último se entiende al visualizar que  $\frac{\beta_{zona,1}}{\beta_{zona,2}} = 0,75$  es la hipótesis que queremos testear.

Al testearlo, el resultado es el siguiente, queda planteada la conclusión,

( 1)	2.zona - 3.zona = 0		
	F( 1, 44) =	5.06	
	Prob > F =	0.0295	

  

( 1)	1b.zona - .75*2.zona = 0		
	F( 1, 44) =	12.45	
	Prob > F =	0.0010	

d) De acuerdo a la tabla, se puede calcular el estadístico:

$$\frac{R^2}{(1 - R^2)} \cdot \left(\frac{N - K}{K}\right) = \frac{0,463}{(1 - 0,463)} \cdot \left(\frac{49 - 3}{3}\right) = 13,23$$

Al mirar una tabla del estadístico, en base a esos grados de libertad, se puede ver que el valor crítico es  $F(3, 46) = 3,859$ . Dado que el valor del estadístico es mayor al valor crítico, podemos rechazar con seguridad la hipótesis nula. Ya que la hipótesis nula es que las zonas geográficas no son conjuntamente significativas (recordar que la hipótesis nula es  $\beta_i = 0 \forall i$  al mismo tiempo), podemos rechazar dicha hipótesis e inclinarnos por decir que la evidencia sugiere que sí son significativas en conjunto las variables geográficas en el modelo estimado. Con esta determinación podremos tomar una decisión respecto a las políticas públicas que realicemos al país, y así salvar al mundo. FIN.