Solución Examen de Introducción a la Econometría

Universidad Carlos III de Madrid

2^a Convocatoria

Curso 2005/2006

Conteste las preguntas siguientes en cuadernillos separados en 2 horas y media

1. [2 puntos] Las desviaciones porcentuales sobre la tendencia de la renta agregada bruta disponible de las familias, Y_d, se descomponen en desviaciones del consumo agregado y del ahorro, $Y_d = C + S$. A partir de una muestra de tamaño $n = 100$ de datos de series temporales, se obtienen los siguientes momentos muestrales: $\hat{E}_n(C) = 1,\ \hat{E}_n(Y_d) = 0,75,\ \widehat{Cov}_n(Y_d, C) = 0,8,\ \widehat{Var}_n(Y_d) = 1,2,\ \widehat{Var}_n(S) = 2/3$.

a. [1 punto] Obtenga por el principio de analogía un estimador del predictor lineal óptimo del consumo agregado dada la renta disponible:

$$PLO\,[C|Y_d] = a_c + b_c Y_d$$

así como de la varianza del error de predicción, $\sigma_c^2 = Var(C - b_c Y_d)$.

b. [0,5 puntos] Considere ahora la predicción del ahorro dada la renta disponible:

$$PLO\,[S|Y_d] = a_s + b_s Y_d$$

Demuestre que $a_c + a_s = 0, b_c + b_s = 1$, y que $\sigma_c^2 = \sigma_s^2$.

c. [0,5 puntos] Si le dan un valor de renta disponible, ¿qué variable preferiría predecir, el consumo o el ahorro? Justifique su respuesta.

SOLUCIÓN: a.

$$\begin{aligned}
\hat{b}_c &= \frac{\widehat{Cov}_n\left(Y_d, C\right)}{\widehat{Var}_n\left(Y_d\right)} = \frac{0,8}{1,2} = \frac{2}{3} \\
\hat{a}_c &= \hat{E}_n(C) - \hat{b}_c \hat{E}_n(Y_d) = 1 - \frac{2}{3} * 0,75 = 0,5. \\
\hat{\sigma}_c^2 &= \widehat{Var}_n(C - b_c Y_d) = \widehat{Var}_n(C)\left(1 - R_c^2\right) \\
&= \frac{2}{3}\left(1 - \frac{\widehat{Cov}_n(C, Y_d)^2}{\widehat{Var}_n(Y_d)}\right) = \frac{2}{3}\left(1 - \frac{0,8^2}{1,2}\right) = 0,311. \\
R_c^2 &= \frac{\widehat{Cov}_n(C, Y_d)^2}{\widehat{Var}_n(C)\widehat{Var}_n(Y_d)}.
\end{aligned}$$

$\frac{2}{3}\left(1 - \frac{0.8^2}{1.2}\right) = 0.311\,11$

b. Tenemos que si sumamos los modelos

$$\begin{aligned}
C &= a_c + b_c Y_d + u_c \\
S &= a_s + b_s Y_d + u_s
\end{aligned}$$

obtenemos

$$C + S = \left(a_c + a_s\right) + \left(b_c + b_s\right) Y_d + u_c + u_s$$

por lo que, usando que $Y_d = C + S$, se deduce que $a_c + a_s = 0,\ b_c + b_s = 1$, $u_c = -u_s$ y por tanto $\sigma_c^2 = \sigma_s^2$.

c. Habría que comparar, además del error absoluto, $\sigma_c^2 = \sigma_s^2$, el relativo de predicción, los R^2 :

$$
\begin{aligned}
R_c^2 &= \frac{\widehat{Cov}_n(C, Y_d)^2}{\widehat{Var}_n(C)\widehat{Var}_n(Y_d)} = \frac{(0,8)^2}{1,0667 * 1,2} = 0,5 \\
R_s^2 &= \frac{\widehat{Cov}_n(S, Y_d)^2}{\widehat{Var}_n(S)\widehat{Var}_n(Y_d)} = \frac{(0,4)^2}{\frac{2}{3} * 1,2} = 0.2 \\
\widehat{Cov}_n(S, Y_d) &= \widehat{Var}_n(Y_d) - \widehat{Cov}_n(C, Y_d) \\
&= 1,2 - 0,8 = 0,4. \\
\widehat{Var}_n(C) &= \widehat{Var}_n(Y_d) + \widehat{Var}_n(S) - 2 * \widehat{Cov}_n(S, Y_d) \\
&= 1,2 + \frac{2}{3} - 2 * 0,4 = 1,0667
\end{aligned}
$$

por lo que es *más* fácil predecir C que S, ya que C tiene más variabilidad y el error relativo de predicción es menor aunque el error absoluto, σ^2, sea igual de grande.

2. **[2 puntos]** Se quiere explicar la nota final obtenida en la selectividad, *selec*, en función de las horas semanales medias de estudio, *horas* y de la asignación a un programa de especial de ayuda al estudio, por medio de un modelo de regresión lineal

$$selec = \beta_0 + \beta_1 horas + \beta_2 prog + u \tag{1}$$

donde *prog* es una variable dummy de tal forma que $prog_i = 1$ si el estudiante i es asignado al programa especial de ayuda al estudio.

a. **[0,5 puntos]** Interprete los coeficientes del modelo (1). ¿Cuál sería la relación entre los coeficientes del modelo (1) y los del modelo lineal basado solamente en *horas*,

$$selec = \delta_0 + \delta_1 horas + v, \tag{2}$$

si se supone que la asignación al programa se ha efectuado aleatoriamente?

b. **[0,5 puntos]** Si en en cambio, se propusiese la asignación al programa favoreciendo a aquellos estudiantes con menor hábito de estudio, ¿cuál sería la relación esperada entre β_1 y δ_1?

c. **[0,5 puntos]** Explica como contrastarías la efectividad del programa de ayuda al estudio dada una muestra. Si sólo hay recursos para incluir a un 10% de los alumnos en el estudio, ¿qué política esperas que sea más efectiva para contrastar la efectividad del programa de ayuda al estudio, la asignación propuesta en (a) o la propuesta en (b)?

d. **[0,5 puntos]** Se quiere modificar el modelo (1) para permitir que la efectividad de tiempo de estudio aumente con la inclusión en el programa. Proponga el modelo adecuado para ello e indique como contrastaría en dicho caso la existencia de un beneficio por la asignación al programa para los estudiantes que dedican 20 horas semanales a estudiar en media.

SOLUCIÓN: a. β_1 indica el incremento medio en la nota de selectividad si se aumenta en una hora el tiempo de estudio a la semana, ceteris paribus, lo que implica no cambiar la asignación o no al programa de ayuda al estudio. β_0 indica la nota media de un estudiante que no dedica ningún tiempo al estudio y no ha sido asignado al programa de ayuda.

Sabemos que en general

$$\delta_1 = \beta_1 + \beta_2 \frac{Cov\,(horas, prog)}{Var\,(horas)}, \tag{3}$$

pero si la asignación al programa de ayuda se efectúa aleatoriamente, entonces *prog* será independiente de *horas* (y de cualquier otra característica del estudiante), por lo que $Cov\,(horas, prog) = 0$ y $\delta_1 = \beta_1$.

b. Podemos esperar que $\beta_1 > 0$ y que también $\beta_2 > 0$, con lo que dada la relación (3) y que la asignación al programa preferencial para estudiantes con *horas* por debajo de su media implica que $Cov\,(horas, prog) < 0$, tendremos que $\delta_1 < \beta_1$, es decir el modelo corto tiene un sesgo negativo e infraestima el efecto del tiempo de estudio sobre la nota de selectividad.

c. Después de la estimación de MCO, la hipótesis nula a contrastar sería $H_0 : \beta_2 = 0$ en contra de la alternativa $H_1 : \beta_2 > 0$, mediante un contraste de la t, rechazando H_0 en favor de H_1 si t es grande y positivo.

La efectividad del contraste va a ser mayor si podemos estimar con más precisión β_2. Si incluimos siempre al 10% de los estudiantes, *prog* es una variable Bernoulli con $p = 0,1$ y por tanto $Var\,(prog) = p\,(1-p) = 0,1 * 0,9 = 0,09$, y su correspondiente suma de cuadrados, SST_{prog}, sería en ambos casos igual. En la situación de (a) tenemos que la varianza estimada del estimador $\hat{\beta}_2$ es

$$\widehat{Var}\left(\hat{\beta}_2\right) = \frac{\hat{\sigma}^2}{n}\frac{1}{SST_{prog}},$$

porque da igual hacer la regresión corta que la larga, mientras que en (b) tenemos que

$$\widehat{Var}\left(\hat{\beta}_2\right) = \frac{\hat{\sigma}^2}{n}\frac{1}{SST_{prog}\left(1 - R^2_{horas,prog}\right)}$$

donde $R^2_{horas,prog}$ es el coeficiente R^2 de la regresión de *prog* sobre *horas*, que en este caso no es cero. Por tanto, si la asignación se hace como en (b), el estimador tendrá mayor varianza y el contraste será menos eficiente.

d. En este caso hay que incluir una interacción:

$$selec = \beta_0 + \beta_1 horas + \beta_2 prog + \beta_3 horas * prog + \varepsilon,$$

donde el parámetro β_3 recoge el retorno adicional medio de una hora extra de estudio para aquellos estudiantes que participan en el programa especial (con respecto a los que no participan). Los estudiantes que dedican 20 horas semanales por término medio tienen un beneficio esperado de ser incluidos en el programa (todo lo demás constante) de $\beta_2 + \beta_3 * 20$. Por tanto la hipótesis nula sería $H_0 : \theta = 0$, donde $\theta = \beta_2 + \beta_3 * 20$, y la alternativa sería $H_1 : \theta > 0$. Para efectuar el contraste habría que hacer un contraste unilateral de la t, donde para estimar θ podríamos usar la estimación de β_2 y β_3 y sus errores estándar y covarianza para estimar el s.e. de $\hat{\theta}$. Alternativamente podríamos estimar directamente θ mediante la ecuación

$$selec = \beta_0 + \beta_1 horas + \theta prog + \beta_3\,(horas - 20) * prog + \varepsilon$$

al sustituir $\beta_2 = \theta - 20\beta_3$.

3. **[3 puntos]** Se dispone de una muestra de 550 empresas textiles y del sector automovilístico con información sobre el valor de su producción Y (en millones de euros), las unidades de trabajo empleadas N (en miles de horas trabajadas) y las unidades de capital utilizado en la producción K. También se tiene una variable, *textil*, que toma valor 1 si la empresa pertenece al sector textil y cero si pertenece al sector automovilístico. Se supone que la función de producción de las empresas es una Cobb-Douglas (es decir, $Y_i = AN_i^{\beta_1}K_i^{\beta_2}e^{U_i}$); aplicando logaritmos se obtiene una versión lineal de esta función de producción $\ln Y_i = \ln A + \beta_1 \ln N_i + \beta_2 \ln K_i + U_i$. Por tanto, se han estimado los siguientes modelos por Mínimos Cuadrados Ordinarios:

$$\widehat{\ln Y_i} = \underset{(0{,}6601)}{-0,5241} \underset{(0{,}1391)}{+0,7366} \ln N_i \underset{(0{,}0731)}{+0,4453} \ln K_i \tag{4}$$
$$n = 550 \quad R^2 = 0,1225 \quad SSR = 541,4375$$

$$\widehat{\ln Y_i} = \underset{(0{,}9505)}{0,6514} \underset{(1{,}3059)}{-1,9003}\ textil \underset{(0{,}2003)}{+0,5334} \ln N_i \underset{(0{,}0950)}{+0,3321} \ln K_i \underset{(0{,}2747)}{+0,3904}\ textil * \ln N_i \underset{(0{,}1492)}{+0,1214}\ textil * \ln K_i \tag{5}$$
$$n = 550 \quad R^2 = 0,1565 \quad SSR = 520,4230$$

$$\widehat{\ln(Y_i/N_i)} = \underset{(0,0835)}{0,0518} + \underset{(0,1097)}{0,3846}\, textil + \underset{(0,0893)}{0,3530}\, \ln(K_i/N_i) + \underset{(0,1387)}{0,0164}\, textil * \ln(K_i/N_i) \tag{6}$$

$$n = 550 \quad R^2 = 0,0949 \quad SSR = 524,2166$$

$$\widehat{\ln(Y_i/N_i)} = \underset{(0,0547)}{0,2884} + \underset{(0,0683)}{0,4131}\, \ln(K_i/N_i) \tag{7}$$

$$n = 550 \quad R^2 = 0,0625 \quad SSR = 542,9477$$

$$\widehat{\ln Y_i} = \underset{(0,5205)}{2,7303} - \underset{(0,9312)}{4,3011}\, textil + \underset{(0,0770)}{0,1854}\, \ln(K_i/N_i) + \underset{(0,2525)}{0,6809}\, textil * \ln N_i + \underset{(0,1089)}{0,3889}\, (textil * \ln K_i + \ln N_i) \tag{8}$$

$$n = 550 \quad R^2 = 0,1460 \quad SSR = 526,9304$$

a. **[0,75 puntos]** Suponga inicialmente que creemos que los parámetros de la función de producción son iguales en ambos sectores industriales. Contraste con un nivel de significación del 5% la hipótesis de que existen rendimientos constantes a escala.

b. **[0,75 puntos]** Ahora suponga por contra que creemos que existen rendimientos constantes a escala. Contraste con un nivel de significación del 5% la hipótesis de que las funciones de producción son iguales en ambos sectores.

c. **[1 punto]** Contraste con un nivel de significación del 5% la hipótesis conjunta de que en ambos sectores existen rendimientos constantes a escala (aunque los parámetros de las funciones de producción son diferentes).

d. **[0,5 puntos]** Un experto industrial afirma que si una empresa del sector automovilístico aumenta las unidades de capital utilizadas en un 2% y además las unidades de trabajo empleadas en un 1% entonces el cambio (porcentual) en su producción será, en promedio, el mismo que el de una empresa textil que aumenta las unidades de capital utilizadas en un 1% y mantiene el trabajo constante. Contraste con un nivel de significación del 5% si existe evidencia empírica a favor de la aseveración del experto.

SOLUCIÓN: **a.** La función de producción es igual en ambos sectores, $\ln Y_i = \beta_0 + \beta_1 \ln N_i + \beta_2 \ln K_i + U_i$, donde los parámetros β_0, β_1 y β_2 son iguales. La hipótesis de rendimientos constantes a escala se expresa en función de los parámetros como

$$H_0 : \beta_1 + \beta_2 = 1 \quad vs. \quad H_1 : \beta_1 + \beta_2 \neq 1$$

Imponiendo esta restricción, se tiene un modelo restringido $\ln Y_i = \beta_0 + (1 - \beta_2) \ln N_i + \beta_2 \ln K_i + U_i$, que reescribiendo y utilizando las propiedades de los logaritmos queda: $\ln(Y_i/N_i) = \beta_0 + \beta_2 \ln(K_i/N_i) + U_i$. Por tanto, el modelo sin restringir en este contexto se correspondería con la estimación (4) y el modelo restringido con (7). El estadístico de contraste es

$$W = qF = \frac{SSR_R - SSR_{SR}}{(SSR_{SR})/n} \underset{H_0}{\overset{a}{\sim}} \chi^2_{(q)}$$

En este caso $q = 1$ y el estadístico tiene valor $W = \frac{542,9477 - 541,4375}{541,4375/550} = 1,534\,1$. El valor crítico de una $\chi^2_{(1)}$ al 5% es $c^*_{0,05} = 3,84$. Puesto que $W < c^*_{0,05}$ no evidencia empírica para rechazar la hipótesis nula al 5%.

b. En este caso suponemos que existen rendimientos constantes a escala. Si se permite que los parámetros sean diferentes entre sectores, el modelo a estimar sería: $\ln(Y_i/N_i) = \beta_0 + \delta_0 textil +$

$\beta_2 \ln(K_i/N_i) + \delta_2 textil * \ln(K_i/N_i) + U_i$, donde δ_0 y δ_2 representan el diferencial en la constante y en la pendiente, respectivamente, entre el sector textil y el automovilístico. En el modelo restringido, no existe ninguna diferencia entre sectores, es decir,

$$H_0 : \delta_0 = \delta_2 = 0 \quad vs. \quad H_1 : H_0 \text{ no es cierta}$$

El modelo sin restringir en este contexto se correspondería con la estimación (6) y el modelo restringido con (7). En este caso, $q = 2$ y el estadístico de contraste se puede calcular como

$$\begin{aligned} W = qF &= \frac{SSR_R - SSR_{SR}}{(SSR_{SR})/n} = \frac{R_{SR}^2 - R_R^2}{(1 - R_{SR}^2)/n} \\ &= \frac{542,9477 - 524,2166}{(524,2166)/550} \simeq \frac{0,0949 - 0,0625}{(1 - 0,0949)/550} \simeq 19,7 \end{aligned}$$

El valor crítico para una $\chi^2_{(2)}$ al 5% es $c^*_{0,05} = 5,99$. Por tanto, existe evidencia empírica para rechazar la hipótesis nula al 5%. ($W > c^*_{0,05}$)

c. El modelo más general (los parámetros pueden ser diferentes entre sectores y no se impone rendimientos constates) sería

$$\ln(Y_i) = \beta_0 + \delta_0 textil + \beta_1 \ln(N_i) + \delta_1 textil * \ln(N_i) + \beta_2 \ln(K_i) + \delta_2 textil * \ln(K_i) + U_i$$

Esto se corresponde con el modelo (5). Si existen rendimientos constates a escala en el sector automovilístico $\beta_1 + \beta_2 = 1$; mientras que que si existen rendimientos constantes a escala en el sector automovilístico: $(\beta_1 + \delta_1) + (\beta_2 + \delta_2) = 1$. Nótese que si ambas condiciones se tienen que dar **conjuntamente**, la hipótesis nula será

$$H_0 : \begin{array}{c} \beta_1 + \beta_2 = 1 \\ \delta_1 + \delta_2 = 0 \end{array} \quad \text{vs.} \quad H_1 : H_0 \text{ no es cierta}$$

El modelo restringido que ha incorporado estas restricciones es (6). Por tanto, el estadístico de contraste sería:

$$W = qF = \frac{SSR_R - SSR_{SR}}{(SSR_{SR})/n} = \frac{524,2166 - 520,4230}{(520,4230)/550} \simeq 4,0092$$

Puesto que $q = 2$, tenemos que compararlo con el valor crítico al 5% de una $\chi^2_{(2)} : c^*_{0,05} = 5,99$. Por tanto, no se puede rechazar la hipótesis nula ($W < c^*_{0,05}$).

d. Utilizando el modelo más general, los cambios en la producción ante cambios en los factores serían:

$$\begin{aligned} \text{en el sector automovilístico,} \quad & \Delta \ln(Y_i) = \beta_1 \Delta \ln(N_i) + \beta_2 \Delta \ln(K_i) \\ \text{en el sector textil,} \quad & \Delta \ln(Y_i) = (\beta_1 + \delta_1)\Delta \ln(N_i) + (\beta_2 + \delta_2)\Delta \ln(K_i) \end{aligned}$$

Por tanto, en el sector automovilístico aumenta las unidades de capital utilizadas en un 2% y además las unidades de trabajo empleadas en un 1% aumentaría la producción en $\Delta \ln(Y_i) = \beta_1 * 1 + \beta_2 * 2$; mientras que aumentar el capital en un 1% en una empresa textil supone un cambio en la producción de $\Delta \ln(Y_i) = (\beta_2 + \delta_2) * 1$. El enunciado dice que estos dos cambios son iguales, luego la hipótesis a contrastar es

$$H_0 : \beta_1 + 2\beta_2 = \beta_2 + \delta_2 \quad vs. \quad H_1 : \beta_1 + 2\beta_2 \neq \beta_2 + \delta_2$$

Si imponemos esta restricción (utilizando que $\beta_1 = \delta_2 - \beta_2$), se tendría el modelo (8). Por tanto, el estadístico de contraste sería:

$$W = qF = \frac{SSR_R - SSR_{SR}}{(SSR_{SR})/n} = \frac{526,9304 - 520,4230}{(520,4230)/550} \simeq 6,877\,2$$

Puesto que $q = 1$, tenemos que compararlo con el valor crítico al 5% de una $\chi^2_{(1)} : c^*_{0,05} = 3,84$. Por tanto, hay evidencia empírica para rechazar la hipótesis nula al 5% ($W > c^*_{0,05}$).

4. [3 puntos] Una empresa multinacional ha sido acusada de discriminar a sus empleadas pagándoles salarios menores que a los empleados. El presidente solicita una muestra aleatoria simple de 250 empleados y empleadas, todos/as con la misma experiencia y años de educación. En la muestra, se disponen de las siguientes variables:

$$salario_i = \text{salario mensual del empleado/a } i \text{ (en euros)}$$

$$mujer_i = \begin{cases} 1 \text{ si el individuo } i \text{ es mujer} \\ 0 \text{ en caso contrario} \end{cases}$$

$$directiv_i = \begin{cases} 1 \text{ si el individuo } i \text{ tiene un puesto de trabajo directivo (toma de decisiones)} \\ 0 \text{ en caso contrario} \end{cases}$$

A partir de esa muestra, el presidente obtiene la siguiente tabla con medias de los salarios para hombres y mujeres (y errores estándar de las respectivas medias entre paréntesis):

Tabla 1

Hombres	Mujeres
$1915,57$	$2132,93$
(49,53)	(79,91)

Algunos trabajadores, que estudiaron y comprendieron los conceptos básicos de la Econometría, no están satisfechos con esta evidencia. Solicitan la misma muestra y, a partir de ella, estiman varios modelos econométricos por Mínimos Cuadrados Ordinarios; como el presidente no los entiende muy bien, confeccionan la siguiente la tabla con medias de salarios (y errores estándar de las respectivas medias entre paréntesis):

Tabla 2

	Hombres	Mujeres
En trabajos directivos	$2994,43$	$2497,97$
	(9,22)	(9,71)
En trabajos no directivos	$1501,87$	$992,16$
	(6,34)	(18,98)

a. [0,75 puntos] Formule un modelo econométrico (incluyendo las variables disponibles en la encuesta y/o variables construidas a partir de ellas) tal que uno de los parámetros pueda interpretarse como la diferencia en el salario promedio entre hombres y mujeres. Utilizando los datos de la Tabla 1, ofrezca valores para la estimación de todos los parámetros en este modelo.

b. [0,75 puntos] Formule otro modelo econométrico (incluyendo las variables disponibles en la encuesta y/o variables construidas a partir de ellas) tal que uno de los parámetros pueda interpretarse como la diferencia en el salario promedio entre hombres en puestos directivos y mujeres en puestos directivos. Utilizando los datos de la Tabla 2, ofrezca valores para la estimación de todos los parámetros en este nuevo modelo.

c. [0,5 puntos] A partir del modelo formulado en el apartado (a) y de los datos de la Tabla 1, contraste con un nivel de significación del 5% la hipótesis de que los hombres y las mujeres tienen el mismo salario.
[Pista: puesto que se utiliza una muestra aleatoria simple, la covarianza (condicional en educación y experiencia) entre el salario medio de un hombre y el salario medio de una mujer es, *ceteris paribus*, cero. Es decir, si $\overline{y}^h$ e $\overline{y}^m$ son el salario medio, *ceteris paribus,* de un hombre y de una mujer respectivamente, entonces $Cov\left(\overline{y}^h, \overline{y}^m\right) = 0$ (donde $Cov(\cdot)$ debe entenderse como covarianza condicional en educación y experiencia)]

d. [0,5 puntos] A partir del modelo formulado en el apartado (b) y de los datos de la Tabla 2, contraste con un nivel de significación del 5% la hipótesis de que los hombres en puestos directivos y las mujeres en puestos directivos tienen el mismo salario.
[Pista: igual que en el apartado anterior, si $\overline{y}^{hd}$, $\overline{y}^{md}$, $\overline{y}^{hn}$ e $\overline{y}^{mn}$ son los salarios medio de hombre directivos, mujeres directivas, hombres no directivos y mujeres no directivas, respectivamente, entonces $Cov\left(\overline{y}^{md}, \overline{y}^{mn}\right) = Cov\left(\overline{y}^{md}, \overline{y}^{hd}\right) = Cov\left(\overline{y}^{md}, \overline{y}^{hn}\right) = Cov\left(\overline{y}^{mn}, \overline{y}^{hd}\right) = Cov\left(\overline{y}^{mn}, \overline{y}^{hn}\right) = Cov\left(\overline{y}^{hd}, \overline{y}^{hn}\right) = 0$]

e. **[0,5 puntos]** Contraste con un nivel de significación del 5% la hipótesis de que los diferenciales salariales entre hombres y mujeres no cambian según el tipo de trabajo; es decir, que la diferencia en el salario promedio entre hombres directivos y mujeres directivas es la misma que la diferencia en el salario promedio entre hombres no directivos y mujeres no directivas.

SOLUCIÓN: **a.** El modelo econométrico sería

$$salario_i = \delta_0 + \delta_1 mujer_i + u_i,$$

donde δ_0 sería el salario esperado de los hombres y $(\delta_0 + \delta_1)$ el salario esperado de las mujeres; por tanto, δ_1 es la diferencia entre el salario esperado entre hombres y mujeres.
A partir de los datos de la Tabla 1, podemos concluir que $\overline{y}^h = \widehat{\delta}_0 = 1915,57$ y, como $\overline{y}^m = \widehat{\delta}_0 + \widehat{\delta}_1 = 2132,93$, la estimación del parámetro δ_1 será

$$\widehat{\delta}_1 = 2132,93 - 1915,57 = 217,\ 36$$

b. Construyamos las variables dummies $md_i = mujer_i * directiv_i$, $mnd_i = mujer_i * (1 - directiv_i)$, $hd_i = (1 - mujer_i) * directiv_i$, $hnd_i = (1 - mujer_i) * (1 - directiv_i)$. El modelo econométrico podría ser

$$salario_i = \beta_0 + \beta_1 md_i + \beta_2 mnd_i + \beta_3 hnd_i + \varepsilon_i$$

(aunque también podría haberse omitido la dummy de md_i en lugar de hd_i, obteniéndose también que el parámetro asociado a hd_i tiene la interpretación del diferencial requerido pero con el signo contrario). Utilizando los resultados de la Tabla 2 es fácil comprobar que los salarios medios estimados son:

$$\begin{aligned}
\widehat{E_n}\left(salario_i|hombre, directivo\right) &= \overline{y}^{hd} = \widehat{\beta}_0 = 2994,43 \\
\widehat{E_n}\left(salario_i|mujer, directivo\right) &= \overline{y}^{md} = \widehat{\beta}_0 + \widehat{\beta}_1 = 2497,97 \\
\widehat{E_n}\left(salario_i|hombre, no\ directivo\right) &= \overline{y}^{hn} = \widehat{\beta}_0 + \widehat{\beta}_3 = 1501,87 \\
\widehat{E_n}\left(salario_i|mujer, no\ directivo\right) &= \overline{y}^{mn} = \widehat{\beta}_0 + \widehat{\beta}_2 = 992,16
\end{aligned}$$

de donde es fácil despejar $\widehat{\beta}_1 = -496,46$; $\widehat{\beta}_2 = -2002,27$; $\widehat{\beta}_3 = -1492,56$

c. La hipótesis nula es $H_0 : \delta_1 = 0$ y el estadístico de este contraste será

$$t = \frac{\widehat{\delta}_1}{se\left(\widehat{\delta}_1\right)} \underset{H_0}{\overset{a}{\sim}} N(0,1)$$

para lo cual necesitamos calcular $se\left(\widehat{\delta}_1\right)$; hay que notar que $\widehat{V}_n\left(\widehat{\delta}_1\right) = \widehat{V}_n\left(\overline{y}^m - \overline{y}^h\right) = \widehat{V}_n\left(\overline{y}^m\right) + \widehat{V}_n\left(\overline{y}^h\right) - 2\widehat{C}_n\left(\overline{y}^m, \overline{y}^h\right)$. A partir de la Tabla 1 y de la pista podemos obtener el valor de estas varianzas y covarianza muestrales: $\widehat{V}_n\left(\widehat{\delta}_1\right) = (79,91)^2 + (49,53)^2 - 0 = 8838,8$. Por tanto,

$$t = \frac{217,\ 36}{\sqrt{8838,8}} = 2,3120$$

Para una hipótesis alternativa bilateral, el valor crítico al 5% es $c^*_{0,05} = 1,96$; puesto que $|t| > c^*_{0,05}$ se rechaza la hipótesis nula de que hombres y mujeres tengan el mismo salario. De hecho, la evidencia empírica indicaría que *las mujeres ganan más*.

d. La hipótesis nula es $H_0 : \beta_1 = 0$ y el estadístico de este contraste será $t = \frac{\widehat{\beta}_1}{se\left(\widehat{\beta}_1\right)} \underset{H_0}{\overset{a}{\sim}} N(0,1)$. Nuevamente para calcular $se\left(\widehat{\beta}_1\right)$, se utiliza que $\widehat{V}_n\left(\widehat{\beta}_1\right) = \widehat{V}_n\left(\overline{y}^{md} - \overline{y}^{hd}\right) = \widehat{V}_n\left(\overline{y}^{md}\right) + \widehat{V}_n\left(\overline{y}^{hd}\right) - 2\widehat{C}_n\left(\overline{y}^{md}, \overline{y}^{hd}\right)$. A partir de la Tabla 2 y de la pista podemos obtener el valor de estas varianzas y covarianza muestrales: $\widehat{V}_n\left(\widehat{\beta}_1\right) = (9,71)^2 + (9,22)^2 - 0 = 179,29$. Por tanto,

$$t = \frac{-496,46}{\sqrt{179,29}} = -37,077$$

Existe una fuerte evidencia empírica para rechazar la hipótesis nula de que hombres directivos y mujeres directivas ganan lo mismo $|t| > c^*_{0,05}$. En este caso, la evidencia empírica indicaría que las mujeres directivas *ganan menos* una vez que se ha controlado por el tipo de puesto de trabajo.

e. El diferencial entre hombres directivos y mujeres directivas es

$$\beta_1 = E\left(salario_i | md_i = 1, hnd_i = mnd_i = 0, hd_i = 0\right) - E\left(salario_i | md_i = hnd_i = mnd_i = 0, hd_i = 1\right)$$

mientras que el diferencial entre hombres no directivos y mujeres no directivas es

$$\begin{aligned} & E\left(salario_i | md_i = 0, mnd_i = 1, hnd_i = 0, hd_i = 0\right) - E\left(salario_i | md_i = 0, mnd_i = 0, hnd_i = 1, hd_i = 0\right) \\ = \quad & (\beta_0 + \beta_2) - (\beta_0 + \beta_3) = \beta_2 - \beta_3. \end{aligned}$$

El enunciado establece que ambos diferenciales son iguales, es decir, la hipótesis a contrastar y el correspondiente estadístico de contraste son:

$$H_0 : \beta_1 = \beta_2 - \beta_3 \quad vs. \quad H_1 : \beta_1 \neq \beta_2 - \beta_3$$

$$t = \frac{\widehat{\beta}_1 - \widehat{\beta}_2 + \widehat{\beta}_3}{se\left(\widehat{\beta}_1 - \widehat{\beta}_2 + \widehat{\beta}_3\right)} \underset{H_0}{\overset{a}{\sim}} N(0,1)$$

Para el cálculo del error estándar utilizamos que $\widehat{V}_n\left(\widehat{\beta}_1 - \widehat{\beta}_1 + \widehat{\beta}_3\right) = \widehat{V}_n\left[\left(\overline{y}^{md} - \overline{y}^{hd}\right) - \left(\overline{y}^{mn} - \overline{y}^{hn}\right)\right]$; puesto que todas las covarianzas son cero $\widehat{V}_n\left(\widehat{\beta}_1 - \widehat{\beta}_1 + \widehat{\beta}_3\right) = \widehat{V}_n\left(\overline{y}^{md}\right) + \widehat{V}_n\left(\overline{y}^{hd}\right) + \widehat{V}_n\left(\overline{y}^{mn}\right) + \widehat{V}_n\left(\overline{y}^{hn}\right) = (9,71)^2 + (9,22)^2 + (18,98)^2 + (6,34)^2 = 579,73$. El valor del estadístico es

$$t = \frac{-496,46 - (-2002,27) + (-1492,56)}{\sqrt{579,73}} = 0,550\,3$$

En conclusión, puesto que $|t| = 0,5503 < c^*_{0,05} = 1,96$ no existe evidencia empírica para rechazar la hipótesis nula.

VALORES CRÍTICOS:

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

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

Recordamos que una t de Student con n grados de libertad se comporta como un $N(0,1)$ para n razonablemente grande ($n > 30$). 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 $\chi^2_{(q)}/q$.