

Estadística para la Economía y la Gestión

IN 3401

Clase 5

21 de octubre de 2009

1 Variables Dummies o cualitativas

2 Omisión de Variables Relevantes

- Impacto sobre el Insesgamiento
- Impacto sobre la Varianza
- Ejemplo

3 Inclusión de Variables Irrelevantes

- Impacto sobre el Insesgamiento
- Impacto sobre la Varianza
- Ejemplo

En el análisis de regresión, la variable dependiente está influida frecuentemente no solo por variables continuas como son el ingreso, producción, precios, etc..., sino también por variables que son esencialmente cualitativas, estos son regresores binarios, es decir, variables que sólo toman el valor 0 o 1.

Dichas variables son llamadas variables dummies, variables dicotómicas o variables ficticias. Muchas veces el regresor es binario porque así fue recogido en la encuesta. Sin embargo, en otros casos el regresor binario ha sido construido a partir de otras variables de los datos.

Algunos ejemplos de variable dummies son: género, raza, religión, nacionalidad, región geográfica, etc.

Por ejemplo, se ha encontrado que manteniendo todos los demás factores constantes, las trabajadoras mujeres ganan menos que sus colegas hombres, y que las personas de color ganan menos que las blancas.

Este patrón puede resultar de discriminación sexual o racial, pero cualquiera sea la razón, las variables cualitativas tales como género o raza sí influyen sobre la variable dependiente.

Por ejemplo, consideremos la siguiente variable dummy para género (mujer/hombre) del individuo. Entonces la variable dummy consistirá en un vector $(n \times 1)$ con elementos 0 o 1 según corresponda. Es decir:

$$d_{1i} = \begin{cases} 1 & \text{Mujer} \\ 0 & \text{Hombre} \end{cases}$$

Pensemos en una ecuación simple de salarios $E(\text{Salario}(W)/\text{Género})$, la cual implica el siguiente modelo:

$$W_i = \beta_0 + \beta_1 d_{1i} + \epsilon_i$$

entonces, dada la especificación escogida para la dummy, tenemos que:

$$\begin{aligned} \beta_0 &= E(W|\text{Hombre}) \\ \beta_0 + \beta_1 &= E(W|\text{Mujer}) \end{aligned}$$

Alternativamente, podríamos haber definido la dummy de la siguiente forma:

$$d_{2i} = \begin{cases} 0 & \text{Mujer} \\ 1 & \text{Hombre} \end{cases}$$

y el modelo

$$W_i = \beta_0 + \beta_1 d_{2i} + \epsilon_i$$

entonces, en esta segunda especificación escogida para la dummy, tenemos que:

$$\begin{aligned} \beta_0 &= E(W|Mujer) \\ \beta_0 + \beta_1 &= E(W|Hombre) \end{aligned}$$

Una tercera forma de definir el modelo sería incluyendo ambas dummies:

$$W_i = \beta_1 d_{1i} + \beta_2 d_{2i} + \epsilon_i$$

con el cual tendríamos que los retornos a ambos géneros serían:

$$\beta_1 = E(W|Mu\tilde{e}r)$$

$$\beta_2 = E(W|Hombre)$$

Los tres modelos anteriores son equivalentes. Note que en el tercer modelo no incluimos término constante ya que ello haría que la matriz X fuese singular y por lo tanto, no invertible (Trampa de las Dummies).

Ejemplo 1

Contamos con datos de ingreso proveniente de la ocupación principal para el año 2000, de acuerdo a zona geográfica de Chile: Norte (de la primera a la cuarta región), Centro (quinta región, sexta región y región metropolitana) y Sur (de la séptima a la duodécima región).

Zona Geográfica	Salario Promedio
Norte	270.154
Centro	296.857
Sur	240.238

Suponga que deseamos averiguar si el salario promedio difiere entre las distintas zonas geográficas.

Consideremos el siguiente modelo de regresión para el salario (y_i):

$$y_i = \beta_0 + \beta_1 D_{1i} + \beta_2 D_{2i} + \epsilon_i$$

D_{1i} : es una variable dummy que toma valor 1 si la persona i vive en el norte y cero sino.

D_{2i} : es una variable dummy que toma valor 1 si la persona i vive en el sur y cero sino.

De esta forma, el salario promedio de los individuos que viven en el norte es:

$$E(Y_i | D_{1i} = 1, D_{2i} = 0) = \beta_0 + \beta_1$$

de igual forma el salario promedio de los individuos que viven en el sur es:

$$E(Y_i | D_{1i} = 0, D_{2i} = 1) = \beta_0 + \beta_2$$

y por último, el salario promedio de los individuos que viven en el centro es:

$$E(Y_i | D_{1i} = 0, D_{2i} = 0) = \beta_0$$

Necesitamos ver si estas diferencias son estadísticamente significativas. El modelo estimado es:

Source	SS	df	MS			
Model	4.9318e+13	2	2.4659e+13	Number of obs = 76935		
Residual	2.3751e+16	76932	3.0872e+11	F(2, 76932) = 79.88		
Total	2.3800e+16	76934	3.0935e+11	Prob > F = 0.0000		
				R-squared = 0.0021		
				Adj R-squared = 0.0020		
				Root MSE = 5.6e+05		
yopraj	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
d1	-26703.8	6710.006	-3.98	0.000	-39855.38	-13552.23
d2	-56618.94	4500.677	-12.58	0.000	-65440.25	-47797.64
_cons	296857.8	2587.352	114.73	0.000	291786.6	301929

La diferencia en los salarios promedios entre regiones es estadísticamente significativa.

Es importante tener claro que las variables dicotómicas simplemente señalaran las diferencias, si es que estas existen, pero no sugieren razones por las cuales estas se presentan.

Los modelos econométricos generalmente son más amplios e introducen tanto variables explicativas continuas como dicotómicas. Por ejemplo, es razonable suponer que, además del género, existen otros factores que explican el salario (educación y experiencia, entre otros)

Especifiquemos nuevamente el modelo como $E(W|\text{Educación, Género})$:

$$W_i = \beta_0 + \beta_1 d_{2i} + \beta_2 E_i + \epsilon_i$$

Dicho modelo presenta un efecto intercepto para el género, es decir, hombres y mujeres poseen diferente intercepto, pero igual pendiente (β_2) en educación (retorno a la educación).

Ahora, si quisiéramos especificar un modelo en que además las pendientes varíen con el género (retornos a la educación diferenciados), tendríamos el siguiente modelo:

$$W_i = \beta_0 + \beta_1 d_{2i} + \beta_2 E_i + \beta_3 d_{2i} \cdot E_i + \epsilon_i$$

donde:

$$E(\text{Salario}(W)|E, \text{Hombre}) = \beta_0 + \beta_1 + (\beta_2 + \beta_3)E$$

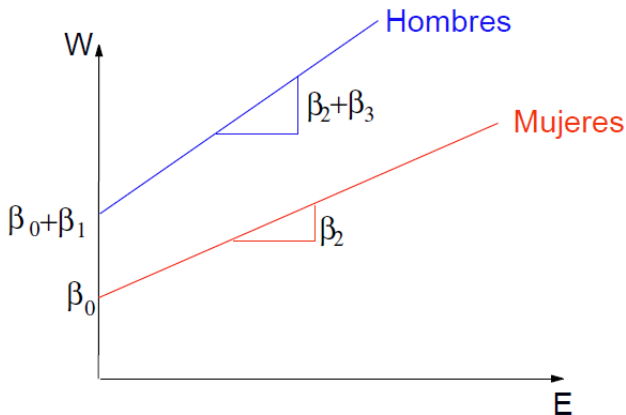
$$E(\text{Salario}(W)|E, \text{Mujer}) = \beta_0 + \beta_2 E$$

Por lo tanto el efecto marginal de la educación es:

$$\text{Hombres: } \frac{\partial E(\text{Salario}(W)|E, \text{Hombre})}{\partial E} = \beta_2 + \beta_3$$

$$\text{Mujeres: } \frac{\partial E(\text{Salario}(W)|E, \text{Mujer})}{\partial E} = \beta_2$$

Gráficamente...



Salario y Educación, diferencia de intercepto y pendiente entre hombres y mujeres

Ejemplo 2

A continuación veremos la aplicación de la llamada "Ecuación de Mincer" para estimar el retorno a la educación. Los datos corresponden a un grupo de jóvenes chilenos egresados de la educación media técnica, los que fueron entrevistados en 1997. La primera figura muestra la estimación de la ecuación de Mincer en su versión original (1974):

$$\ln(W_i) = \alpha + \beta \text{Educacion}_i + \epsilon_i$$

. regress logw edu_historia if trabajahoy==1						
Source	SS	df	MS			
Model	2.31709498	1	2.31709498	Number of obs =	626	
Residual	152.696431	624	.244705819	F(1, 624) =	9.47	
				Prob > F =	0.0022	
				R-squared =	0.0149	
				Adj R-squared =	0.0134	
Total	155.013526	625	.248021642	Root MSE =	.49468	
logw	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
edu_historia	.0697799	.0226767	3.08	0.002	.0252479	.1143119
_cons	11.04943	.2807663	39.35	0.000	10.49807	11.60079

Consideremos ahora una versión más completa del modelo en que incluimos la experiencia y una dummy que toma el valor 1 si el individuo es una mujer:

```
. regress logw exp edu_historia dmujer if trabajahoy==1
```

Source	SS	df	MS			
Model	20.7029712	3	6.9009904	Number of obs =	626	
Residual	134.310555	622	.215933368	F(3, 622) =	31.96	
				Prob > F =	0.0000	
				R-squared =	0.1336	
				Adj R-squared =	0.1294	
				Root MSE =	.46469	
Total	155.013526	625	.248021642			

logw	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
exp	.0980998	.0184241	5.32	0.000	.0619188	.1342808
edu_historia	.1513598	.0269237	5.62	0.000	.0984874	.2042321
dmujer	-.2180018	.0390985	-5.58	0.000	-.2947828	-.1412208
_cons	9.716256	.391983	24.79	0.000	8.946485	10.48603

Note que el retorno a la educación sigue siendo positivo, mientras que la dummy para mujer es negativa (Qué significa que el parámetro sea negativo?).

Veamos a continuación, la misma especificación, sólo que esta vez la dummy se define como 1 si el individuo es hombre:

```
. regress logw exp edu_historia dhombre if trabajahoy==1
```

Source	SS	df	MS			
Model	20.7029712	3	6.9009904	Number of obs = 626		
Residual	134.310555	622	.215933368	F(3, 622) = 31.96		
				Prob > F = 0.0000		
				R-squared = 0.1336		
				Adj R-squared = 0.1294		
				Root MSE = .46469		
Total	155.013526	625	.248021642			

logw	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
exp	.0980998	.0184241	5.32	0.000	.0619188	.1342808
edu_historia	.1513598	.0269237	5.62	0.000	.0984874	.2042321
dhombre	.2180018	.0390985	5.58	0.000	.1412208	.2947828
_cons	9.498254	.3825447	24.83	0.000	8.747018	10.24949

Cómo es el parámetro de la dummy para el hombre comparado con el de la mujer? Qué pasa con la estimación del resto de los parámetros?.

Omisión Variable Relevante

Considere el siguiente modelo poblacional (expresado en desvíos con respecto a la media):

$$Y = X_1\beta_1 + X_2\beta_2 + \epsilon$$

Suponga ahora que el investigador se equivoca y estima el siguiente modelo:

$$Y = X_1\beta_1 + \epsilon$$

Estimando el modelo *incorrecto* obtenemos:

$$\begin{aligned}\hat{\beta}_1 &= (X_1'X_1)^{-1}X_1'Y \\ &= \beta_1 + (X_1'X_1)^{-1}X_1'X_2\beta_2 + (X_1'X_1)^{-1}X_1'\epsilon\end{aligned}$$

por lo cual:

$$\begin{aligned}E(\hat{\beta}_1) &= \beta_1 + (X_1'X_1)^{-1}X_1'X_2\beta_2 \\ &= \beta_1 + Z\beta_2\end{aligned}$$

Ello implica que por lo general, la omisión de variables relevantes (que pertenecen al modelo poblacional), causará que los parámetros estimados sea sesgados.

Ello no sucederá, sólo en el caso que $Z = 0$ (es decir que X_1 y X_2 sea ortogonales) o si $\beta_2 = 0$ (aunque dicho caso es contradictorio, dado que implicaría que la variable no pertenece al modelo poblacional).

La dirección del sesgo es difícil de obtener, sin embargo, el análisis se simplifica si pensamos en β_1 y β_2 como escalares. En dicho caso:

$$E(\hat{\beta}_1) = \beta_1 + \frac{\text{Cov}(X_1, X_2)}{V(X_1)}\beta_2$$

La dirección del sesgo depende de como covarien las variables incluidas con respecto a las excluidas y del signo del parámetro omitido.

Estimando el modelo *incorrecto*, el estimador de la varianza será:

$$V(\hat{\beta}_1|X_1) = \sigma^2(X_1'X_1)^{-1}$$

mientras que si hubiéramos estimado el modelo correcto, se puede demostrar que la varianza del estimador insesgado de β_1 ($\hat{\beta}_1^*$) correspondería a:

$$V(\hat{\beta}_1^*|X_1, X_2) = \sigma^2(X_1'M_2X_1)^{-1}$$

donde $M_2 = I - X_2'(X_2'X_2)^{-1}X_2'$. Luego, comparamos las inversas de ambas matrices:

$$(V(\hat{\beta}_1|X_1))^{-1} - (V(\hat{\beta}_1^*|X_1, X_2))^{-1} = \sigma^{-2}(X_1'X_2(X_2'X_2)^{-1}X_2'X_1)$$

tal que se puede demostrar que dicha matriz es definida positiva.

Por lo tanto, el omitir variables relevantes implica que los parámetros estimados serán sesgados y que sus varianzas serán menores.

Suponga que un investigador quiere estimar el retorno a la educación y que el modelo verdadero (obviamente es un caso ilustrativo) está dado por:

$$W_i = \beta_1 E_i + \beta_2 EXP_i + \epsilon_i$$

Donde W_i corresponde al logaritmo del salario del individuo i , E_i corresponde a los años de educación del individuo i , EXP_i corresponde a los años de experiencia laboral del individuo i y ϵ_i corresponde a un término de error bien comportado.

Sin embargo este investigador utiliza el siguiente modelo para su estimación.

$$W_i = \beta_1 E_i + \epsilon_i$$

Los resultados del modelo verdadero son:

```
reg lnw esc exp
```

Source	SS	df	MS			
Model	11944.2608	2	5972.13038			
Residual	31137.3188	63559	.489896298			
Total	43081.5796	63561	.677798958			

Number of obs =	63562
F(2, 63559) =	12190.60
Prob > F =	0.0000
R-squared =	0.2772
Adj R-squared =	0.2772
Root MSE =	.69993

lnw	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
esc	.1222713	.0007833	156.10	0.000	.1207361 .1238065
exp	.0185953	.0002238	83.10	0.000	.0181567 .0190339
_cons	10.21893	.0114195	894.87	0.000	10.19655 10.24132

Los resultados del modelo estimado son:

```
reg lnw esc
```

Source	SS	df	MS			
Model	8561.14296	1	8561.14296			
Residual	34520.4366	63560	.543115742			
Total	43081.5796	63561	.677798958			

Number of obs =	63562
F(1, 63560) =	15763.02
Prob > F =	0.0000
R-squared =	0.1987
Adj R-squared =	0.1987
Root MSE =	.73696

lnw	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
esc	.0863413	.0006877	125.55	0.000	.0849934 .0876892
_cons	10.98386	.0071161	1543.52	0.000	10.96991 10.99781

Podemos ver el parámetro que acompaña a la variable años de educación es menor en el modelo estimado que en el modelo verdadero. Esta dirección del sesgo se puede explicar por el signo del parámetro que acompaña a la variable experiencia en el modelo verdadero y a la relación existente entre educación y experiencia en el mercado laboral.

Inclusión Variable Irrelevante

Considere ahora el siguiente modelo poblacional:

$$Y = X_1\beta_1 + \epsilon$$

Suponga ahora que el investigador se equivoca y estima el siguiente modelo:

$$Y = X_1\beta_1 + X_2\beta_2 + \epsilon$$

Estimando el modelo *incorrecto* obtenemos:

$$\begin{aligned}\hat{\beta}_1 &= (X_1' M_2 X_1)^{-1} X_1' M_2 Y \\ &= \beta_1 + (X_1' M_2 X_1)^{-1} X_1' M_2 \epsilon\end{aligned}$$

donde M_2 se define igual que el la sección anterior. Entonces:

$$E(\hat{\beta}_1) = \beta_1$$

Con el mismo razonamiento, se puede demostrar que:

$$\begin{aligned} E(\hat{\sigma}^2) &= E\left(\frac{\hat{\epsilon}'\hat{\epsilon}}{n - k_1 - k_2}\right) \\ &= \sigma^2 \end{aligned}$$

es decir, la inclusión de variable irrelevantes no causa sesgo en los parámetros estimados, ni en la varianza de los errores estimados. Bajo dichos resultados, pareciera que es mejor poner muchos regresores en nuestro modelo. Sin embargo, nos falta estudiar que sucede con la varianza de los parámetros estimados.

Recordemos que:

$$\hat{\beta}_1 = \beta_1 + (X_1' M_2 X_1)^{-1} X_1' M_2 \epsilon$$

con lo cual, la varianza estimada:

$$V(\hat{\beta}_1 | X_1, X_2) = \sigma^2 (X_1' M_2 X_1)^{-1}$$

mientras que la varianza verdadera es:

$$V(\hat{\beta}_1^* | X_1) = \sigma^2 (X_1' X_1)^{-1}$$

entonces, como probamos con anterioridad, la varianza verdadera es menor que la varianza estimada. Ello implica que el incluir regresores adicionales, aumenta la varianza de nuestros parámetros estimados, lo cual se traduce en parámetros menos eficientes.

Suponga que un investigador quiere estimar el retorno a la educación y que el modelo verdadero (obviamente es un caso ilustrativo) está dado por:

$$W_i = \beta_1 + \beta_2 E_i + \epsilon_i$$

Donde W_i corresponde al logaritmo del salario del individuo i , E_i corresponde a los años de educación del individuo i y ϵ_i corresponde al término de error bien comportado.

Sin embargo este investigador utiliza el siguiente modelo para su estimación.

$$W_i = \beta_1 + \beta_2 E_i + \beta_3 D_i + \epsilon_i$$

Donde D_i corresponde a una variable dicotómica que toma el valor 1 si el individuo fuma y 0 si no fuma.

Los resultados del modelo verdadero son:

```
reg lnw esc
```

Source	SS	df	MS		
Model	8561.14296	1	8561.14296		
Residual	34520.4366	63560	.543115742		
Total	43081.5796	63561	.677798958		

Number of obs =	63562
F(1, 63560) =	15763.02
Prob > F =	0.0000
R-squared =	0.1987
Adj R-squared =	0.1987
Root MSE =	.73696

lnw	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
esc	.0863413	.0006877	125.55	0.000	.0849934 .0876892
_cons	10.98386	.0071161	1543.52	0.000	10.96991 10.99781

Los resultados del modelo estimado son:

```
reg lnw esc dfumo
```

Source	SS	df	MS		
Model	8561.25517	2	4280.62759		
Residual	34520.3244	63559	.543122522		
Total	43081.5796	63561	.677798958		

Number of obs =	63562
F(2, 63559) =	7881.51
Prob > F =	0.0000
R-squared =	0.1987
Adj R-squared =	0.1987
Root MSE =	.73697

lnw	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
esc	.0863539	.0006883	125.47	0.000	.0850049 .0877029
dfumo	-.0027275	.0060008	-0.45	0.649	-.0144891 .009034
_cons	10.9848	.0074121	1482.02	0.000	10.97028 10.99938

Podemos ver no existe una variación importante en los parámetros del modelo estimado y el modelo verdadero. Sin embargo, tal como habíamos demostrado, la varianza de los parámetros aumenta disminuyendo entonces la eficiencia.