Universidad Carlos III de Madrid

Estadística II: Introducción a la Econometría

Examen Final, Convocatoria Ordinaria, Curso 2009-2010.

7 de Junio de 2010

1. La Oficina de Estadísticas Laborales de Estados Unidos realiza todos los años una encuesta, conocida como CPS, para estudiar la situación del mercado de trabajo norteamericano. Entre las variables que se pueden construir a partir de la encuesta se encuentran el logaritmo neperiano del salario nominal a la hora ($lwage$), los años potenciales de experiencia laboral ($exper$), los años efectivos de educación ($educa$), una variable binaria que toma valor 1 si el encuestado es mujer ($female$), una binaria que toma valor 1 si el encuestado es hispano ($hispanic$), una binaria que toma valor 1 si el encuestado es negro ($black$), y una binaria que toma valor 1 si el encuestado no nació en EEUU ($alien$). Con los datos disponibles se estiman los siguientes modelos lineales por MCO:

   Modelo 1:

$$\widehat{lwage} = \underset{(.005)}{.83} + \underset{(.0002)}{.02}\, exper - \underset{(4x10^{-6})}{.0003}\, exper^2 + \underset{(.0002)}{.09}\, educa$$
$$- \underset{(.002)}{.03}\, hispanic - \underset{(.002)}{.13}\, black - \underset{(.001)}{.04}\, alien$$
$$n = 1174705; \quad SCE = 315338.7323$$

   Modelo 2:

$$\widehat{lwage} = \underset{(.004)}{.93} + \underset{(.0002)}{.02}\, exper - \underset{(4x10^{-6})}{.0003}\, exper^2 + \underset{(.0002)}{.09}\, educa - \underset{(.0009)}{.25}\, female$$
$$- \underset{(.002)}{.04}\, hispanic - \underset{(.002)}{.11}\, black - \underset{(.001)}{.04}\, alien$$
$$n = 1174705; \quad SCE = 297451.9783$$

   Modelo 3: Se generan nuevas dummies interactuando $female$ y $hispanic$:

   $nohisfem = female * (1 - hispanic)$; $hisfem = female * hispanic$; $hismale = (1 - female) * hispanic$

$$\widehat{lwage} = \underset{(.005)}{.93} + \underset{(.0002)}{.02}\, exper - \underset{(4x10^{-6})}{.0003}\, exper^2 + \underset{(.0002)}{.09}\, educa$$
$$- \underset{(.001)}{.25}\, nohisfem - \underset{(.003)}{.04}\, hismale - \underset{(.003)}{.29}\, hisfem - \underset{(.002)}{.11}\, black - \underset{(.001)}{.04}\, alien$$
$$n = 1174705; \quad SCE = 297451.8690$$

   (a) Calcule el diferencial salarial medio exacto, en tanto por ciento, entre los trabajadores hispanos y los no hispanos, ambos no nacidos en EEUU, ceteris paribus, utilizando los resultados de la estimación del Modelo 1. Contraste que dicho diferencial es significativamente negativo.

   (b) De nuevo, utilizando los resultados de la estimación del Modelo 1, calcule el diferencial salarial medio exacto, en tantos por ciento, entre un trabajador negro con 16 años de educación (es decir, con un título universitario) y un trabajador no negro con 8 años de educación (es decir, solo con educación primaria) ambos hispanos, no nacidos en EEUU y permaneciendo constantes el resto de factores que pudieran afectar.

(c) Usando el Modelo 1, ¿podemos concluir que el efecto parcial de experiencia sobre el salario es siempre positivo? Si en lugar de información cuantitativa sobre la experiencia potencial sólo tuviésemos disponible una clasificación de los trabajadores en las categorías 0 (experiencia baja), 1 (experiencia media), 2 (experiencia alta), explique cómo reformularía el modelo para permitir contrastar que cambiar de 0 a 1 pueda tener una compensación salarial diferente que pasar de 1 a 2.

(d) Calcule usando el Modelo 3, la diferencia salarial esperada entre hombres y mujeres para trabajadores hispanos en función del resto de variables explicativas. Repite el ejercicio para trabajadores no hispanos. Contraste que *ceteris paribus* los diferenciales salariales por sexo no difieren entre trabajadores de origen hispano y el resto.

2. Suponga que $\log(wage)$ mide el logaritmo del salario mensual, *educ* el número de años de educación, y *abil* el cociente de inteligencia (IQ). Considere el modelo lineal

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 abil + u. \tag{1}$$

(a) Interprete el coeficiente β_1 si se satisface el supuesto $E(u|educ, abil) = 0$. ¿Es razonable que ese supuesto se cumpla? ¿Qué ocurriría si $u = educ^2 + \varepsilon$, con ε independiente de *educ* y *abil*? ¿Y si $u = educ^2 * \varepsilon$, ε independiente de *educ* y *abil*, y de media cero?

(b) Si en lugar de (1) se estima el modelo de regresión simple

$$\log(wage) = \gamma_0 + \gamma_1 educ + v, \tag{2}$$

explique bajo qué condiciones la estimación MCO del parámetro de la variable *educ* es un estimador insesgado de β_1. En ese caso indique cómo construiría un intervalo de confianza para β_1, y si preferiría estimar (1) ó (2) para construir dicho intervalo.

(c) Suponga ahora que se consigue información sobre el cociente de inteligencia de los trabajadores de la muestra para estimar (1). Interprete la relación entre los estimadores MCO de β_1 y γ_1. Si la covarianza muestral entre el nivel educativo y el cociente de inteligencia es positiva, ¿cuál espera que sea mayor?

3. El siguiente modelo explica los logaritmos de los precios de la vivienda sobre la superficie de la vivienda $(sqrft)$, superficie del terreno $(lotsize)$, ambos en pies cuadrados, y número de dormitorios $(bdrms)$,

$$\log(price) = \beta_0 + \beta_1 sqrft + \beta_2 lotsize + \beta_3 bdrms + u.$$

Considere las siguientes salidas de Gretl para responder a las preguntas.

Modelo 1: MCO, usando las observaciones 1–88
Variable dependiente: lprice

	Coeficiente	Desv. Típica	Estadístico t	Valor p
const	4,75938	0,0935361	50,8828	0,0000
sqrft	0,000364117	4,20076e–005	8,6679	0,0000
lotsize	5,60179e–006	2,03772e–005	0,27490	,
bdrms	0,0252387	0,0285928	0,8827	,

Media de la vble. dep.	5,633180	D.T. de la vble. dep.	0,303573
Suma de cuad. residuos	3,028430	D.T. de la regresión	0,189876
R^2	0,622277	R^2 corregido	0,608787
$F(3, 84)$	46,12847	Valor p (de F)	1,02e–17

.

Modelo 2: MCO, usando las observaciones 1–88
Variable dependiente: lprice

	Coeficiente	Desv. Típica	Estadístico t	Valor p
const	5,03649	0,126347	39,8622	0,0000
bdrms	0,167226	0,0344746	4,8507	0,0000

Media de la vble. dep.	5,633180	D.T. de la vble. dep.	0,303573
Suma de cuad. residuos	6,295240	D.T. de la regresión	0,270556
R^2	0,214823	R^2 corregido	0,205693
$F(1, 86)$	23,52939	Valor p (de F)	5,43e–06

(a) Explique la hipótesis que contrasta el estadístico $F(3, 84)$ en el Modelo 1, cómo se construye el estadístico de contraste y el significado de su p-valor.

(b) Contraste la significación individual y conjunta de las variables $sqrft$ y $lotsize$. Explique la diferencia entre ambos procedimientos y las conclusiones obtenidas.

(c) Explique detenidamente cómo obtener un intervalo de confianza para el cambio porcentual en los precios (*price*) cuando se añade un dormitorio de 150 pies cuadrados a una casa y no varía el terreno disponible mediante la salida de un modelo de regresión reparametrizado.

VALORES CRITICOS:

$N(0,1)$
$\Pr(N(0,1) > 0,005) = 2,576$
$\Pr(N(0,1) > 0,01) = 2,326$
$\Pr(N(0,1) > 0,025) = 1,960$
$\Pr(N(0,1) > 0,05) = 1,645$
$\Pr(N(0,1) > 0,10) = 1,282$

$\chi^2_{(1)}$	$\chi^2_{(2)}$	$\chi^2_{(3)}$
$\Pr\left(\chi^2_{(1)} > 0,01\right) = 6,63$	$\Pr\left(\chi^2_{(2)} > 0,01\right) = 9,21$	$\Pr\left(\chi^2_{(3)} > 0,01\right) = 11,34$
$\Pr\left(\chi^2_{(1)} > 0,05\right) = 3,84$	$\Pr\left(\chi^2_{(2)} > 0,05\right) = 5,99$	$\Pr\left(\chi^2_{(3)} > 0,05\right) = 7,81$
$\Pr\left(\chi^2_{(1)} > 0,10\right) = 2,71$	$\Pr\left(\chi^2_{(2)} > 0,10\right) = 4,61$	$\Pr\left(\chi^2_{(3)} > 0,10\right) = 6,25$

Recordamos que una t de Student con n grados de libertad se comporta como un $N(0,1)$ para n razonablemente grande ($n \geq 20$). Por otro lado, una F de $Fisher$ con q grados de libertad en el numerador y n grados de libertad en el denominador se comporta como una $\chi^2_{(q)}/q$.

Universidad Carlos III de Madrid

Estadística II: Introducción a la Econometría

Examen Final, Convocatoria Ordinaria, Curso 2009-2010.

7 de Junio de 2010

SOLUCIÓN

1. La Oficina...

   a. *Calcule el diferencial salarial medio exacto, en tantos por ciento, entre los trabajadores hispanos y los no hispanos, ambos no nacidos en EEUU, ceteris paribus, utilizando los resultados de la estimación del Modelo 1. Contraste que dicho diferencial es significativamente negativo.*

   Para calcular el diferencial salarial medio,

   $$\frac{wage_{hispanic} - wage_{hispanic}}{wage_{hispanic}} \times 100,$$

   tenemos en cuenta que si $\widehat{\beta}_b$ es la estimación de la dummy de raza negra para la regresión de los logaritmos de los salarios entonces:

   $$\ln(\widehat{wage_{hispanic}}) - \ln(\widehat{wage_{hispanic}}) = \widehat{\beta}_h.$$

   Tomando antilogaritmos y substrayendo por uno:

   $$\begin{aligned}\frac{wage_{hispanic} - wage_{hispanic}}{wage_{hispanic}} \times 100 &= \left(\exp(\widehat{\beta}_h) - 1\right) \times 100 \\ &= (\exp(-0.03) - 1) \times 100 \\ &= -2.955 \approx -3\%.\end{aligned}$$

   Las hipótesis nula y alternativa son

   $$\begin{aligned}H_0 &: \beta_h = 0 \\ H_1 &: \beta_h < 0,\end{aligned}$$

   y podemos comprobar como el estadístico

   $$t = \frac{\hat{\beta}_h}{se\left(\hat{\beta}_h\right)} = \frac{-0.03}{.002} = -15$$

   es significativo a cualquier nivel de significación habitual.

   b. *De nuevo, utilizando los resultados de la estimación del Modelo 1, calcule el diferencial salarial medio exacto, en tantos por ciento, entre un trabajador negro con 16 años de educación (es decir, con un título universitario) y un trabajador no negro con 8 años de educación (es decir, solo con educación primaria) ambos hispanos, no nacidos en EEUU y permaneciendo constantes el resto de factores que pudieran afectar.*

Sea $\ln(\widehat{wage}_{b,16})$ el logaritmo del salario esperado para una persona negra con 16 años de educación y $\ln(\widehat{wage}_{h,8})$ el logaritmo del salario esperado para una persona no negra con 8 años de educación. En este caso:

$$\begin{aligned}\ln(\widehat{wage}_{b,16}) - \ln(\widehat{wage}_{nb,8}) &= \widehat{\beta}_b + \widehat{\beta}_{educ}(16-8) \\ &= \widehat{\beta}_b + 8\widehat{\beta}_e.\end{aligned}$$

Por tanto:

$$\begin{aligned}\frac{wage_{b,16} - wage_{nb,8}}{wage_{nb,8}} x100 &= \left(\exp(\widehat{\beta}_b + 8\widehat{\beta}_e) - 1\right) \times 100 \\ &= (\exp\{-.13 + 8 \times (.09)\} - 1) \times 100 \\ &= (\exp\{.59\} - 1) \times 100 \\ &= 80.4\%.\end{aligned}$$

c. *Usando el Modelo 1, ¿podemos concluir que el efecto parcial de experiencia sobre el salario es siempre positivo? Si en lugar de información cuantitativa sobre la experiencia potencial sólo tuviésemos disponible una clasificación de los trabajadores en las categorías 0 (experiencia baja), 1 (experiencia media), 2 (experiencia alta), explique cómo reformularía el modelo para permitir contrastar que cambiar de 0 a 1 pueda tener una compensación salarial diferente que pasar de 1 a 2.*

Para valores pequeños de *exper* el efecto es positivo, pero para valores mayores que

$$exper^* = \frac{-\beta_{exper}}{2\beta_{exper^2}} = \frac{0.02}{2 * .0003} = \frac{1}{.03} \approx 33.3$$

el efecto es negativo.

Podemos construir dos variables binarias, $exper1$ y $exper2$ que indiquen si se tiene experiencia media y alta, respectivamente, mientras que experiencia baja sería el grupo de referencia. El cambio de 1 a 2 sería igual a $\beta_{exper2} - \beta_{exper1}$ que no tiene porqué ser igual a β_{exper1} : H_0 sería $\beta_{exper2} - \beta_{exper1} = \beta_{exper1}$, es decir $\mathrm{H}_0 : \beta_{exper2} = 2\beta_{exper1}$.

d. *Calcule usando el Modelo 3 la diferencia salarial esperada entre hombres y mujeres para trabajadores hispanos en función del resto de variables explicativas. Repite el ejercicio para trabajadores no hispanos. Contraste que ceteris paribus los diferenciales salariales por sexo no difieren entre trabajadores de origen hispano y el resto.*

$$\begin{aligned}E[lwage|\,\mathrm{hom}, educ, exper, hispanic, black] &= \beta_0 + \beta_1 exper + \beta_2 exper^2 + \beta_3 educa + \beta_4 nohisfem \\ &\quad + \beta_5 hismale + \beta_6 hisfem + \beta_7 black + \beta_8 alien\end{aligned}$$

Para hispanos tenemos que

$$\begin{aligned}E[lwage|\,\mathrm{hom}, educ, exper, hispanic, black] &= \beta_0 + \beta_1 exper + \beta_2 exper^2 + \beta_3 educa \\ &\quad + \beta_5 + \beta_7 black + \beta_8 alien\end{aligned}$$

$$\begin{aligned}E[lwage|muj, educ, exper, hispanic, black] &= \beta_0 + \beta_1 exper + \beta_2 exper^2 + \beta_3 educa \\ &\quad + \beta_6 + \beta_7 black + \beta_8 alien\end{aligned}$$

y por tanto el diferencial es $\beta_5 - \beta_6 \equiv \beta_{hismale} - \beta_{hisfem}$, cp.

Para no hispanos

$$E\left[lwage|\,\text{hom}, educ, exper, hispanic, black\right] = \beta_0 + \beta_1 exper + \beta_2 exper^2 + \beta_3 educa + \beta_7 black + \beta_8 alien$$

$$E\left[lwage|muj, educ, exper, hispanic, black\right] = \beta_0 + \beta_1 exper + \beta_2 exper^2 + \beta_3 educa + \beta_4 + \beta_7 black + \beta_8 alien$$

y el diferencial es $-\beta_4 \equiv -\beta_{nhisfem}$.
La hipótesis nula formulada en el modelo 3 es:

$$H_0 : \beta_{nhisfem} + \beta_{hismal} = \beta_{hisfem}$$

Se contrasta utilizando la suma de residuos al cuadrado (SCE) de los modelos 2 (modelo restringido) y del modelo 3 (modelo no restringido):

$$W = F = \frac{SCE_R - SCE_{NR}}{(SCE_{NR})/n} = \frac{SCE_R - SCE_{NR}}{(SCE_{NR})} * n$$
$$= \frac{297451.9783 - 297451.8690}{297451.8690} * (1174705) = 0.431\,65$$

que se compara con el valor crítico de una chi-cuadrado con un grado de libertad. Puesto que $\Pr\left(\chi^2_{(1)} > 0,01\right) = 6,63$, el estadístico F no es mayor que el valor crítico: NO hay evidencia empírica para rechazar la hipótesis nula con un nivel de significación del 1%.

2. Suponga...

(a) *Interprete el coeficiente β_1 si se satisface el supuesto $E(u|educ, abil) = 0$. ¿Es razonable que ese supuesto se cumpla? ¿Qué ocurriría si $u = educ^2 + \varepsilon$, con ε independiente de educ y abil? ¿Y si $u = educ^2 * \varepsilon$, ε independiente de educ y abil, y de media cero?*

$100\beta_1$ es el incremento porcentual esperado (aproximado) en el salario debido al aumento de un año de educación, ceteris paribus. El supuesto se cumplirá si los otros factores que determinan *wage*, aparte de *educ* y *abil*, no cambian su efecto medio cuando cambian *educ* o *abil*. Si experiencia está relacionada con educación, por ejemplo, el supuesto no se cumpliría.

Si $u = educ^2 + \varepsilon$, entonces obtenemos que $E\left(u|educ, abil\right) = E\left(educ^2 + \varepsilon|educ, abil\right) = educ^2$ y el supuesto no se cumple.

Si $u = educ^2 * \varepsilon$, entonces obtenemos que $E\left(u|educ, abil\right) = E\left(educ^2 * \varepsilon|educ, abil\right) = educ^2 E\left(\varepsilon|educ, abil\right) = educ^2 E\left(\varepsilon\right) = 0$ y el supuesto sí se cumple.

(b) *Si en lugar de (1) se estima el modelo de regresión simple*

$$\log(wage) = \gamma_0 + \gamma_1 educ + v, \tag{3}$$

explique bajo qué condiciones la estimación MCO del parámetro de la variable educ es un estimador insesgado de β_1. En ese caso indique cómo construiría un intervalo de confianza para β_1, y si preferiría estimar (1) ó (2) para construir dicho intervalo.

Se obtendría que no hay sesgo si $E\left[v|educ\right] = E\left[\beta_2 abil + u|educ\right] = \beta_2 E\left[abil|educ\right]$ si bien $\beta_2 = 0$ (*abil* no tiene efecto parcial sobre *wage* una vez que se controla *educ*) o si

bien $E[abil|educ] = 0$, es decir los niveles de inteligencia no cambian con educación (y por tanto la covarianza entre ambas variables es cero).

Si no hay sesgo, el intervalo de confianza sería

$$\hat{\gamma}_1 \pm se(\hat{\gamma}_1) z_{\alpha/2}$$

donde

$$se(\hat{\gamma}_1) = \left(\frac{\hat{\sigma}^2}{SCT_{educ}}\right)^{1/2} = \left(\frac{\hat{\sigma}^2}{\sum_{i=1}^{n}\left(educ_i - \overline{educ}\right)^2}\right)^{1/2}$$

y $\hat{\sigma}^2$ es la varianza residual. En regresión múltiple habría que sustituir SCT_{educ} por la SCE de la regresión de *educ* sobre *abil*, que es más pequeña que SCT_{educ}, y por tanto error estándar será más grande y el IC más ancho para un mismo α. (Esto se hace suponiendo que $\hat{\sigma}^2$ no varía porque *abil* no es significativa y $\beta_2 = 0$).

(c) *Suponga ahora que se consigue información sobre el cociente de inteligencia de los trabajadores de la muestra para estimar (1). Interprete la relación entre los estimadores MCO de β_1 y γ_1. Si la covarianza muestral entre el nivel educativo y el cociente de inteligencia es positiva, ¿cuál espera que sea mayor?*

La relación es

$$\hat{\gamma}_1 = \hat{\beta}_1 + \hat{\beta}_2\hat{\delta}_{educ|abil} = \hat{\beta}_1 + \hat{\beta}_2\frac{\widehat{Cov}(educ, abil)}{\widehat{Var}(educ)}.$$

La diferencia entre $\hat{\gamma}_1$ y $\hat{\beta}_1$ es el efecto de no controlar por *abil* : hay un feedback de *educ* a *abil*, y éste a través de $\hat{\beta}_2$ produce un efecto sobre *wage*. Si $\widehat{Cov}(educ, abil) > 0$ y también esperamos que $\hat{\beta}_2 > 0$, entonces el sesgo de $\hat{\gamma}_1$ es positivo: sobre estimará β_1 al recoger el efecto conjunto de mayor educación y también mayor habilidad ligada a esos mayores niveles educativos.

3. Considere...

(a) *Explique la hipótesis que contrasta el estadístico $F(3, 84)$ en el Modelo 1, cómo se construye el estadístico de contraste y el significado de su p-valor.*

Es el contraste de significación global del modelo:

$$\begin{aligned} H_0 &: \beta_1 = \beta_2 = \beta_3 = 0 \\ H_1 &: H_0 \text{ es falsa} \end{aligned}$$

y el estadístico es

$$F = \frac{R^2}{1-R^2}\frac{n-k-1}{k} \sim_{asy} F_{k,n-k-1}$$

que en este caso es significativo, concluyendo que R^2 no es cero y las variables explicativas contribuyen a explicar la variable dependiente.

(b) *Contraste la significación individual y conjunta de las variables $sqrft$ y lotsize. Explique la diferencia entre ambos procedimientos y las conclusiones obtenidas.*

La significación individual implica el contraste

$$\begin{aligned} H_0^j &: \beta_j = 0 \\ H_1^j &: \beta_j \neq 0 \end{aligned}$$

mediante tests de la t ($sqrft$ es muy significativa, mientras que $lotsize$ no es significativa a los niveles habituales y por tanto la hipótesis H_0^1 se rechaza, pero H_0^2 no se pueden rechazar, es decir imponer un valor $\beta_2 = 0$ no implica un modelo significativamente peor si el otro parámetro se permite que sea diferente de cero), mientras que el test conjunto impone simultáneamente ambas restricciones,

$$\begin{aligned} H_0 &: \beta_1 = \beta_2 = 0 \\ H_1 &: H_0 \text{ es falsa} \end{aligned}$$

y se lleva a cabo mediante un test de la F, o equivalentemente

$$\begin{aligned} W &= qF = \frac{SCE_r - SCE_{nr}}{SCR_{nr}} (n - k - 1) \\ &= \frac{6,295240 - 3,02843}{3,02843} (88 - 3 - 1) \approx 90.612 \end{aligned}$$

que comparado con una χ^2 es claramente significativo.

(c) *Explique detenidamente cómo obtener un intervalo de confianza para el cambio porcentual en los precios (price) cuando se añade un dormitorio de 150 pies cuadrados a una casa y no varía el terreno disponible mediante la salida de un modelo de regresión reparametrizado.*

El cambio requerido es $150\beta_1 + \beta_3$, y el intervalo de confianza sería

$$150\hat{\beta}_1 + \hat{\beta}_3 \pm se\left(150\hat{\beta}_1 + \hat{\beta}_3\right) z_{\alpha/2}.$$

El estimador $\hat{\beta}_1 150 + \hat{\beta}_3$ se puede obtener directamente de la salida de MCO, pero no el se porque es necesaria la covarianza entre $\hat{\beta}_1$ y $\hat{\beta}_1 150 + \hat{\beta}_3$. En su lugar podemos definir $\theta = 150\beta_1 + \beta_3$ y sustituir $\beta_3 = \theta - 150\beta_1$ en el modelo:

$$\begin{aligned} \log(price) &= \beta_0 + \beta_1 sqrft + \beta_2 lotsize + \{\theta - 150\beta_1\} bdrms + u \\ &= \beta_0 + \beta_1 \{sqrft - 150 bdrms\} + \beta_2 lotsize + \theta bdrms + u \\ &= \beta_0 + \beta_1 x_1^* + \beta_2 lotsize + \theta bdrms + u \end{aligned}$$

donde $x_1^* = sqrft - 150bdrms$, y el coeficiente de $bdrms$ nos proporciona $\hat{\theta} = \hat{\beta}_1 150 + \hat{\beta}_3$ y su correspondiente se para construir el intervalo de confianza

$$\hat{\theta} \pm se\left(\hat{\theta}\right) z_{\alpha/2}.$$