EXAMEN EXTRAORDINARIO DE ECONOMETRÍA

Conteste cada pregunta en un cuadernillo diferente en dos horas y media.
Todas las cuestiones (a) , (b) , etc., tienen la misma puntuación.

1. Un modelo para estimar los efectos del tabaquismo en el ingreso anual (quizás a través de ausencias laborales debido a enfermedades o efectos sobre la productividad) es

$$\log(income) = \beta_0 + \beta_1 cigs + \beta_2 educ + \beta_3 age + \beta_4 age^2 + u_1 \quad (1)$$

donde *cigs* es el número de cigarrillos que se consumen al día, en promedio, y *educ* (años de educación) y *age* (edad) se suponen exógenas.

Para reflejar el hecho de que el consumo de cigarrillos podría estar determinado conjuntamente con los ingresos, se especifica una ecuación de demanda de cigarrillos,

$$cigs = \gamma_0 + \gamma_1 \log(income) + \gamma_2 educ + \gamma_3 age + \gamma_4 age^2 + \gamma_5 \log(cigpric) + \gamma_6 restaurn + u_2, \quad (2)$$

donde *cigpric* es el precio de una cajetilla de cigarrillos (en centavos) y *restaurn* es una variable binaria igual a la unidad si la persona vive en un estado cuyos restaurantes restringen el consumo de tabaco, y se supone que éstas son variables exógenas al individuo.

   (a) ¿Cómo se interpretan β_1 y γ_1? ¿Qué signos se esperarían para γ_5 y γ_6? ¿Bajo qué supuestos estarían las ecuaciones (1) y (2) identificadas, si es posible que lo estén?

   (b) Realize los contrastes que sean factibles para comprobar las dos condiciones para la identificación de las ecuaciones (1) y (2) dada la información en la Tabla 1. Comente cualquier supuesto adicional que se utilize.

   (c) Compare la estimación de la ecuación de ingreso mediante MCO y mediante MC2E. ¿Qué conclusiones puede extraer? Proporcione una estimación de la variación esperada en los ingresos de un aumento del consumo de tabaco en un cigarrillo al día. ¿Podría realizar un contraste válido con la información proporcionada para determinar si la variación es negativa?

Table 1: Tabla de Regresiones

	(1)	(2)	(3)	(4)	(5)	(6)
Var. dependiente:	log(*income*)	cigs	cigs	log(*income*)	hatu2	hatu2
cigs	0.00173					
	(0.00143)					
educ	0.0604	-0.450	-0.472	0.0397	-0.000301	-1.15e-10
	(0.00745)	(0.156)	(0.156)	(0.0115)	(0.0103)	(0.0103)
age	0.0577	0.823	0.824	0.0938	-0.000269	-3.60e-10
	(0.00920)	(0.136)	(0.137)	(0.0172)	(0.0115)	(0.0116)
age^2	-0.000631	-0.00959	-0.00958	-0.00105	0.00000214	3.17e-12
	(0.0000984)	(0.00144)	(0.00146)	(0.000197)	(0.000123)	(0.000124)
log(*cigpric*)		-0.351			0.439	
		(6.027)			(0.392)	
restaurn		-2.736			-0.0145	
		(1.001)			(0.0675)	
hatcigs				-0.0421		
				(0.0188)		
Constante	7.795	1.580	-0.332	7.781	-1.784	1.05e-08
	(0.208)	(25.19)	(3.226)	(0.208)	(1.668)	(0.259)
Observaciones	807	807	807	807	807	807
R^2	0.165	0.071	0.044	0.169	0.002	0.001

Errores estándar en paréntesis. Todas las regresiones están ajustadas por MCO.
hatcigs son las predicciones de cigs en la regresión de la columna (2).
hatu2 son los residuos del ajuste MC2E de la función de demanda, ecuación (2), usando lcigpric y restaurn como instrumentos para cigs

2. Se desea estimar un modelo de probabilidad lineal para estudiar los determinantes que llevan a la subscripción de un plan de pensiones,

$$pp = \beta_0 + \beta_1 \log(income) + \beta_2 age + \beta_3 age^2 + \beta_4 male + \beta_5 married + \beta_6 male * married + u, \quad (3)$$

donde $pp = 1$ si el individuo ha subscrito un plan y 0 en caso contrario, *income* es el ingreso anual, *age* la edad en años, *married* es una variable binaria indicando si el individuo está casado y *male* una variable binaria indicando si el individuo es un hombre. El modelo (3) se ha estimado con una muestra de 9275 individuos y los resultados están resumidos en la Tabla 2, junto con los de otros modelos relacionados.

(a) Interprete el coeficiente β_1 en (3) y proporcione un intervalo de confianza al 95% para el efecto sobre *pp* de un aumento de los ingresos del 10%.

(b) Determine si la probabilidad de subscribir un plan de pensiones de un hombre soltero es diferente a la de una mujer soltera usando el modelo (3). ¿Cuál es la diferencia estimada entre dicha probabilidad para un hombre y una mujer, ambos casados?

(c) Explique por qué con la información proporcionada no es posible llevar a cabo un contraste de hipótesis válido acerca de si la probabilidad de subscribir un plan de pensiones depende del género y estado civil del individuo.

(d) Finalmente, se decide estimar modelos diferentes para hombres y para mujeres incluyendo como variables explicativas $\log(income), age, age^2$ y *married*. Proporcione la ecuación estimada para los hombres usando los resultados de la columna (3).

Table 2: Tabla de Regresiones

	(1)	(2)	(3)
Variable dependiente:	pp	pp	pp
log(*income*)	0.229	0.217	0.231
	(0.00838)	(0.00765)	(0.00948)
age	0.0137	0.0142	0.0158
	(0.00347)	(0.00346)	(0.00394)
age^2	-0.000160	-0.000165	-0.000181
	(0.0000398)	(0.0000397)	(0.0000448)
male	-0.0151		0.197
	(0.0143)		(0.177)
married	-0.0360		-0.0374
	(0.0116)		(0.0118)
male*married	-0.00875		-0.00145
	(0.0244)		(0.0258)
log(*income*) $*$ *male*			-0.0112
			(0.0203)
age*male			-0.00761
			(0.00867)
age^2*male			0.0000747
			(0.000102)
Constante	-0.774	-0.770	-0.830
	(0.0723)	(0.0718)	(0.0824)
Observaciones	9275	9275	9275
R^2	0.084	0.083	0.084

Errores estándar robustos en paréntesis
Todas las regresiones están ajustadas por MCO

3. Suponga que el modelo verdadero que se desea estimar para explicar los resultados de un examen es

$$Testscore_i = \beta_0 + \beta_1 STR_i + \beta_2 \textit{Effort}_i + u_i$$

donde STR es una variable continua que aumenta cuando el tamaño de la clase aumenta. *Effort* es una variable continua que aumenta cuando el estudiante pone más esfuerzo. Finalmente, el error u incluye todos los otros factores que se supone están incorrelados con las otras variables explicativas del modelo. Suponga que $\beta_1 > 0$ y $\beta_2 > 0$, $\text{Cov}(STR, \textit{Effort}) > 0$, $\text{Cov}(STR, u) = 0$, $\text{Cov}(\textit{Effort}, u) = 0$.

Sin embargo, el económetra estima la siguiente especificación,

$$Testscore_i = \gamma_0 + \gamma_1 STR_i + e_i. \tag{4}$$

(a) Discuta si la estimación MCO de γ_1 en (4) está sesgada o no para β_1, y si es así, cuál es el signo del sesgo.

(b) Imagine que en lugar de *Effort* se observa un indicador de la asistencia a clase, CA_i, para el que se sabe que $\text{Cov}(CA, \textit{Effort}) > 0$, por lo que se cumple que

$$\textit{Effort}_i = \alpha_0 + \alpha_1 CA_i + v_i, \quad \text{Cov}\left(CA, v\right) = 0, \alpha_1 > 0.$$

También se sabe que

$$\text{Cov}\left(STR, v\right) = 0.$$

Interprete esta condición.

(c) ¿Cuál sería el signo esperado de δ_3 en el siguiente modelo de regresión?

$$Testscore_i = \delta_0 + \delta_1 STR_i + \delta_3 CA_i + w_i$$

¿Cree que el estimador MCO de δ_1 sería sesgado o insesgado para β_1? ¿Por qué?

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 $\mathbb{P}\left(Z \leq Z_\alpha\right) = \alpha$ 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 χ^2_m como una chi-cuadrado con m grados de libertad.

EXAMEN EXTRAORDINARIO DE ECONOMETRÍA SOLUCIONES

1. *Un modelo para estimar los efectos del tabaquismo en el ingreso anual (quizás a través de ausencias laborales debido a enfermedades o efectos sobre la productividad) es*

$$\log(income) = \beta_0 + \beta_1 cigs + \beta_2 educ + \beta_3 age + \beta_4 age^2 + u_1 \quad (5)$$

donde cigs es el número de cigarrillos que se consumen al día, en promedio, y educ (años de educación) y age (edad) se suponen exógenas.

Para reflejar el hecho de que el consumo de cigarrillos podría estar determinado conjuntamente con los ingresos, se especifica una ecuación de demanda de cigarrillos,

$$cigs = \gamma_0 + \gamma_1 \log(income) + \gamma_2 educ + \gamma_3 age + \gamma_4 age^2 + \gamma_5 \log(cigpric) + \gamma_6 restaurn + u_2, \quad (6)$$

donde cigpric es el precio de una cajetilla de cigarrillos (en centavos) y restaurn es una variable binaria igual a la unidad si la persona vive en un estado cuyos restaurantes restringen el consumo de tabaco, y se supone que éstas son variables exógenas al individuo.

(a) *¿Cómo se interpretan β_1 y γ_1?*

[15%] β_1 : si se consume un cigarrillo más al día, el ingreso anual se incrementa (aproximadamente) en un $100\beta_1\%$ en promedio, todo lo demás igual.

[15%] γ_1 : si se incrementa el ingreso en un 1%, la demanda de cigarrillos aumenta (aproximadamente) en $\gamma_1/100$ unidades (cigarrillos) en promedio, todo lo demás igual.

[20%] *¿Qué signos se esperarían para γ_5 y γ_6?*

γ_5 debería ser negativo (si aumenta el precio, la demanda de cigarrillos baja) y γ_6 debería ser negativo (si hay restricciones al consumo, la demanda de cigarrillos debería bajar, todo lo demás igual).

¿Bajo qué supuestos estarían las ecuaciones (5) y (6) identificadas, si es posible que lo estén?

[30%] La ecuación *(5)* está identificada si los dos posibles instrumentos, $\log(cigpric)$ y $restaurn$, son exógenos en esa ecuación, es decir, se cumple que

$$H_0^{(exog)} : \text{Cov}\left(log\left(cigpric\right), u_1\right) = 0 \text{ y } \text{Cov}\left(restaurn, u_1\right) = 0,$$

y si al menos uno de ellos es relevante, es decir, si en la forma reducida de $cigs$,

$$cigs = \pi_0 + \pi_1 educ + \pi_2 age + \pi_3 age^2 + \pi_4 \log(cigpric) + \pi_5 restaurn + v$$

la hipótesis nula

$$H_0^{(no\ rel)} : \pi_4 = \pi_5 = 0$$

es falsa, por lo que

$$H_1^{(no\ rel)} : \pi_4 \neq 0 \ \text{ y/o } \ \pi_5 \neq 0$$

es cierta.

[20%] La ecuación *(6)* no puede estar identificada porque no hay instrumentos exógenos que puedan usarse para el regresor endógeno $\log(income)$.

(b) *Realize los contrastes que sean factibles para comprobar las dos condiciones para la identificación de las ecuaciones (5) y (6) dada la información en la Tabla 1. Comente cualquier supuesto adicional que se utilize.*

[50%] Para contrastar $H_0^{(exog)}$ en contra de

$$H_1^{(exog)} : \text{Cov}\left(log\left(cigpric\right), u_1\right) \neq 0 \text{ y/o } \text{Cov}\left(restaurn, u_1\right) \neq 0$$

podemos emplear el contraste de sobre-identificación de Hansen/Sargan, que se basa en el estádístico F para contrastar la hipótesis

$$H_0 : \delta_4 = \delta_5 = 0$$

en contra de

$$H_1 : \delta_4 \neq 0 \ \text{ y/o } \ \delta_5 \neq 0$$

en la regresión de los residuos MC2E sobre todas las variables exógenas,

$$\hat{u}_2 = \delta_0 + \delta_1 educ + \delta_2 age + \delta_3 age^2 + \delta_4 \log(cigpric) + \delta_5 restaurn + e.$$

Para ello podemos construir el estadístico F sólo válido para homocedasticidad (que es un supuesto adicional), comparando las columnas (5) = modelo no restringido y (6) = modelo restringido,

$$F = \frac{R^2_{no\ res} - R^2_{res}}{1 - R^2_{no\ res}} \frac{n - k_{no\ res} - 1}{m} = \frac{0.002 - 0.001}{1 - 0.002} \frac{807 - 5 - 1}{2} = 0.401\,3$$

donde $k_{no\ res} = 5$ es el número de regresores en la ecuación no restringida y $m = 2$ es igual al número de restricciones (igual al número de instrumentos) y el estadístico J,

$$J = mF = 2 \cdot 0.401\,3 = 0.8026.$$

Dado el número de regresores endógenos, $k = 1$, se compara J con el valor crítico de una $\chi^2_{m-k} = \chi^2_1$, donde $m - k = 1$ es el grado de sobreidentificación, que al nivel de significación del 5% es igual a 3.86. Por lo tanto el estadístico J no es significativo al 5%, y no podemos rechazar la hipótesis nula de exogeneidad.

[50%] Para contrastar la hipótesis nula

$$H_0^{(no\ rel)} : \pi_4 = \pi_5 = 0$$

en contra de

$$H_1^{(no\ rel)} : \pi_4 \neq 0 \ \text{ y/o } \ \pi_5 \neq 0$$

usamos también un estadístico F sólo válido para homocedasticidad, comparando las columnas (2) = modelo no restringido y (3) = modelo restringido,

$$F = \frac{R^2_{no\ res} - R^2_{res}}{1 - R^2_{no\ res}} \frac{n - k_{no\ res} - 1}{q} = \frac{0.071 - 0.044}{1 - 0.071} \frac{807 - 5 - 1}{2} = 11.640,$$

donde $k_{no\ res} = 5$ es el número de regresores en la ecuación no restringida y $q = 2$ es igual al número de restricciones (igual al número de instrumentos) y comparando el valor de F con el valor crítico de una $\chi^2_q/2 = \chi^2_2/2$, que al nivel de significación del 5% es igual a $5.99/2 \approx 3.00$, el estadístico F es significativo, rechazando $H_0^{(no\ rel)}$ y concluyendo que los instrumentos son relevantes, y además $F > 10$, por lo que podemos afirmar que los instrumentos no son débiles.

(c) *Compare la estimación de la ecuación de ingreso mediante MCO y mediante MC2E. ¿Qué conclusiones puede extraer?*

[40%] La estimación MCO (columna 1) y la estimación MC2E (columna 4, segunda etapa de MC2E) de la ecuación de ingreso dan resultados muy diferentes: el coeficiente estimado para *cigs* cambia de signo (0.00173 y −0.0421, respectivamente), por lo que podemos sospechar que la estimación MCO es inconsistente por un problema de endogeneidad de *cigs* en esa ecuación.

Proporcione una estimación de la variación esperada en los ingresos de un aumento del consumo de tabaco en un 1 cigarrillo al día.

[30%] Si $\Delta cigs = 1$, entonces los ingresos aumentarían en un $100 \cdot (-0.0421)\,\%$ en promedio, es decir bajarían en un 4,2%, usando la estimación MC2E.

¿Podría realizar un contraste válido con la información proporcionada para determinar si la variación es negativa?

[30%] No, porque los errores estándar de la estimación de la segunda etapa de MC2E obtenidos por MCO en la columna (4) reemplazando *cigs* por *hatcigs* no son válidos al no tener en cuenta los efectos de la primera etapa (se deberían usar e.e. específicos).

2. *Se desea estimar un modelo de probabilidad lineal para estudiar los determinantes que llevan a la subscripción de un plan de pensiones,*

$$pp = \beta_0 + \beta_1 \log(income) + \beta_2 age + \beta_3 age^2 + \beta_4 male + \beta_5 married + \beta_6 male * married + u, \quad (7)$$

donde $pp = 1$ si el individuo ha subscrito un plan y 0 en caso contrario, income es el ingreso anual, age la edad en años, married es una variable binaria indicando si el individuo está casado y male una variable binaria indicando si el individuo es un hombre. El modelo (7) *se ha estimado con una muestra de 9275 individuos y los resultados están resumidos en la Tabla 2, junto con los de otros modelos relacionados.*

(a) *Interprete el coeficiente β_1 en* (7) *y proporcione un intervalo de confianza al 95% para el efecto sobre pp de un aumento de los ingresos del 10%.*

[50%] Si el ingreso aumenta en un 1%, la probabilidad de que el individuo subscriba un plan de pensiones aumenta en aproximadamente en $\beta_1/100$ (o en β_1 puntos porcentuales si expresamos la probabilidad en tanto por cien) todo lo demás igual.

[50%] Intervalo de confianza al 95%, usando los resultados de la columna (1)

$$\begin{aligned} 10\hat{\beta}_1/100 \pm 1.96 s.e.\left(10\hat{\beta}_1/100\right) &= \hat{\beta}_1/10 \pm 1.96 s.e.\left(\hat{\beta}_1\right)/10 \\ 0.229/10 \pm 1.96 * 0.00838/10 &: [0.0213, 0.0245] \end{aligned}$$

(b) *Determine si la probabilidad de subscribir un plan de pensiones de un hombre soltero es diferente a la de una mujer soltera usando el modelo* (7).

[50%] Comparando

$$\begin{aligned} \Pr(pp = 1|\ \text{hombre soltero}) &= \Pr(pp = 1|\ male = 1, married = 0, age, income) \\ &= \beta_0 + \beta_1 \log(income) + \beta_2 age + \beta_3 age^2 + \beta_4 \end{aligned}$$

y

$$\begin{aligned} \Pr(pp = 1|\ \text{mujer soltera}) &= \Pr(pp = 1|\ male = 0, married = 0, age, income) \\ &= \beta_0 + \beta_1 \log(income) + \beta_2 age + \beta_3 age^2 \end{aligned}$$

tomando los demás factores como fijos, el contraste buscado es

$$\begin{aligned} H_0 &: \beta_4 = 0 \\ H_1 &: \beta_4 \neq 0 \end{aligned}$$

que se realiza mediante un estadístico t usando los errores estándar robustos de los resultados de la columna (1)

$$t = \frac{\hat{\beta}_4}{e.e.\left(\hat{\beta}_4\right)} = \frac{-0.0151}{0.0143} = -1.0559$$

por lo que no podemos rechazar H_0 a los niveles habituales de significatividad: no se rechaza que las dos probabilidades son iguales.

¿Cuál es la diferencia estimada entre dicha probabilidad para un hombre y una mujer, ambos casados?

[50%] Las probabilidades tomando como fijos los demás factores son

$$\begin{aligned} \Pr(pp = 1|\ \text{hombre casado}) &= \Pr(pp = 1|\ male = 1, married = 1, age, income) \\ &= \beta_0 + \beta_1 \log(income) + \beta_2 age + \beta_3 age^2 + \beta_4 + \beta_5 + \beta_6 \end{aligned}$$

y

$$\begin{aligned}\Pr\left(pp=1|\text{ mujer casada}\right) &= \Pr\left(pp=1|\ male=0, married=1, age, income\right)\\ &= \beta_0+\beta_1\log\left(income\right)+\beta_2 age+\beta_3 age^2+\beta_5,\end{aligned}$$

por lo que usando las estimaciones de la columna (1)

$$\begin{aligned}\widehat{\Pr}\left(pp=1|\text{ hombre casado}\right)-\widehat{\Pr}\left(pp=1|\text{ mujer casada}\right) &= \hat{\beta}_4+\hat{\beta}_6\\ &= -0.0151-0.00875\\ &= -0.023\,85\end{aligned}$$

un hombre casado tiene una probabilidad 0.02385 menor de subscribir un plan que una mujer casada, todo lo demás igual.

(c) *Explique por qué con la información proporcionada no es posible llevar a cabo un contraste de hipótesis válido acerca de si la probabilidad de subscribir un plan de pensiones depende del género y estado civil del individuo.*

Porque aunque se podría calcular el valor del estadístico F sólo válido para homocedasticidad para contrastar

$$H_0: \beta_4=\beta_5=\beta_6=0$$

comparando los R^2 de las columnas (1) y (2), este contraste no es válido para el modelo lineal de probabilidad porque los errores de este modelo siempre son hetorocedásticos y necesitaríamos en su lugar un estadístico F robusto para hacer un contraste válido.

(d) *Finalmente, se decide estimar modelos diferentes para hombres y para mujeres incluyendo como variables explicativas* $\log\left(income\right), age, age^2$ *y married. Proporcione la ecuación estimada para los hombres usando los resultados de la columna (3).*

La probabilidad es

$$\begin{aligned}\widehat{\Pr}\left(pp=1|\text{ hombre}\right) &= \widehat{\Pr}\left(pp=1|\ male=1, age, married, income\right)\\ &= (-0.830+0.197)+(0.231-0.0112)\log\left(income\right)\\ &\quad +(0.0158-0.00761)\,age+(-0.000181+0.0000747)\,age^2\\ &\quad +(-0.0374-0.00145)\,married\\ &= -0.633\ +0.219\,8\log\left(income\right)\\ &\quad +0.008\,19age-1.063\times10^{-4}age^2-0.038\,85married\end{aligned}$$

3. *Suponga que el modelo verdadero que se desea estimar para explicar los resultados de un examen es*

$$Testscore_i = \beta_0 + \beta_1 STR_i + \beta_2 Effort_i + u_i$$

donde STR es una variable continua que aumenta cuando el tamaño de la clase aumenta. Effort es una variable continua que aumenta cuando el estudiante pone más esfuerzo. Finalmente, el error u incluye todos los otros factores que se supone están incorrelados con las otras variables explicativas del modelo. Suponga que $\beta_1 > 0$ *y* $\beta_2 > 0$, $Cov\,(STR, Effort) > 0$, $Cov(STR, u) = 0$, $Cov(Effort, u) = 0$.

Sin embargo, el económetra estima la siguiente especificación,

$$Testscore_i = \gamma_0 + \gamma_1 STR_i + e_i. \tag{8}$$

(a) *Discuta si la estimación MCO de* γ_1 *en* (8) *está sesgada o no para* β_1*, y si es así, cuál es el signo del sesgo.*

[50%] Se debe comprobar si se cumplen las dos condiciones para sesgo por Variables Omitidas (VO):

- $\text{Cov}(STR, Effort) \neq 0$ (en realidad > 0).

- $Cov\,(Effort, e) \neq 0$, que es cierta porque comparando la ecuación estimada (8) con la verdadera tenemos que

$$e_i = \beta_2 Effort_i + u_i$$

y podemos calcular $Cov\,(Effort, e) = \beta_2 Var\,(Effort) > 0$ porque $\beta_2 > 0$, y por tanto la estimación MCO de γ_1 estará sesgada para β_1.

También podemos comprobar que

$$\begin{aligned} \text{Cov}\,(e_i, STR_i) &= \text{Cov}\,(\beta_2 Effort_i + u_i, STR_i) \\ &= \beta_2 \text{Cov}\,(Effort_i, STR_i) + \text{Cov}\,(u_i, STR_i) \\ &= \beta_2 \text{Cov}\,(Effort_i, STR_i) > 0, \end{aligned}$$

por lo que no se cumple el supuesto 1 de MCO.

[50%] El sesgo de estimación es

$$Sesgo\,(\hat{\gamma}_1) = E\,(\hat{\gamma}_1 | \mathbf{X}) - \beta_1 \approx \frac{\text{Cov}\,(e_i, STR_i)}{\text{Var}\,(STR_i)} = \beta_2 \frac{\text{Cov}\,(Effort_i, STR_i)}{\text{Var}\,(STR_i)} > 0,$$

es decir $\hat{\gamma}_1$ sobreestima sistemáticamente el valor verdadero de β_1.

(b) *Imagine que en lugar de Effort se observa un indicador de la asistencia a clase,* CA_i*, para el que se sabe que* $Cov\,(CA, Effort) > 0$*, por lo que se cumple que*

$$Effort_i = \alpha_0 + \alpha_1 CA_i + v_i, \quad \text{Cov}\,(CA, v) = 0, \alpha_1 > 0.$$

También se sabe que

$$\text{Cov}\,(STR, v) = 0.$$

Interprete esta condición.

La incorrelación entre STR y v indica que la parte de *Effort* que no puede explicar CA, tampoco está correlada con STR, es decir que STR no es una variable omitida en la regresión de *Effort* sobre CA, y el error de reemplazar *Effort* por CA no estará relacionado con STR (lo que convierte a CA en una variable de control válida para reemplazar la variable omitida *Effort* y poder estimar el valor verdadero del coeficiente de STR).

(c) *¿Cuál sería el signo esperado de δ_3 en el siguiente modelo de regresión?*

$$Testscore_i = \delta_0 + \delta_1 STR_i + \delta_3 CA_i + w_i$$

[30%] En esta regresión el coeficiente δ_3 va a capturar la correlación del error $e_i = \beta_2 Effort_i + u_i$ con CA_i, por lo que esperaríamos que su signo sea el de $\text{Cov}(\beta_2 Effort_i, CA_i) = \beta_2 \text{Cov}(Effort_i, CA_i) > 0$, ya que en principio CA no está omitida el modelo verdadero y por tanto $\text{Cov}(CA, u) = 0$.

¿Cree que el estimador MCO de δ_1 sería sesgado o insesgado para β_1? ¿Por qué?

[30%] Usando el modelo verdadero y sustituyendo la ecuación de *Effort* en términos de CA,

$$\begin{aligned} Testscore_i &= \beta_0 + \beta_1 STR_i + \beta_2 Effort_i + u_i \\ &= \beta_0 + \beta_1 STR_i + \beta_2 (\alpha_0 + \alpha_1 CA_i + v_i) + u_i \\ &= \underbrace{(\beta_0 + \beta_2 \alpha_0)}_{=\delta_0} + \underbrace{\beta_1}_{=\delta_1} STR_i + \underbrace{\beta_2 \alpha_1}_{=\delta_3} CA_i + \underbrace{(\beta_2 v_i + u_i)}_{=w_i}, \end{aligned}$$

[40%] donde el error $w_i = \beta_2 v_i + u_i$ está incorrelado con STR_i (porque $\text{Cov}(STR, u) = 0$ y $\text{Cov}(STR, v) = 0$) y también esperamos que esté incorrelado con CA_i porque $\text{Cov}(CA, v) = 0$ y porque podemos suponer que $\text{Cov}(CA, u) = 0$ ya que CA no es una variable omitida en el modelo verdadero. En ese caso el modelo satisface los supuestos del modelo de regresión múltiple con la correspondiente definición de los coeficientes, y el estimador MCO de δ_1 estimaría consistentemente sin sesgo el coeficiente β_1 de STR en el modelo original, mientras que $\delta_3 = \beta_2 \alpha_1 = (+)(+) > 0$ como habíamos anticipado.