# EXAMEN FINAL DE ECONOMETRÍA

Conteste cada pregunta en un cuadernillo diferente en dos horas y media. Todas los ejercicios tienen la misma puntuación.

1. Se desea estimar la siguiente ecuación de salarios

$$log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 exper^2 + abil + u, \tag{1}$$

donde *wage* es el salario mensual, *educ* son los años de educación, *exper* son los años de experiencia laboral y $u$ satisface los supuestos habituales del modelo de regresión lineal múltiple, pero no se puede observar la capacidad del trabajador (*abil*).

Se dispone de observaciones para las calificaciones de dos exámenes ($test1$ y $test2$) que son indicadores de la capacidad (*abil*). Se supone que las calificaciones se pueden escribir como

$$test1 = \gamma_1 abil + e1, \quad \text{Cov}(abil, e1) = 0$$

y

$$test2 = \delta_1 abil + e2, \quad \text{Cov}(abil, e2) = 0,$$

donde $\gamma_1 > 0$ y $\delta_1 > 0$. Dado que la capacidad es la que afecta al salario, se puede suponer que $test1$ y $test2$ no están correlados con $u$, y también se supone que $e1$ y $e2$ no están correlados con ninguna de las variables explicativas en (1).

(a) Explique por qué una regresión MCO de (1) con *abil* omitida producirá estimadores inconsistentes y razonar si $test1$ y $test2$ son instrumentos válidos.

(b) Si se escribe *abil* en términos de la calificación del primer examen y se inserta el resultado en (1), se obtiene

$$log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 exper^2 + \alpha_1 test1 + v. \tag{2}$$

Determine el valor de $\alpha_1$, escriba $v$ en función de $u$ y $e1$, y demuestre que $test1$ es endógena en esta ecuación. ¿Producirá una regresión MCO de (2) estimadores consistentes de $\beta_1$?

(c) Si además se supone que $e1$ y $e2$ no están correlados entre sí, ¿usaría $test2$ preferentemente como una variable de control adicional o como un instrumento para $test1$ en (2)? Razone la respuesta.

(d) Considere la ecuación (2) y las estimaciones de la Tabla 1. Contraste, si es posible, si $test2$ es un instrumento relevante para $test1$. Contraste, si es posible, si $test2$ es exógena. ¿Qué información nos proporcionan las estimaciones sobre si $test1$ es endógena o exógena (suponiendo que $test2$ es exógena)?

Table 1: Tabla de Regresiones

| | (1) | (2) | (3) | (4) | (5) | (6) |
|---|---|---|---|---|---|---|
| Var. Dependiente: | log(*wage*) | log(*wage*) | log(*wage*) | test1 | test2 | log(*wage*) |
| educ | 0.0780 | 0.0573 | 0.0478 | 2.637 | 1.258 | 0.00965 |
| | (0.00680) | (0.00792) | (0.00860) | (0.243) | (0.116) | (0.0178) |
| exper | 0.0163 | 0.0157 | 0.0179 | 0.239 | -0.290 | 0.0145 |
| | (0.0140) | (0.0140) | (0.0138) | (0.398) | (0.222) | (0.0154) |
| exper$^2$ | 0.000152 | 0.000165 | -0.0000685 | -0.0181 | 0.0307 | 0.000194 |
| | (0.000588) | (0.000591) | (0.000587) | (0.0167) | (0.00916) | (0.000656) |
| test1 | | 0.00579 | 0.00468 | | 0.146 | 0.0191 |
| | | (0.000984) | (0.000999) | | (0.0155) | (0.00424) |
| test2 | | | 0.00758 | 0.524 | | |
| | | | (0.00206) | (0.0614) | | |
| Constante | 5.517 | 5.214 | 5.194 | 47.02 | 2.672 | 4.514 |
| | (0.125) | (0.131) | (0.128) | (3.874) | (2.336) | (0.239) |
| Observaciones | 935 | 935 | 935 | 935 | 935 | 935 |
| $R^2$ | 0.131 | 0.162 | 0.176 | 0.322 | 0.267 | . |

Errores estándar robustos entre paréntesis
Todas las regresiones se ajustan por MCO, excepto (6), que se ajusta por MC2E con test2 como instrumento para test1.

2. Se desea estimar esta ecuación

$$sleep = \beta_0 + \beta_1 totwrk + \beta_2 educ + \beta_3 age + \beta_4 age^2 + \beta_5 yngkid + u$$

para explicar los minutos de sueño nocturnos (por semana), *sleep*, de una muestra de trabajadores, hombres y mujeres, en términos de *totwrk* (minutos trabajados por semana), *educ* (años de educación), *age* (edad en años) y *yngkid* (que es una variable binaria igual a uno si en el hogar hay niños menores de tres años). Se supone que $u$ satisface los supuestos habituales del modelo de regresión, incluyendo homoscedasticidad condicional. Usando las estimaciones pertinentes de la Tabla 2, responder a las siguientes preguntas.

(a) Contraste si el mismo modelo de regresión es apropiado tanto para hombres como para mujeres y si hay discriminación en contra de las mujeres en el reparto de las tareas de cuidado de los niños.

(b) Contraste si el efecto de *age* sobre *sleep* depende del género y encuentre el nivel de *age* para el que el valor esperado de *sleep* es mínimo para las mujeres, todos los demás factores fijos.

(c) Construya e interprete un intervalo de confianza al 95% para el efecto sobre *sleep* del incremento de un año de educación para un hombre.

(d) Contraste si el efecto medio sobre *sleep* de un año adicional de edad (*age*) es igual al efecto de un año menos de educación (*educ*) para hombres de 20 años de edad, todos los demás factores fijos.

Table 2: Tabla de Regresiones

| | (1) | (2) | (3) | (4) | (5) |
|---|---|---|---|---|---|
| Variable Dependiente: | sleep | sleep | sleep | sleep | sleep |
| totwrk | -0.146 | -0.163 | -0.182 | -0.183 | -0.182 |
| | (0.0191) | (0.0207) | (0.0293) | (0.0291) | (0.0293) |
| educ | -11.14 | -11.71 | -13.05 | -13.87 | -7.731 |
| | (5.747) | (5.748) | (7.767) | (7.646) | (11.58) |
| age | -8.124 | -8.697 | 7.157 | -9.230 | |
| | (11.86) | (11.79) | (13.63) | (11.78) | |
| $\text{age}^2$ | 0.126 | 0.128 | -0.0448 | 0.133 | |
| | (0.137) | (0.136) | (0.156) | (0.136) | |
| yngkid | 17.15 | -0.0228 | 60.38 | 39.54 | 60.38 |
| | (53.93) | (53.91) | (64.52) | (62.57) | (64.52) |
| female | | -87.75 | 590.5 | -226.2 | 590.5 |
| | | (35.54) | (541.6) | (162.4) | (541.6) |
| totwrkf*female | | | 0.0422 | 0.0381 | 0.0422 |
| | | | (0.0412) | (0.0406) | (0.0412) |
| educ*female | | | 2.847 | 5.748 | 2.847 |
| | | | (11.53) | (11.01) | (11.53) |
| age*female | | | -37.51 | | -37.51 |
| | | | (24.91) | | (24.91) |
| $\text{age}^2$*female | | | 0.413 | | 0.413 |
| | | | (0.289) | | (0.289) |
| yngkid*female | | | -178.7 | -128.0 | -178.7 |
| | | | (117.6) | (109.6) | (117.6) |
| $\text{age} - \text{educ}$ | | | | | 7.157 |
| | | | | | (13.63) |
| $\text{age}^2 - 41 * \text{educ}$ | | | | | -0.0448 |
| | | | | | (0.156) |
| Constante | 3825.4 | 3928.6 | 3648.2 | 4010.0 | 3648.2 |
| | (259.3) | (257.9) | (323.0) | (278.7) | (323.0) |
| Observaciones | 706 | 706 | 706 | 706 | 706 |
| $R^2$ | 0.115 | 0.123 | 0.131 | 0.126 | 0.131 |

Errores estándar entre paréntesis
Todas las regresiones se ajustan por MCO.

3. Considere un modelo de regresión lineal simple,

$$Y_i = \beta_0 + \beta_1 X_i + u_i,$$

y un instrumento binario $Z_i$ para $X_i$.

(a) Demuestre que el estimador MC2E de $\beta_1$ se puede escribir como

$$\hat{\beta}_1^{MC2E} = \frac{\bar{Y}_1 - \bar{Y}_0}{\bar{X}_1 - \bar{X}_0}$$

donde $\bar{Y}_1$ y $\bar{X}_1$ son las medias muestrales de $Y_i$ y $X_i$ (respectivamente) para la parte de la muestra en la que $Z_i = 1$ y $\bar{Y}_0$ y $\bar{X}_0$ son las medias muestrales de $Y_i$ y $X_i$ (respectivamente) para la parte de la muestra en la que $Z_i = 0$.

Pista: definiendo $n_1$ como el número de observaciones para las que $Z_i = 1$ y $n_0$ como el número de observaciones para las que $Z_i = 0$, $n = n_1 + n_0$, se puede escribir

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i = \frac{1}{n}\left(\sum_{i:Z_i=1} Y_i + \sum_{i:Z_i=0} Y_i\right) = \frac{n_1}{n}\bar{Y}_1 + \frac{n_0}{n}\bar{Y}_0.$$

Considere un modelo simple para estimar el efecto de poseer un ordenador personal (PC) sobre la nota media del grado para alumnos universitarios de último año,

$$GPA_i = \beta_0 + \beta_1 PC_i + u_i,$$

donde $PC_i$ es una variable binaria indicando la posesión de un ordenador personal.

(b) ¿Por qué podría estar correlada la posesión de un PC con $u_i$? Explique porque $PC_i$ está probablemente correlada con los ingresos anuales de los padres. ¿Esto significa que los ingresos de los padres es una buena variable instrumental para $PC_i$? ¿Por qué sí o por qué no?

(c) Suponga que, hace cuatro años, la universidad concedió becas para comprar ordenadores a la mitad de los nuevos estudiantes, y que los estudiantes que recibieron la beca fueron seleccionados aleatoriamente. Explique cómo usaría esta información para construir una variable instrumental para $PC_i$.

En particular, si le dicen que

• entre los estudiantes que recibieron las becas, un 90% de ellos tiene un PC y el grupo tiene un $GPA$ medio de 3.05

• entre los estudiantes que no recibieron las becas, un 75% de ellos tiene un PC y el grupo tiene un $GPA$ medio de 2.75.

¿Cuál sería su estimador $\hat{\beta}_1^{MC2E}$?

(d) Ahora imagine que la universidad sólo dió becas a estudiantes (seleccionados aleatoriamente) cuyos padres tenían una renta por debajo de un determinado nivel (y se dispone de la lista de estudiantes que cumplían este requisito, aunque no se disponen de datos sobre la renta de los padres). ¿Cómo modificaría su modelo y/o estrategia de estimación para obtener estimadores consistentes de $\beta_1$?

**ALGUNOS VALORES CRÍTICOS:** $Z_{0.90} = 1.282$, $Z_{0.95} = 1.645$, $Z_{0.975} = 1.96$, $\chi^2_{2,0.95} = 5.99$, $\chi^2_{2,0.975} = 7.378$, $\chi^2_{3,0.95} = 7.815$, $\chi^2_{3,0.975} = 9.348$, $\chi^2_{4,0.95} = 9.488$, $\chi^2_{4,0.975} = 11.143$, $\chi^2_{5,0.95} = 11.071$, $\chi^2_{5,0.975} = 12.833$, $\chi^2_{6,0.95} = 12.592$, $\chi^2_{6,0.975} = 14.449$, donde $\mathbb{P}(Z \leq Z_\alpha) = \alpha$ y $\mathbb{P}\left(\chi^2_m \leq \chi^2_{m,\alpha}\right) = \alpha$, $Z$ está distribuida como una normal estándar con media cero y varianza uno, y $\chi^2_m$ como una chi-cuadrado con $m$ grados de libertad.

# EXAMEN FINAL DE ECONOMETRÍA
# SOLUCIONES

1. *Se desea estimar la siguiente ecuación de salarios*

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 exper^2 + abil + u, \tag{3}$$

*donde wage es el salario mensual, educ son los años de educación, exper son los años de experiencia laboral y u satisface los supuestos habituales del modelo de regresión lineal múltiple, pero no se puede observar la capacidad del trabajador* ($abil$).

*Se dispone de observaciones para las calificaciones de dos exámenes (*$test1$ *y* $test2$*) que son indicadores de la capacidad (*$abil$*). Se supone que las calificaciones se pueden escribir como*

$$test1 = \gamma_1 abil + e1, \quad \text{Cov}\,(abil, e1) = 0$$

*y*

$$test2 = \delta_1 abil + e2, \quad \text{Cov}\,(abil, e2) = 0,$$

*donde* $\gamma_1 > 0$ *y* $\delta_1 > 0$*. Dado que la capacidad es la que afecta al salario, se puede suponer que* $test1$ *y* $test2$ *no están correlados con* $u$*, y también se supone que* $e1$ *y* $e2$ *no están correlados con ninguna de las variables explicativas en* (3)*.*

(a) *Explique por que una regresión MCO de* (3) *con* $abil$ *omitida producirá estimadores inconsistentes y razonar si* $test1$ *y* $test2$ *son instrumentos válidos.*

[50%] Es razonable pensar que $abil$ estará correlada con alguna de los regresores incluídos, en particular con $educ$, por lo que se estaría incumpliendo el supuesto $E\left[u|\, educ, exper\right] = 0$ ya que $\text{Cov}(u, educ) \neq 0$.

[50%] De igual forma, las dos ecuaciones para $test1$ y para $test2$, también implican que $\text{Cov}(test1, abil) \neq 0$ y $\text{Cov}(test2, abil) \neq 0$, y por tanto estarían correladas con el error $abil + u$, es decir, no serían exógenas (aunque previsiblemente estarían correladas con $educ$ y otros regresores potencialmente endógenos).

(b) *Si se escribe* $abil$ *en términos de la calificación del primer examen y se inserta el resultado en* (3)*, se obtiene*

$$log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 exper^2 + \alpha_1 test1 + v. \tag{4}$$

*Determine el valor de* $\alpha_1$*, escriba* $v$ *en función de* $u$ *y* $e1$*, y demuestre que* $test1$ *es endógena en esta ecuación. ¿Producirá una regresión MCO de* (4) *estimadores consistentes de* $\beta_1$*?*

[50%] Despejando $abil$ se obtiene

$$abil = \frac{1}{\gamma_1} test1 - \frac{1}{\gamma_1} e1$$

y sustituyendo en (3) se obtiene

$$\begin{aligned} \log(wage) &= \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 exper^2 + \frac{1}{\gamma_1} test1 - \frac{1}{\gamma_1} e1 + u, \\ \alpha_1 &= \frac{1}{\gamma_1}, \quad v = u - \frac{1}{\gamma_1} e1. \end{aligned}$$

[50%] En este caso, los regresores originales, $educ$, $exper$ y $exper^2$ están incorrelados con $v$ porque lo están con $e1$ y con $u$. $test1$ también está incorrelado con $u$, pero no con $e1$, ya que $\text{Cov}(test1, e1) = Var\,(e1) > 0$, por lo que $test1$ es endógena en (4) y en ese caso el estimador MCO de todos los coeficientes, incluyendo $\beta_1$, serán inconsistentes.

**Nota**. Con ese argumento general sería suficiente para la nota completa. Se podría argumentar que $test1$ es una variable de control válida, pero se debería demostrar/argumentar que $test1$ cumple las condiciones de una variable de control (es decir, una vez que se condiciona por $test1$, el valor esperado del error $v$ no cambia cuando cambia $educ$), aunque en realidad no hay un argumento para justificar esto, el problema deriva de que la parte del error $v$ que no explica $test1$, puede que esté correlada con $educ$, a pesar de que $v$ no lo está.

Para la demostración de que en particular el MCO de $\beta_1$ es consistente, la clave sería comprobar que cuando sustituimos la regresión del error $v$ sobre la posible variable de control $test1$,

$$v = \eta_0 + \eta_1 test1 + s, \quad Cov\,(test1, s) = 0, \quad \eta_1 \neq 0$$

donde $\eta_1 \neq 0$ porque hemos concluido que $test1$ es endógena, en la regresión con $test1$ y obtenemos un modelo con error $s$,

$$\begin{aligned} log(wage) &= \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 exper^2 + \alpha_1 test1 + \eta_0 + \eta_1 test1 + s \\ &= \beta_0 + \eta_0 + \beta_1 educ + \beta_2 exper + \beta_3 exper^2 + (\alpha_1 + \eta_1)\, test1 + s, \end{aligned}$$

se cumple que

$$Cov\,(educ, s) = 0 \tag{5}$$

(e igual para $exper$ y $exper^2$).

De esta regresión MCO obviamente no se puede esperar estimar consistentemente el coeficiente verdadero de $test1$, $\alpha_1$, si no $\alpha_1 + \eta_1 \neq \alpha_1$, pero además este problema se transmite a la estimación de los otros coeficientes, excepto si se cumple (5), lo cual no es cierto porque

$$\begin{aligned} Cov\,(educ, s) &= Cov\,(educ, v - \eta_1 test1) \\ &= Cov\,(educ, -\eta_1 test1) \quad \text{porque} \quad Cov\,(educ, v) = 0 \\ &= -\eta_1 Cov\,(educ, test1) \\ &= -\eta_1 Cov\,(educ, \gamma_1 abil + e1) \\ &= -\eta_1 Cov\,(educ, \gamma_1 abil) \quad \text{porque} \quad Cov\,(educ, e1) = 0 \\ &= -\eta_1 \gamma_1 Cov\,(educ, abil) \end{aligned}$$

que es diferente de cero porque esperamos que $educ$ está correlada con $abil$ y $\eta_1\gamma_1 \neq 0$.

(c) *Si además se supone que $e1$ y $e2$ no están correlados entre sí, ¿usaría $test2$ preferentemente como una variable de control adicional o como un instrumento para $test1$ en (4)? Razone la respuesta.*

[50%] Para saber si $test2$ es un buen instrumento hay que comprobar la exogeneidad de $test2$ en (4) y su relevancia para el regresor endógeno en (4) que es $test2$ :

- exogeneidad: $Cov\,(test2, v) = 0$, que se cumple porque $test2$ está incorrelado con $u$ (no está omitida en (3)) y con $e1$, ya que se asume que el error $e1$ no depende de ningún regresor en (3) .

- relevancia: $\text{Cov}(test2, test1) = \gamma_1 \delta_1 Var\,(abil) \neq 0$.

[50%] Por tanto $test2$ sería un instrumento válido, pero por esa razón no podría ser una buena variable de control porque no aportaría ninguna información sobre factores omitidos contenidos en el error $v$, ya que $Cov\,(test2, v) = 0$.

(d) *Considere la ecuación (2) y las estimaciones de la Tabla 1. Contraste, si es posible, si $test2$ es un instrumento relevante para $test1$.*

[25%] Para hacer el contraste hay que comprobar que $test2$ es significativa en la forma reducida de $test1$, regresando $test1$ sobre todas las variables exógenas

$$test1 = \pi_0 + \pi_1 educ + \pi_2 exper + \pi_3 exper^2 + \pi_4 test2 + w.$$

Se realizaría el contraste de

$$\begin{aligned} H_0 &: \pi_4 = 0 \\ H_1 &: \pi_4 \neq 0 \end{aligned}$$

con un contraste $t$.

[25%] Usando el output de la regresión (4)

$$t_4 = \frac{\hat{\pi}_4}{se\left(\hat{\pi}_4\right)} = \frac{0.524}{0.0614} = 8.534$$

que es significativo comparado con cualquier valor crítico de una $N\left(0,1\right)$, por lo que $test2$ es relevante.

*Contraste, si es posible, si $test2$ es exógena.*

[25%] El contraste de $Cov\left(test2, v\right) = 0$ no se puede realizar porque la ecuación está exactamente identificada al existir un sólo instrumento, $test2$, para el regresor endógeno, $test1$.

*¿Qué información nos proporcionan las estimaciones sobre si $test1$ es endógena o exógena (suponiendo que $test2$ es exógena)?*

[25%] Si $test1$ fuese exógena, entonces los estimadores MCO deberían ser consistentes, al igual que los estimadores MC2E, ya que el instrumento se supone exógeno. En este caso comparando las regresiones (2) y (6) en la Tabla 1 podemos ver que ciertos coeficientes, como el de *educ*, cambian sustancialmente, indicando que posiblemente algo vaya mal en la regresión MCO si damos por buena la regresión MC2E.

2. *Se desea estimar esta ecuación*

$$sleep = \beta_0 + \beta_1 totwrk + \beta_2 educ + \beta_3 age + \beta_4 age^2 + \beta_5 yngkid + u$$

*para explicar los minutos de sueño nocturnos (por semana), sleep, de una muestra de trabajadores, hombres y mujeres, en términos de totwrk (minutos trabajados por semana), educ (años de educación), age (edad en años) y yngkid (que es una variable binaria igual a uno si en el hogar hay niños menores de tres años). Se supone que u satisface los supuestos habituales del modelo de regresión, incluyendo homoscedasticidad condicional. Usando las estimaciones pertinentes de la Tabla 2, responder a las siguientes preguntas.*

(a) *Contraste si el mismo modelo de regresión es apropiado tanto para hombres como para mujeres y si hay discriminación en contra de las mujeres en el reparto de las tareas de cuidado de los niños.*

[25%] Para este contraste, debemos contrastar en el modelo que incluye el regresor binario $female$ y las interacciones de $female$ con todos los regresores, si todas esas variables que dependen de $female$ son conjuntamente significativas, es decir, contrastar en

$$\begin{aligned} sleep &= \beta_0 + \beta_1 totwrk + \beta_2 educ + \beta_3 age + \beta_4 age^2 + \beta_5 yngkid \\ &\quad + \beta_6 female + \beta_7 totwrk * female + \beta_8 educ * female + \beta_9 age * female \\ &\quad + \beta_{10} age^2 * female + \beta_{11} yngkid * female + u \end{aligned}$$

las hipótesis

$$\begin{aligned} H_0 &: \beta_6 = \cdots = \beta_{11} = 0 \\ H_1 &: H_0 \text{ es falsa.} \end{aligned}$$

[25%] Para ello, podemos realizar un contraste $F$ bajo el supuesto de homoscedasticidad,

$$F = \frac{R_{ur}^2 - R_r^2}{1 - R_{ur}^2} \frac{n-k-1}{q} = \frac{0.131 - 0.115}{1 - 0.131} \frac{706 - 11 - 1}{6} = 2.13$$

comparando el modelo restringido (1) con el no restringido (3). El valor crítico al 5% se obtiene a partir de la distribución $\chi^2(6)/6$ que aproxima la distribución del estadístico $F$ para muestras grandes, es decir $cv = 12.592/6 = 2.099$, por lo que el estatístico $F$ es significativamente diferente de cero al nivel del 5%, y podemos rechazar (marginalmente) la hipótesis $H_0$, concluyendo que la regresión para las mujeres es diferente de la regresión para los hombres.

**Note**. Contrastar las hipótesis

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

en un modelo con sólo $female$,

$$\begin{aligned} sleep &= \beta_0 + \beta_1 totwrk + \beta_2 educ + \beta_3 age + \beta_4 age^2 + \beta_5 yngkid \\ &\quad + \beta_6 female + u, \end{aligned} \tag{6}$$

no es correcto porque (6) no permite regresiones (completamente) separadas para mujeres y hombres, sino una versión restringida del modelo general donde $H_1^*$ es solo una desviación particular de la hipótesis de regresiones iguales, una vez que las restricciones $\beta_7 = \cdots = \beta_{11} = 0$ son impuestas sin justificación.

[25%] Para contrastar discriminación en contra de las mujeres, se deberá contrastar

$$\begin{aligned} H_0 &: \beta_{11} = 0 \\ H_1 &: \beta_{11} < 0 \end{aligned}$$

donde la alternativa indica que las mujeres duermen menos en promedio que los hombres cuando hay niños presentes, mediante un contraste $t$,

[25%]

$$t_{11} = \frac{\hat{\beta}_{11}}{se\left(\hat{\beta}_{11}\right)} = \frac{-178.7}{117.6} - 1.519\,6$$

que no es significativo al nivel del 5%, para el que el valor crítico unilateral de la $N(0,1)$ es $-1.645$, por lo que no hay suficiente evidencia que apoye tal discriminación.

(b) *Contraste si el efecto de age sobre sleep depende del género y encuentre el nivel de age para el que el valor esperado de sleep es mínimo para las mujeres, todos los demás factores fijos.*

[30%] Las hipótesis que deben ser contrastadas son

$$\begin{aligned} H_0 &: \beta_9 = \beta_{10} = 0 \\ H_0 &: \beta_9 \neq 0 \;\text{ o }\; \beta_{10} \neq 0 \end{aligned}$$

mediante un contraste $F$ bajo el supuesto de homocedasticidad,

[30%]

$$F = \frac{R_{ur}^2 - R_r^2}{1 - R_{ur}^2} \frac{n-k-1}{q} = \frac{0.131 - 0.126}{1 - 0.131} \frac{706 - 11 - 1}{2} = 2,$$

comparando el modelo restringido (4) con el no restringido (3). El valor crítico al 5% se obtiene a partir de la distribución $\chi^2(2)/2$ que aproxima la distribución del estadístico $F$ para muestras grandes, es decir $cv = 5.99/2 = 2.99$, por lo que el estatístico $F$ no es significativamente diferente de cero al nivel del 5%, concluyendo que no hay evidencia empírica suficiente para apoyar que el efecto de *age* sobre *sleep* es diferente para cada género.

[40%] Para las mujeres, la parte del valor esperado de *sleep* que depende de *age* es

$$(\beta_3 + \beta_9)\, age + (\beta_4 + \beta_{10})\, age^2$$

y, dado que $\hat{\beta}_4 + \hat{\beta}_{10} > 0$, el valor de *age* que lo hace mínimo es

$$age^*_{female} = -\frac{\hat{\beta}_3 + \hat{\beta}_9}{2\left(\hat{\beta}_4 + \hat{\beta}_{10}\right)} = -\frac{7.157 - 37.51}{2 * (-0.0448 + 0.413)} = 41.22 \text{ años.}$$

(c) *Construya e interprete un intervalo de confianza al 95% para el efecto sobre sleep del incremento de un año de educación para un hombre.*

[75%] Este efecto se describe mediante el coeficiente $\beta_2$, por lo que el intervalo de confianza es

$$\hat{\beta}_2 \pm 1.96 se\left(\hat{\beta}_2\right)$$

es decir, usando la regresión (3) se obtiene

$$-13.05 \pm 1.96 \cdot 7.767 \quad \text{o} \quad [-28.273,\; 2.173\,3]$$

[25%] indicando que este efecto no es significativamente diferente de cero al nivel del 5%.

(d) *Contraste si el efecto medio sobre sleep de un año adicional de edad (age) es igual al efecto de un año menos de educación (educ) para hombres de 20 años de edad, todos los demás factores fijos.*

[30%] Los dos efectos son

$$\begin{aligned}
& E\left[sleep|\, age=21, male, x\right] - E\left[sleep|\, age=20, male, x\right] \\
= \;& \beta_3\left(21\right)+\beta_4\left(21\right)^2-\left(\beta_3\left(20\right)+\beta_4\left(20\right)^2\right) \\
= \;& \beta_3+41\beta_4,
\end{aligned}$$

y

$$E\left[sleep|\, educ-1, male, x\right] - E\left[sleep|\, educ, male, x\right] = \beta_2\left(educ-1\right)-\beta_2 educ = -\beta_2,$$

respectivamente, y la igualdad entre ellos implica que

$$H_0 : \theta = 0$$

donde

$$\theta = \left(\beta_3+41\beta_4\right)-\left(-\beta_2\right)=\beta_2+\beta_3+41\beta_4$$

[40%] y sustituyendo $\beta_2 = \theta - \left(\beta_3+41\beta_4\right)$ en el modelo se obtiene

$$\begin{aligned}
sleep \;=\;& \beta_0+\beta_1 totwrk+\left\{\theta-\left(\beta_3+41\beta_4\right)\right\} educ+\beta_3 age+\beta_4 age^2+\beta_5 yngkid \\
& +\beta_6 female+\beta_7 totwrk*female+\beta_8 educ*female+\beta_9 age*female \\
& +\beta_{10} age^2*female+\beta_{11} yngkid*female+u
\end{aligned}$$

o equivalentemente,

$$\begin{aligned}
sleep \;=\;& \beta_0+\beta_1 totwrk+\theta educ+\beta_3\left(age-educ\right)+\beta_4\left(age^2-41educ\right)+\beta_5 yngkid \\
& +\beta_6 female+\beta_7 totwrk*female+\beta_8 educ*female+\beta_9 age*female \\
& +\beta_{10} age^2*female+\beta_{11} yngkid*female+u
\end{aligned}$$

por lo que para contrastar $H_0$ en contra de $H_1 : \theta \neq 0$ se usa un contraste $t$ del coeficiente de *educ* en la regresión (5) de la Tabla 2,

[30%]

$$t_\theta = \frac{\hat{\theta}}{se\left(\hat{\theta}\right)} = \frac{-7.731}{11.58} = -0.667\,62$$

que no es significativo comparado con el valor crítico de la $N\left(0,1\right)$ a cualquier nivel de significación estándar, por lo que no es posible rechazar la hipótesis nula de igualdad entre ambos efectos.

3. *Considere un modelo de regresión lineal simple,*

$$Y_i = \beta_0+\beta_1 X_i+u_i,$$

*y un instrumento binario $Z_i$ para $X_i$.*

(a) *Demuestre que el estimador MC2E de $\beta_1$ se puede escribir como*

$$\hat{\beta}_1^{MC2E} = \frac{\bar{Y}_1-\bar{Y}_0}{\bar{X}_1-\bar{X}_0}$$

*donde $\bar{Y}_1$ y $\bar{X}_1$ son las medias muestrales de $Y_i$ y $X_i$ (respectivamente) para la parte de la muestra en la que $Z_i = 1$ y $\bar{Y}_0$ y $\bar{X}_0$ son las medias muestrales de $Y_i$ y $X_i$ (respectivamente) para la parte de la muestra en la que $Z_i = 0$.*

*Pista: definiendo $n_1$ como el número de observaciones para las que $Z_i = 1$ y $n_0$ como el número de observaciones para las que $Z_i = 0$, $n = n_1 + n_0,$ se puede escribir*

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i = \frac{1}{n}\left(\sum_{i:Z_i=1} Y_i + \sum_{i:Z_i=0} Y_i\right) = \frac{n_1}{n}\bar{Y}_1 + \frac{n_0}{n}\bar{Y}_0.$$

[50%, 10% por la primera expresión de MC2E] Por la definición del estimador MC2E

$$\begin{aligned}\hat{\beta}_1^{MC2E} &= \frac{\widehat{Cov}\left(Y,Z\right)}{\widehat{Cov}\left(X,Z\right)} = \frac{\frac{1}{n}\sum_{i=1}^{n} Y_i Z_i - \bar{Y}\bar{Z}}{\frac{1}{n}\sum_{i=1}^{n} X_i Z_i - \bar{X}\bar{Z}} \\ &= \frac{\frac{1}{n}\sum_{i:Z_i=1} Y_i - \bar{Y}\frac{n_1}{n}}{\frac{1}{n}\sum_{i:Z_i=1} X_i - \bar{X}\frac{n_1}{n}} = \frac{\frac{n_1}{n}\bar{Y}_1 - \bar{Y}\frac{n_1}{n}}{\frac{n_1}{n}\bar{X}_1 - \bar{X}\frac{n_1}{n}} = \frac{\bar{Y}_1 - \bar{Y}}{\bar{X}_1 - \bar{X}}\end{aligned}$$

porque

$$\bar{Z} = \frac{1}{n}\sum_{i=1}^{n} Z_i = \frac{1}{n}\sum_{i:Z_i=1} 1 = \frac{n_1}{n}$$

donde $n_1$ es el número de observaciones para las que $Z_i = 1$, y

$$\bar{Y}_1 = \frac{1}{n_1}\sum_{i:Z_i=1} Y_i, \quad \bar{X}_1 = \frac{1}{n_1}\sum_{i:Z_i=1} X_i.$$

Si $n_0$ es el número de observaciones para las que $Z_i = 0$, $n = n_1 + n_0,$ se puede escribir

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i = \frac{1}{n}\left(\sum_{i:Z_i=1} Y_i + \sum_{i:Z_i=0} Y_i\right) = \frac{n_1}{n}\bar{Y}_1 + \frac{n_0}{n}\bar{Y}_0$$

[50%] e igualmente para $\bar{X}$, obteniendo

$$\begin{aligned}\hat{\beta}_1^{MC2E} &= \frac{\bar{Y}_1 - \bar{Y}}{\bar{X}_1 - \bar{X}} = \frac{\bar{Y}_1 - \frac{n_1}{n}\bar{Y}_1 - \frac{n_0}{n}\bar{Y}_0}{\bar{X}_1 - \frac{n_1}{n}\bar{X}_1 - \frac{n_0}{n}\bar{X}_0} = \frac{\bar{Y}_1\left(1 - \frac{n_1}{n}\right) - \frac{n_0}{n}\bar{Y}_0}{\bar{X}_1\left(1 - \frac{n_1}{n}\right) - \frac{n_0}{n}\bar{X}_0} \\ &= \frac{\bar{Y}_1\frac{n_0}{n} - \frac{n_0}{n}\bar{Y}_0}{\bar{X}_1\frac{n_0}{n} - \frac{n_0}{n}\bar{X}_0} = \frac{\bar{Y}_1 - \bar{Y}_0}{\bar{X}_1 - \bar{X}_0}.\end{aligned}$$

*Considere un modelo simple para estimar el efecto de poseer un ordenador personal (PC) sobre la nota media del grado para alumnos universitarios de último año,*

$$GPA_i = \beta_0 + \beta_1 PC_i + u_i,$$

*donde $PC_i$ es una variable binaria indicando la posesión de un ordenador personal.*

(b) *¿Por qué podría estar correlada la posesión de un PC con $u_i$? Explique porque $PC_i$ está probablemente correlada con los ingresos anuales de los padres. ¿Esto significa que los ingresos de los padres es una buena variable instrumental para $PC_i$? ¿Por qué sí o por qué no?*

[30%] La renta de los padres puede estar correlada con muchos factores causales incluidos en $u$ que describen diferentes aspectos de la educación previa y el acceso a oportunidades de aprendizaje que afectan al $GPA$, y también estaría correlada con la propiedad de un PC, todo lo demás igual, debido a la disponibilidad de un presupuesto mayor.

[30%] Esto implica que $u$ y $PC$ estarían correlados a través de los ingresos

[40%] y, por lo tanto, $PC$ es endógeno en la ecuación. En resumen, el ingreso de los padres estaría correlado con $PC$ (relevancia) pero también con $u$ (por lo tanto, no es exógeno), por lo que no sería un instrumento válido.

(c) *Suponga que, hace cuatro años, la universidad concedió becas para comprar ordenadores a la mitad de los nuevos estudiantes, y que los estudiantes que recibieron la beca fueron seleccionados aleatoriamente. Explique cómo usaría esta información para construír una variable instrumental para $PC_i$.*

[40%] Se debería construir una variable instrumental $Z_i$ de tal forma que $Z_i = 1$ si el estudiante recibió la beca y $Z_i = 0$ si no fue así. Se espera que $Z_i$ está correlada con $PC_i$ porque recibir la beca incentiva la compra del ordenador, todo lo demás igual, incluso si no todo el mundo que recibió la beca lo compró (o si hay estudiantes que sin recibir la beca tienen un PC).

[30%] A la vez, $Z_i$ debería ser independiente de cualquier factor contenido en $u_i$ porque fue asignado aleatoriamente, y por tanto es exógena.

*En particular, si le dicen que*

- *entre los estudiantes que recibieron las becas, un 90% de ellos tiene un PC y el grupo tiene un $GPA$ medio de 3.05*
- *entre los estudiantes que no recibieron las becas, un 75% de ellos tiene un PC y el grupo tiene un $GPA$ medio de 2.75.*

*¿Cuál sería su estimador $\hat{\beta}_1^{MC2E}$?*

[30%]

$$\hat{\beta}_1^{MC2E} = \frac{\bar{Y}_1 - \bar{Y}_0}{\bar{X}_1 - \bar{X}_0} = \frac{3.05 - 2.75}{0.90 - 0.75} = 2.$$

(d) *Ahora imagine que la universidad sólo dió becas a estudiantes (seleccionados aleatoriamente) cuyos padres tenían una renta por debajo de un determinado nivel (y se dispone de la lista de estudiantes que cumplían este requisito, aunque se disponen de datos sobre la renta de los padres). ¿Cómo modificaría su modelo y/o estrategia de estimación para obtener estimadores consistentes de $\beta_1$?*

[30%] En este caso la variable $Z_i$ definida anteriormente estaría correlada (negativamente) con algunos factores en $u_i$ relacionados con la renta de los padres, ya que $Z_i$ se asignó de forma diferente en función de esa renta.

[30%] Sin embargo, si se construye otra variable binaria $W_i = 1$ para los estudiantes que cumplían los requisitos para la beca, y $W_i = 0$ para los que no, y se incluye en la regresión estimada junto con $PC_i$,

$$GPA_i = \beta_0 + \beta_1 PC_i + \beta_2 W_i + v_i$$

[40%] ahora $Z_i$ estaría incorrelada con los factores causales restantes incluidos en el nuevo término de error $v_i$ porque $Z_i$ fue asignada independientemente de ellos (condicionando en $W_i$) y sería un instrumento válido.