EXAMEN DE ECONOMETRÍA

UNIVERSIDAD CARLOS III DE MADRID

CURSO 2015-16

Responda todas las preguntas en 2 horas y media. Valores críticos al final del examen.

1 A *partir de una muestra aleatoria de compra-venta de 258 viviendas en España en un periodo determinado se ha estudiado la relación entre el precio de la vivienda en euros* (Y) *y su superficie en metros cuadrados* (X_1)*, número de habitaciones* (X_2)*, si la vivienda es de obra nueva o de segunda mano (información cualitativa recogida por la variable ficticia* X_3 *igual a* 0 *si es de segunda mano e igual a* 1 *si es de nueva), si tiene garaje (información cualitativa recogida por la variable ficticia* X_4 *que toma el valor* 0 *si tiene garaje y* 1 *si no tiene), si el municipio al que pertenece está calificado como turístico (información cualitativa recogida por la variable ficticia* X_5*, que toma el valor* 0 *si es turístico y el valor* 1 *si no lo es). Se considera el siguiente modelo*

$$\log(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_1 X_2 + \beta_7 X_4 X_5 + U,$$

donde los $\beta's$ *son parámetros desconocidos y* U *un término de error. Supongamos que se cumplen todos los supuestos clásicos (condiciones del Teorema de Gauss-Markov). A continuación presentamos los resultados del ajuste del modelo, así como los de otros dos modelos restringidos (errores estándar en paréntesis).*

Variable dependiente: $\log(Y)$			
V. Explicativa	*Modelo 1*	*Modelo 2*	*Modelo 3*
Constante	0.005 (0.01)	0.004 (0.01)	0.004 (0.01)
X_1	0.014 (0.0002)	0.015 (0.0002)	0.015 (0.0002)
X_2	0.043 (0.001)	0.046 (0.001)	0.046 (0.001)
X_3	0.026 (0.002)	0.025 (0.002)	0.021 (0.002)
X_4	0.063 (0.003)	0.062 (0.002)	0.059 (0.002)
X_5	−0.032 (0.004)	−0.031 (0.004)	
X_1X_2	0.022 (0.01)	0.019 (0.01)	0.019 (0.01)
X_4X_5	−0.053 (0.03)		
R^2	0.7601	0.7571	0.7533

Todos los intervalos de confianza son al 95% *y los contrastes al* 5% *de significación.*

a. 0.5 *Cuantificar el efecto sobre el precio de un incremento en la superficie de* $10m^2$ *en una vivienda de* $100m^2$ *en términos del número de habitaciones. ¿En qué unidades se mide dicho efecto? ¿Es este efecto de igual magnitud para las viviendas nuevas y para las de segunda mano?*

Tenemos que

$$\begin{aligned}
\Delta E\left[\log(Y)|\,\mathbf{X}\right] &= E\left[\log(Y)|\,\mathbf{X}+\mathbf{\Delta X}\right] - E\left[\log(Y)|\,\mathbf{X}\right] \\
&= \beta_0 + \beta_1\left(X_1+10\right) + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6\left(X_1+10\right)X_2 + \beta_7 X_4 X_5 \\
&\quad - \left(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_1 X_2 + \beta_7 X_4 X_5\right) \\
&= \beta_1 * 10 + \beta_6 * 10 X_2
\end{aligned}$$

y por tanto

$$\begin{aligned}\Delta E\left[\log(\widehat{Y})|\, X_1,\ldots,X_5\right] &= 10\left(\hat{\beta}_1+\hat{\beta}_6 X_2\right)\\ &= 10\left(\hat{\beta}_1+\hat{\beta}_6 X_2\right)\\ &= 10\left(0.014+0.022*X_2\right)\\ &= 0.14+0.22*X_2\end{aligned}$$

Aproximando

$$\Delta E\left[\log(Y)|\, X_1,\ldots,X_5\right] \approx \frac{\Delta E\left[Y|\, X_1,\ldots,X_5\right]}{Y}.$$

se estima que una vivienda de x_2 habitaciones incrementará su precio en aproximadamente un $(14+22x_2)\%$ al aumentar su tamaño en 10m^2 manteniendo constantes el resto de variables explicativas.

El efecto es independiente del hecho de que las viviendas sean nuevas o de segunda mano.

b. 0.5 *Contraste si la variación en el precio provocada por construir una habitación adicional es independiente de la superficie de la vivienda.*

El efecto anterior $\beta_1*10+\beta_6*10X_2$ no depende de X_2 cuando $\beta_6=0$, por lo que se necesita el contraste de la interacción entre X_1 y X_2,

$$\begin{aligned}H_0 &: \beta_6=0\\ H_1 &: \beta_6\neq 0\end{aligned}$$

mediante el estadístico t

$$t_{\beta_6}=\frac{\hat{\beta}_6}{s.e.\left(\hat{\beta}_6\right)}=\frac{0.022}{0.01}=2.2$$

que es significativo comparado con el valor crítico 1.96 al 5% de significación de una normal estándard, por lo que se rechaza H_0 y confirmamos que la variación del precio depende de la superficie de la vivienda.

c. 0.75 *¿Tiene influencia sobre el precio el que una vivienda se encuentre en una zona turística? Interprete el valor estimado de este efecto ¿Cambiaría su respuesta con un nivel de significación del 1%?*

El efecto de pertenecer a una zona turística se mide mediante los parámetros β_5 y β_7, por lo que se pide hacer el contraste conjunto

$$\begin{aligned}H_0 &: \beta_5=\beta_7=0\\ H_1 &: \beta_5\neq 0 \text{ y/o } \beta_7\neq 0\end{aligned}$$

mediante un contraste de la F. Como el error es homoscedástico (condiciones Gauss-Markov), podemos usar el contraste de la F clásico,

$$\begin{aligned}F &= \frac{R^2_{no\ r}-R^2_r}{1-R^2_{no\ r}}\frac{n-k-1}{q}\\ &= \frac{0.7601-0.7533}{1-0.7601}\frac{258-7-1}{2}\\ &= 3.5431\end{aligned}$$

y compararlo con una $\chi^2_2/2$ con valor crítico al 5% de 3: como $F>vc$, se rechaza la hipótesis nula, y pertenecer a una zona turística afecta al precio de la vivienda.

d. 0.75 *Sabiendo que la covarianza estimada entre el estimador del coeficiente de X_1 y el de X_1X_2 es de* 0.00051, *proporcione un intervalo de confianza para el efecto causal de un aumento de la superficie de 10m^2 en viviendas de una habitación.*

$$
\begin{aligned}
& \left[10\left(\hat{\beta}_1+\hat{\beta}_6\right) \pm 1.96 * 10 * s.e.\left(\hat{\beta}_1+\hat{\beta}_6\right)\right] \\
= & \left[10\left(\hat{\beta}_1+\hat{\beta}_6\right) \pm 1.96 * 10 * \sqrt{\widehat{Var}\left(\hat{\beta}_1\right)+\widehat{Var}\left(\hat{\beta}_6\right)+2\widehat{Cov}\left(\hat{\beta}_1,\hat{\beta}_6\right)}\right] \\
= & 10\left(0.014+0.022\right) \pm 1.96 * 10 * \sqrt{0.0002^2+0.01^2+2 * 0.00051} \\
= & [-0.296, 1.016]
\end{aligned}
$$

es decir, el efecto está entre aproximadamente un -29.1% y un $101,6\%$ al nivel de confianza del 95%.

2. E*n un estudio a nivel nacional sobre la relación entre los alquileres medios de viviendas (ALQ) (en euros) y su precio (PRECIO) (en miles de euros), controlando por el hecho de que la vivienda se encuentre en un nucleo urbano, utilizando el modelo*

$$ALQ = \beta_0 + \beta_1 PRECIO + \beta_2 URBANA + U,$$

donde $URBANA$ toma el valor 1 si la vivienda se encuentra en una zona urbana y 0 en otro caso, los $\beta's$ son parámetros y U es un término de error, se han realizado los siguientes ajustes utilizando una muestra aleatoria de tamaño $n = 50$ de los municipios del país.

	(1)	*(2)*	*(3)*	*(4)*
V. Dependiente	*ALQ*	*PRECIO*	*PRECIO*	*ALQ*
Constante	125.9 (14.19)	−18.67 (12.00)	7.225 (8.936)	120.7 (15.71)
URBANA	0.525 (0.249)	0.182 (0.115)	0.616 (0.131)	0.0815 (0.305)
PRECIO	1.5121 (0.228)			2.240 (0.339)
INGRESO		2.731 (0.682)		
REG2		−5.095 (4.122)		
REG3		−1.778 (4.073)		
REG4		13.41 (4.048)		
R^2	0.669	0.691	0.317	0.599
SCR	20259.6	3767.6	8322.2	24565.7

a. 0.50 *¿Por qué podríamos pensar que PRECIO es una variable endógena? ¿Cuáles serían las consecuencias sobre los estimadores MCO del modelo?*

PRECIO sería endógena porque pueden existir factores que afectan al alquiler, aparte del precio y de si la vivienda es urbana (ej. las calidades de la construcción, ambiente en la vecindad, colegios cercanos, infraestructuras en los alrededores, etc) que están contenidos en el error y pueden a la vez estar correlacionados con el precio. Si ese es el caso, los estimadores MCO del modelo serían inconsistentes porque están sesgados.

b. 0.50 *Se consideran 5 instrumentos: renta familiar media ($INGRESO$ en miles de euros) y cuatro variables binarias para describir la región del país ($REG1$, $REG2$, $REG3$ y $REG4$). Sabemos con certeza que estos instrumentos son independientes del término de error. Contraste la relevancia de estos instrumentos. Establezca si estos instrumentos son débiles o no.*

Para responder a las preguntas debemos usar el estadístico F de la primera etapa (regresión (2)) que contrasta la hipótesis

$$H_0 : \beta_{INGRESO} = \beta_{REG2} = \beta_{REG3} = \beta_{REG4} = 0$$

en contra de la alternativa de que algún coeficiente es diferente de cero (notar que no se emplea $REG1$ porque habría muticolinealidad entre los instrumentos, por lo que realidad sólo hay 4 variables instrumentales no dependientes linealmente).

Teniendo en cuenta que los instrumentos son independientes del término de error, se calcula el estadístico F para la hipótesis conjunta comparando la regresión (2) con la (3) que impone la hipótesis nula de que los instrumentos no son significativos en la regresión con todas las variables exógenas, suponiendo que no hay heterocedasticidad condicional,

$$\begin{aligned} F &= \frac{R^2_{no\ r} - R^2_r}{1 - R^2_{no\ r}} \frac{n-k-1}{q} \\ &= \frac{0.691 - 0.317}{1 - 0.691} \frac{50-5-1}{4} \\ &= 13.314. \end{aligned}$$

El estadístico F es significativo al 5% comparado con una $\chi^2_4/4$ (vc $= 9.49/4 = 2.37$), no se puede rechazar H_0 y por tanto los instrumentos son relevantes, además este F de la primera etapa es mayor que 10, y por tanto no son instrumentos débiles.

c. 0.75 *¿Cuál de los ajustes de precios, el (2) o el (3), se utiliza en la segunda etapa de la estimación por mínimos cuadrados en dos etapas (MC2E)? La columna (4) ofrece la estimación MC2E. Los errores estándar, R^2 y SCR son los que ofrece el paquete GRETL al estimar los coeficientes β por MCO utilizando como variable dependiente ALQ y como variables explicativas los precios predichos en la regresión de la primera etapa. ¿Es posible realizar un contraste de significación conjunta de la variable URBANA y PRECIO utilizando la información disponible? Justifique sus respuestas.*

Se usaría la regresión (2), regresión de PRECIOS sobre todas las variables exógenas, incluyendo regresores exógenos e instrumentos (omitiendo una variable binaria para evitar la multicolinealidad).

Los errores estándard de (4) no son válidos porque no tienen en cuenta que se incluye un regresor "estimado" y se debería usar la fórmula específica para MC2E.

Tampoco es posible realizar un contraste conjunto (usando por ejemplo el contraste F de regresión basado en el coeficiente de determinación) porque el R^2 tiene el mismo problema que los errores estándar.

d. 0.75 *Explique cómo contrastaría que todos los instrumentos utilizados son exógenos. Proponga un estadístico para el contraste y establezca el criterio para rechazar la hipótesis nula de exogeneidad de todos los instrumentos.*

Se utilizaría el contraste de sobreidentificación J de Sargan, que consiste en realizar un ajuste MCO donde la variable dependiente son los residuos del ajuste MC2E (usando el modelo original, no la segunda etapa) y las variables explicativas son todas las variables exógenas del modelo ($URBANA$, $INGRESO$, $REG2$, $REG3$, $REG4$). Entonces, calcularemos el estadístico F de significación conjunta de las $m = 4$ variables instrumentales, y calcularemos el estadístico $J = mF = 4F$, que está distribuido como una Ji-cuadrado con $m - k = 4 - 1 = 3$ grados de libertad (número de variables exógenas fuera del modelo menos número de variables endógenas en el modelo = condiciones de sobreidentificación). Si el estadístico F es significativo se rechaza la hipótesis nula de que los instrumentos son conjuntamente exógenos.

3. *Considere la siguiente curva de Engel Working-Lesser para el gasto en alimentación,*

$$Y = \beta_0 + \beta_1 D + \beta_2 \ln X + \beta_3 M + \beta_4 D \ln X + \beta_5 DM + U, \tag{1}$$

donde Y es el porcentaje de gasto en alimentación respecto al total por una unidad familiar en miles de euros en un año, X es gasto total en el año, M es el número de miembros del hogar, D es una variable que toma el valor 1 si la familia reside en una ciudad de más de 10.000 habitantes y 0 en otro caso, $\beta's$ son parámetros y U un término de error. Por otro lado, se considera el siguiente modelo alternativo

$$\ln Y = \gamma_0 + \gamma_1 D + \gamma_2 \ln X + \gamma_3 M + \gamma_4 D \ln X + \gamma_5 DM + V,$$

donde los $\gamma's$ son parámetros y V un término de error.

a. 0.50 *¿En cuál de los dos modelos la elasticidad de Y respecto a X para una familia que reside en una ciudad de 20.000 habitantes es una constante, esto es, no dependende de ninguna variable del modelo? Justifique su respuesta.*

La elasticidad está dada por

$$\begin{aligned}\xi_{Y,X} &= \frac{X}{Y}\frac{dY}{dX} \\ &= \frac{d\ln Y}{d\ln X} \\ &= X\frac{d\ln Y}{dX} \\ &= \frac{1}{Y}\frac{dY}{d\ln X}\end{aligned}$$

Para el primer modelo $\xi_{Y,X} = (\beta_2 + \beta_4)/Y$, mientras que para el segundo $\xi_{Y,X} = (\gamma_2 + \gamma_4)$ ya que $D = 1$. Es constante solo para el segundo modelo.

b. 0.50 *¿Para que valores de los parámetros γ y β la elasticidad de Y respecto a X es idéntica en los dos modelos para una familia que reside en un pueblo de 100 habitantes que gasta en alimentación el 10% de su presupuesto?*

Tiene que ocurrir que

$$\frac{\beta_2 + \beta_4 * 0}{10} = \gamma_2 + \gamma_4 * 0 \quad \Longrightarrow \quad \beta_2 = 10\gamma_2,$$

ya que ahora $D = 0$.

c. 0.50 *¿Se puede comparar el ajuste de los dos modelos utilizando el coeficiente de determinación? Justifique su respuesta.*

No, porque la variable dependiente de cada modelo es diferente y por tanto la suma total de cuadrados es diferente en cada modelo.

d. 1.50 *Los gustos de las familias son muy diferentes para las familias que residen en ciudades pequeñas y grandes. Este hecho implica que la varianza de Y sea diferente dependiendo del valor que tome D, manteniendo constante X y M; esto lo podemos expresar utilizando notación matemática como $Var(Y|\, M, X, D) = \sigma_1^2 D + \sigma_2^2(1 - D)$. Explique las consecuencias de esta heterogeneidad sobre las inferencias realizadas por MCO utilizando la salida habitual que ofrece GRETL para el modelo (1) ¿Serán válidas estas inferencias sobre los parámetros β_2 y β_3 utilizando únicamente la subpoblación de las familias residentes en municipios de menos de 10.000 habitantes? Explique cómo contrastaría $H_0 : \beta_2 = \beta_3$ vs $H_1 : \beta_2 \neq \beta_3$ utilizando esta subpoblación.*

Existe heterocedasticidad y los contrastes e intervalos de confianza obtenidos con la salida de GRETL habitual son inválidos, ya que usan errores estandar no robustos.

Usando solamente la muestra condicional a $D = 0$, todas observaciones tendrían la misma varianza condicional en las otras variables y la inferencia asintótica basada en errores estandar habituales sería correcta (aunque no eficiente).

Contraste $H_0 : \beta_2 = \beta_3$: posibles respuestas: contraste F no robusto comparando R^2 o SCR del modelo restringido fijando $\beta_2 = \beta_3$, que supone reemplazar los dos regresores $\ln X$ y M por $(\ln X + M)$:

$$Y = \beta_0 + \beta_1 D + \beta_2 (\ln X + M) + \beta_4 D \ln X + \beta_5 DM + U,$$

o sustituir β_2 o β_3 por $\theta = \beta_2 - \beta_3$ y hacer un contraste de la t habitual sobre θ en

$$Y = \beta_0 + \beta_1 D + \theta \ln X + \beta_3 (\ln X + M) + \beta_4 D \ln X + \beta_5 DM + U$$

o hacer un contraste F robusto si se sigue sospechando heterocedasticidad condicional en otra variable [esto último debería recibir menos puntuación].

4. *Queremos estimar la relación causal de desempeñar un trabajo en la administración pública respecto al hecho de ser hombre o mujer y a los años de educación. Para ello consideramos una muestra de 700*

personas empleadas seleccionadas aleatoriamente de un distrito municipal a los que se les preguntó si estaban empleados por el gobierno ($GOV = 1$ si trabajan en la administración y cero en otro caso), su género ($MALE = 1$ si hombre y 0 si mujer) y los años de educación ($EDUC$). Con estos datos estimamos un modelo de probabilidad lineal que proporciona el siguiente ajuste (errores estándar en paréntesis)

$$\widehat{GOV} = \underset{(0.027)}{0.152} + \underset{(0.003)}{0.035}\,EDUC - \underset{(0.025)}{0.050}\,MALE$$

a. 0.50 *Proporcione una expresión para la media y varianza condicional de GOV dados EDUC y MALE en términos de los* verdaderos parámetros desconocidos *bajo la especificación del modelo lineal de probabilidad.*

$$\begin{aligned} E(GOV|\,EDUC, MALE) &= \beta_0 + \beta_1 EDUC + \beta_2 MALE \\ VAR(GOV|\,EDUC, MALE) &= [\beta_0 + \beta_1 EDUC + \beta_2 MALE]\,[1 - \beta_0 - \beta_1 EDUC - \beta_2 MALE] \end{aligned}$$

b. 0.50 *Interprete el resultado del ajuste para un hombre con 16 años de educación.*

Estimamos que la probabilidad de que un hombre con 16 años de educación trabaje para la administración es de $0.152 + 0.035 * 16 - 0.050 = 0.662$.

c. 0.75 *¿Cuánto de más probable es el hecho de encontrarnos a una mujer trabajando para la administración respecto a encontrarnos con un hombre con el mismo nivel educativo?*

Estimamos que la diferencia entre la probabilidad de encontrarnos una mujer trabajando para el gobierno y encontrarnos a un hombre con el mismo nivel de educación es de $\hat{\beta}_2 = 0.05$. Es el 5% más probable encontrarnos a una mujer.

d. 0.75 *Explique las limitaciones del modelo lineal de probabilidad y cómo estas dificultades son superadas por los modelos probit y logit.*

El modelo lineal de probabilidad puede producir estimaciones de la probabilidad negativas, o mayores que uno. Además los efectos parciales respecto a cualquier variable explicativa son constantes para cualquier valor de la variable, lo que no es muy lógico al tratar de explicar probabilidades.

Estas limitaciones son superadas por los modelos probit y logit, que nunca pueden dar lugar a estimaciones de probabilidades negativas o mayores que uno. En el modelo logit,

$$\Pr(GOV = 1|\,EDUC, MALE) = \frac{1}{1 + e^{-\beta_0 - \beta_1 EDUC - \beta_2 MALE}}$$

y

$$\ln\left(\frac{\Pr(GOV = 1|\,EDUC, MALE)}{\Pr(GOV = 0|\,EDUC, MALE)}\right) = \beta_0 + \beta_1 EDUC + \beta_2 MALE,$$

que proporciona una interpretación directa de los coeficientes. Por ejemplo, si la educación aumenta en un año, la variación porcentual de encontrar una persona trabajando para el gobierno respecto a que no trabaje para el mismo (la razón de probabilidades u "odds ratio") varía en un $100\beta_1\%$.

Valores Críticos:

$\Pr(F_{1,\infty} > 3.84) = 0.050$	$\Pr\left(\ \chi^2_1 > 3.84\right) = 0.050$	$\Pr(N(0,1) > 1.645) = 0.050$
$\Pr(F_{1,\infty} > 5.02) = 0.025$	$\Pr\left(\ \chi^2_1 > 5.02\right) = 0.025$	$\Pr(N(0,1) > 1.960) = 0.025$
$\Pr(F_{1,\infty} > 6.63) = 0.010$	$\Pr\left(\ \chi^2_1 > 6.63\right) = 0.010$	$\Pr(N(0,1) > 2.326) = 0.010$
$\Pr(F_{1,\infty} > 7.88) = 0.005$	$\Pr\left(\ \chi^2_1 > 7.88\right) = 0.005$	$\Pr(N(0,1) > 2.576) = 0.005$
$\Pr(F_{2,\infty} > 3.00) = 0.050$	$\Pr\left(\ \chi^2_2 > 5.99\right) = 0.050$	$\Pr(t_2 > 6.31) = 0.050$
$\Pr(F_{2,\infty} > 3.69) = 0.025$	$\Pr\left(\ \chi^2_2 > 7.38\right) = 0.025$	$\Pr(t_2 > 2.92) = 0.025$
$\Pr(F_{2,\infty} > 4.61) = 0.010$	$\Pr\left(\ \chi^2_2 > 9.21\right) = 0.010$	$\Pr(t_2 > 6.96) = 0.010$
$\Pr(F_{2,\infty} > 5.30) = 0.005$	$\Pr\left(\ \chi^2_2 > 10.6\right) = 0.005$	$\Pr(t_2 > 9.92) = 0.005$
$\Pr(F_{4,\infty} > 2.37) = 0.050$	$\Pr\left(\ \chi^2_4 > 9.49\right) = 0.050$	$\Pr(t_4 > 2.13) = 0.050$
$\Pr(F_{4,\infty} > 2.79) = 0.025$	$\Pr\left(\ \chi^2_4 > 11.1\right) = 0.025$	$\Pr(t_4 > 2.78) = 0.025$
$\Pr(F_{4,\infty} > 3.32) = 0.010$	$\Pr\left(\ \chi^2_4 > 13.3\right) = 0.010$	$\Pr(t_4 > 3.74) = 0.010$
$\Pr(F_{4,\infty} > 3.71) = 0.005$	$\Pr\left(\ \chi^2_4 > 14.8\right) = 0.005$	$\Pr(t_4 > 4.60) = 0.005$
$\Pr(F_{7,\infty} > 2.01) = 0.050$	$\Pr\left(\ \chi^2_7 > 14.1\right) = 0.050$	$\Pr(t_7 > 1.89) = 0.050$
$\Pr(F_{7,\infty} > 2.29) = 0.025$	$\Pr\left(\ \chi^2_7 > 16.0\right) = 0.025$	$\Pr(t_7 > 2.36) = 0.025$
$\Pr(F_{7,\infty} > 2.51) = 0.010$	$\Pr\left(\ \chi^2_7 > 17.6\right) = 0.010$	$\Pr(t_7 > 3.00) = 0.010$
$\Pr(F_{7,\infty} > 2.90) = 0.005$	$\Pr\left(\ \chi^2_7 > 20.3\right) = 0.005$	$\Pr(t_7 > 3.50) = 0.005$