TEMA 4: *Variables binarias*

Econometría I

M. Angeles Carnero

Departamento de Fundamentos del Análisis Económico

Curso 2011-12

Variables binarias

Las variables que hemos estado considerando hasta este momento eran variables cuantitativas (salarios, renta, experiencia laboral, años de educación, etc.). En este tema vamos a ver cómo se pueden incorporar factores cualitativos en los modelos de regresión. El sexo o la raza de un individuo, la región en donde vive, el sector industrial al que pertenece una empresa, etc. son factores cualitativos que aparecen con frecuencia en los modelos empíricos.
Los factores cualitativos aparecen a menudo bajo la forma de información binaria: Un individuo es hombre o mujer, un individuo ha participado o no en un programa de formación profesional, una empresa ofrece o no un plan de pensiones a sus trabajadores, etc. Estas variable cualitativas se representan mediante variables binarias que toman los valores cero y uno y se denominan variables ficticias o, utilizando el término en inglés, variables dummy.

Modelos con un único factor cualitativo:

- Vamos a considerar, por ejemplo, las variables ficticias o variables dummy de sexo:

$$
\begin{aligned}
Hombre_t &= \begin{cases} 1 \text{ si el individuo t es hombre} \\ 0 \text{ si el individuo t es mujer} \end{cases} \\
Mujer_t &= \begin{cases} 1 \text{ si el individuo t es mujer} \\ 0 \text{ si el individuo t es hombre} \end{cases}
\end{aligned}
$$

- y consideremos el modelo para el salario por hora en función de la experiencia laboral que vimos en el Tema 1, en el que ahora incorporamos la dummy de ser hombre.

$$salario_t = \beta_1 + \beta_2 Hombre_t + \beta_3 exp_t + u_t \qquad (1)$$

- Si calculamos en este modelo la media del salario para hombres y mujeres que tengan los mismos años de experiencia laboral tenemos

$$\begin{aligned} E(salario_t/Hombre_t = 1) &= \beta_1 + \beta_2 + \beta_3 exp \\ E(salario_t/Hombre_t = 0) &= \beta_1 + \beta_3 exp \end{aligned} \qquad (1b)$$

- Al incorporar la variable ficticia $Hombre_t$ lo que estamos haciendo es permitir que el término independiente del modelo pueda ser distinto para hombres y mujeres, ya que para hombres el término constante es $\beta_1 + \beta_2$ mientras que para las mujeres es β_1, y por tanto β_2 refleja las posibles diferencias en el término constante entre hombres y mujeres.
- Además,

$$\beta_2 = E(salario_t/t \text{ es Hombre}) - E(salario_t/t \text{ es Mujer})$$

y por tanto β_2 mide la diferencia en el salario medio entre hombres y mujeres con la misma experiencia laboral. La hipótesis de ausencia de diferencias entre hombres y mujeres sería $\beta_2 = 0$.

- Alternativamente podríamos haber incorporado en el modelo la dummy de ser mujer:

$$salario_t = \alpha_1 + \alpha_2 Mujer_t + \alpha_3 exp_t + u_t \qquad (2)$$

Si calculamos ahora la media del salario para hombres y mujeres que tienen los mismos años de experiencia laboral tenemos

$$\begin{aligned} E(salario_t / Mujer_t = 0) &= \alpha_1 + \alpha_3 exp \\ E(salario_t / Mujer_t = 1) &= \alpha_1 + \alpha_2 + \alpha_3 exp \end{aligned} \qquad (2b)$$

y por tanto

$$\alpha_2 = E(salario_t / t \text{ es mujer}) - E(salario_t / t \text{ es hombre})$$

es decir α_2 mide la diferencia en el salario medio entre mujeres y hombres con la misma experiencia laboral.

- Los modelos

$$salario_t = \beta_1 + \beta_2 Hombre_t + \beta_3 exp_t + u_t \qquad (1)$$

$$salario_t = \alpha_1 + \alpha_2 Mujer_t + \alpha_3 exp_t + u_t \qquad (2)$$

son equivalentes y tendremos que

$$\left.\begin{array}{l} \beta_1 + \beta_2 + \beta_3 exp = \alpha_1 + \alpha_3 exp \\ \beta_1 + \beta_3 exp = \alpha_1 + \alpha_2 + \alpha_3 exp \end{array}\right\} \Rightarrow \left\{\begin{array}{l} \beta_3 = \alpha_3 \\ \beta_1 + \beta_2 = \alpha_1 \\ \beta_1 = \alpha_1 + \alpha_2 \end{array}\right.$$

Por lo tanto, obviamente $\alpha_2 = -\beta_2$. Esta relación entre los parámetros de los modelos (1) y (2) también se verifica para los estimadores MCO de los dos modelos como ilustra el siguiente ejemplo

Ejemplo 1:

Utilizando una muestra para Estados Unidos de 526 individuos se han estimado por MCO los modelos (1) y (2) obteniéndose los siguientes resultados

$$\begin{aligned}
\widehat{salario}_t &= \underset{(0{,}2845)}{4{,}145} + \underset{(0{,}3022)}{2{,}481}\, Hombre_t + \underset{(0{,}0111)}{0{,}0269} exp_t \\
\widehat{salario}_t &= \underset{(0{,}2862)}{6{,}626} - \underset{(0{,}3022)}{2{,}481}\, Mujer_t + \underset{(0{,}0111)}{0{,}0269} exp_t
\end{aligned}$$

donde *salario* es el salario por hora en dólares y *exp* es la experiencia potencial en años (edad menos años de educación menos 6).

- Podemos comprobar que efectivamente

$$\begin{aligned}
&\widehat{\beta}_3 = \widehat{\alpha}_3 = 0{,}0269 \\
&\widehat{\beta}_1 + \widehat{\beta}_2 = 4{,}145 + 2{,}481 = 6{,}626 = \widehat{\alpha}_1 \\
&\widehat{\alpha}_1 + \widehat{\alpha}_2 = 6{,}626 - 2{,}481 = 4{,}145 = \widehat{\beta}_1
\end{aligned}$$

  Según este modelo el salario por hora de los hombres es en media 2,481 dólares mayor que el de las mujeres con la misma experiencia.

- Si queremos contrastar la hipótesis nula de que no hay diferencias en el salario medio por hora entre hombres y mujeres tenemos que contrastar en el modelo (1)

$$H_0 : \beta_2 = 0$$
$$H_1 : \beta_2 \neq 0$$

- El estadístico de contraste bajo la hipótesis de normalidad de los errores es

$$t = \frac{\widehat{\beta}_2}{SE(\widehat{\beta}_2)} \sim t_{523} \simeq N(0,1) \qquad \text{Bajo } H_0$$

- En este ejemplo $t = 2{,}481/0{,}3022 = 8{,}21$. Puesto que $|t| = 8{,}21 > z_{0{,}025} = 1{,}96$, podemos rechazar H_0 al 5 %, y por tanto concluimos que existen diferencias en el salario medio entre hombres y mujeres. Nótese que alternativamente podríamos haber utilizado el segundo modelo para hacer el contraste (contrastando $H_0 : \alpha_2 = 0; H_1 : \alpha_2 \neq 0$) y que la conclusión que obtendríamos sería idéntica ya que el valor absoluto del estadístico de contraste sería el mismo.

- Si quisieramos contrastar la hipótesis de que los hombres ganan más que las mujeres (e.d. que existe discriminación por sexo en el mercado laboral), tendríamos que contrastar

$$H_0 : \beta_2 = 0$$
$$H_1 : \beta_2 > 0$$

- Si rechazamos H_0 podemos afirmar que el salario medio de los hombres es mayor que el salario medio de las mujeres con la misma experiencia laboral. En este contraste unilateral, el estadístico de contraste es el mismo pero la región de rechazo cambia. Puesto que $t = 8{,}21 > z_{0{,}05} = 1{,}64$, rechazamos la hipótesis nula de que no existen diferencias salariales entre hombres y mujeres frente a la alternativa unilateral de que el salario medio de los hombres es mayor que el salario medio de las mujeres con la misma experiencia laboral. En el modelo (2), este mismo contraste se realizaría con el parámetro α_2 con las siguientes hipótesis:

$$H_0 : \alpha_2 = 0$$
$$H_1 : \alpha_2 < 0$$

- Por último también podemos escribir el mismo modelo incluyendo las dos dummies de sexo (Hombre y Mujer) y no incluyendo el término constante, es decir podemos especificar el modelo

$$salario_t = \delta_1 Hombre_t + \delta_2 Mujer_t + \delta_3 exp_t + u_t \qquad (3)$$

- Si calculamos ahora la media del salario para hombres y mujeres que tienen los mismos años de experiencia laboral tenemos

$$\begin{aligned} E(salario_t / Hombre_t = 1, Mujer_t = 0) &= \delta_1 + \delta_3 exp \\ E(salario_t / Hombre_t = 0, Mujer_t = 1) &= \delta_2 + \delta_3 exp \end{aligned} \qquad (3b)$$

- y por tanto

$$\delta_1 - \delta_2 = E(salario_t / t \text{ es hombre}) - E(salario_t / t \text{ es mujer})$$

$\delta_1 - \delta_2$ mide la diferencia en el salario medio entre hombres y mujeres con la misma experiencia laboral.

- Los modelos (1), (2) y (3) son equivalentes y tendremos

$$\left.\begin{array}{l} \delta_1 + \delta_3 exp = \beta_1 + \beta_2 + \beta_3 exp = \alpha_1 + \alpha_3 exp \\ \delta_2 + \delta_3 exp = \beta_1 + \beta_3 exp = \alpha_1 + \alpha_2 + \alpha_3 exp \end{array}\right\} \Rightarrow$$

$$\Rightarrow \left\{\begin{array}{l} \delta_3 = \beta_3 = \alpha_3 \\ \delta_1 = \beta_1 + \beta_2 = \alpha_1 \\ \delta_2 = \beta_1 = \alpha_1 + \alpha_2 \end{array}\right.$$

- Esta relación entre los parámetros del modelo (3) y los de los modelos (1) y (2) también se verifica para los estimadores MCO como ilustra el siguiente ejemplo

Ejemplo 1 (Continuación):

- En base a la misma muestra del ejemplo 1 se ha estimado ahora el modelo (3) obteniendo los siguientes resultados

$$\widehat{salario}_t = \underset{(0{,}2862)}{6{,}626}\, Hombre_t + \underset{(0{,}2846)}{4{,}145}\, Mujer_t + \underset{(0{,}0111)}{0{,}0269} exp_t$$

Podemos comprobar que efectivamente

$$\begin{aligned} \widehat{\beta}_3 &= \widehat{\alpha}_3 = 0{,}0269 = \widehat{\delta}_3 \\ \widehat{\beta}_1 + \widehat{\beta}_2 &= 4{,}145 + 2{,}481 = 6{,}626 = \widehat{\alpha}_1 = \widehat{\delta}_1 \\ \widehat{\alpha}_1 + \widehat{\alpha}_2 &= 6{,}626 - 2{,}481 = 4{,}145 = \widehat{\beta}_1 = \widehat{\delta}_2 \end{aligned}$$

- Nótese que la diferencia estimada en el salario medio entre hombres y mujeres con la misma experiencia laboral viene dada ahora por $\widehat{\delta}_1 - \widehat{\delta}_2 = 2{,}481$, que es exactamente el mismo resultado que obtuvimos antes.

- Para contrastar ahora la hipótesis nula de que no hay diferencias en el salario medio por hora entre hombres y mujeres tenemos que contrastar

$$H_0 : \delta_1 - \delta_2 = 0$$
$$H_1 : \delta_1 - \delta_2 \neq 0$$

El estadístico de contraste bajo la hipótesis de normalidad de los errores es

$$t = \frac{\widehat{\delta}_1 - \widehat{\delta}_2}{SE(\widehat{\delta}_1 - \widehat{\delta}_2)} \sim t_{523} \simeq N(0,1) \qquad \text{Bajo } H_0$$

- Para calcular el error estándar de $\widehat{\delta}_1 - \widehat{\delta}_2$ necesitamos la $\widehat{cov(\widehat{\delta}_1, \widehat{\delta}_2)}$ ya que

$$\begin{aligned} SE(\widehat{\delta}_1 - \widehat{\delta}_2) &= \sqrt{\widehat{var(\widehat{\delta}_1)} + \widehat{var(\widehat{\delta}_2)} - 2 * \widehat{cov(\widehat{\delta}_1, \widehat{\delta}_2)}} \\ &= \sqrt{\left(SE(\widehat{\delta}_1)\right)^2 + \left(SE(\widehat{\delta}_2)\right)^2 - 2 * \widehat{cov(\widehat{\delta}_1, \widehat{\delta}_2)}} \end{aligned}$$

- Sabiendo que $\widehat{cov(\widehat{\delta}_1, \widehat{\delta}_2)} = 0{,}0358$

$$SE(\widehat{\delta}_1 - \widehat{\delta}_2) = \sqrt{0{,}2862^2 + 0{,}2846^2 - 2 * 0{,}0358} = 0{,}3022$$

que coincide con el error estandar de $\widehat{\beta}_2$ y por tanto obtenemos de nuevo el mismo valor para el estadístico de contraste.

- En la práctica, puesto que los modelos son equivalentes, estimaríamos solo uno de ellos y es muy importante tener presente que aunque las conclusiones en base a los tres modelos son las mismas, la interpretación de los parámetros estimados son diferentes.
- **Nota:** No podemos especificar un modelo que incluya las dos dummies de sexo y el término constante ya que habría multicolinealidad exacta puesto que $Hombre_t + Mujer_t = 1$ para todas las observaciones.
- Si la variable dependiente está en logarítmos los coeficientes de las variables ficticias miden diferencias en porcentaje.

Ejemplo 2:

- En base a la misma muestra del ejemplo 1 se ha estimado ahora el siguiente modelo para el logaritmo del salario por hora

$$\log\widehat{(salario)}_t = \underset{(0{,}0404)}{1{,}355} + \underset{(0{,}0429)}{0{,}393}\, Hombre_t + \underset{(0{,}00158)}{0{,}00376} exp_t$$

- Ahora el coeficiente de la dummy de ser hombre mide la diferencia en porcentaje en el salario medio por hora entre hombres y mujeres con la misma experiencia, y por tanto, en base a estos resultados, se estima que los hombres tienen en media un salario que es un 39,3 % mayor que el de las mujeres con la misma experiencia laboral.

Modelos con un factor cualitativo con más de dos valores:

Supongamos ahora que queremos analizar si existen diferencias en el salario por hora entre distintas regiones. La región de residencia es también un factor cualitativo, pero a diferencia del sexo que solo puede tomar dos valores, los países están generalmente divididos en más de dos regiones. Supongamos que tenemos J regiones en un país, entonces podemos definir J variables ficticias

$$
\begin{aligned}
R_{1t} &= \begin{cases} 1 \text{ si el individuo } t \text{ vive en la región } 1 \\ 0 \text{ si el individuo } t \text{ no vive en la región } 1 \end{cases} \\
R_{2t} &= \begin{cases} 1 \text{ si el individuo } t \text{ vive en la región } 2 \\ 0 \text{ si el individuo } t \text{ no vive en la región } 2 \end{cases} \\
&\vdots \\
R_{Jt} &= \begin{cases} 1 \text{ si el individuo } t \text{ vive en la región } J \\ 0 \text{ si el individuo } t \text{ no vive en la región } J \end{cases}
\end{aligned}
$$

y considerar el modelo

$$salario_t = \beta_1 + \beta_2 R_{2t} + \beta_3 R_{3t} + ... + \beta_J R_{Jt} + \beta_{J+1} exp_t + u_t$$

donde hemos excluido la dummy de la región 1. Si calculamos en este modelo la media del salario para individuos que tengan los mismos años de experiencia laboral pero que vivan en distintas regiones tenemos

$$E(salario_t / t \text{ vive en la región } 1) = E(salario_t / R_{2t} = 0, R_{3t} = 0, ...R_{Jt} = 0)$$
$$= \beta_1 + \beta_{J+1} exp$$

$$E(salario_t / t \text{ vive en la región } 2) = E(salario_t / R_{2t} = 1, R_{3t} = 0, ...R_{Jt} = 0)$$
$$= \beta_1 + \beta_2 + \beta_{J+1} exp$$

$$E(salario_t / t \text{ vive en la región } 3) = E(salario_t / R_{2t} = 0, R_{3t} = 1, ...R_{Jt} = 0)$$
$$= \beta_1 + \beta_3 + \beta_{J+1} exp$$
$$\vdots$$
$$E(salario_t / t \text{ vive en la región } J) = E(salario_t / R_{2t} = 0, R_{3t} = 0, ...R_{Jt} = 1)$$
$$= \beta_1 + \beta_J + \beta_{J+1} exp$$

- Al incorporar las variables ficticias regionales estamos permitiendo que el término independiente del modelo pueda ser distinto para distintas regiones, ya que para la región 1 el término constante es β_1, para la región 2 es $\beta_1 + \beta_2$, .., y para la región J es $\beta_1 + \beta_J$, y por tanto, $\beta_2, .., \beta_J$ reflejan las posibles diferencias en el término constante entre individuos que viven en la regiones 2,...J, en comparación con los individuos que viven en la región 1.
- Además

$$
\begin{aligned}
\beta_2 &= E(salario_t / t \text{ en región } 2) - E(salario_t / t \text{ en región } 1) \\
\beta_3 &= E(salario_t / t \text{ en región } 3) - E(salario_t / t \text{ en región } 1) \\
&\vdots \\
\beta_J &= E(salario_t / t \text{ en región } J) - E(salario_t / t \text{ en región } 1)
\end{aligned}
$$

y por tanto β_2 mide la diferencia en el salario medio entre los individuos que viven en la región 2 comparados con los que viven en la región 1 con la misma experiencia laboral.

- Tenemos entonces que en este modelo las diferencias entre individuos que viven en distintas regiones se reflejan simplemente en el término constante y la hipótesis de ausencia de diferencias entre regiones sería $\beta_2 = \beta_3 = ... = \beta_J = 0$.
- También podríamos escribir el modelo incluyendo la dummy de la región 1 y excluyendo la dummy de cualquier otra región. Por ejemplo, si excluyeramos la de la región 2 los coeficientes de las demas dummies reflejarían diferencias en el salario medio entre las distintas regiones y la región 2.
- Los modelos definidos excluyendo cualquiera de las dummies regionales son equivalentes y podriamos obtener los estimadores MCO y las varianzas estimadas de cualquiera de los modelos utilizando los resultados para uno de los otros análogamente al ejemplo 1.

Ejemplo 3:

- En base a la misma muestra del ejemplo 1 se ha estimado ahora el siguiente modelo para el salario por hora

$$\widehat{salario}_t = \underset{(0{,}383)}{5{,}853} - \underset{(0{,}462)}{0{,}712} Norte_t - \underset{(0{,}430)}{1{,}057} Sur_t + \underset{(0{,}5121)}{0{,}217}\, Oeste_t + \underset{(0{,}0117)}{0{,}0329} exp_t$$

  siendo la categoría omitida la región Este. En base a estos resultados tenemos que el salario medio por hora para individuos con la misma experiencia laboral es 0,712 dolares menor en la región Norte que en la Este, 1,057 dolares menor en la región Sur que en la Este y 0,217 dolares mayor en la región Oeste que en la Este.

- Consideremos ahora el efecto de la educación sobre los salarios. Algunas bases de datos proporcionan información sobre los años de educación de los individuos y en ese caso podemos especificar el modelo

$$salario_t = \beta_1 + \beta_2 educ_t + \beta_2 exp_t + u_t$$

  donde $educ_t$ son los años de educación.

- En este modelo, β_2 es el efecto marginal de la educación sobre el salario. Un año más de educación supone un incremento en el salario que es el mismo cuando pasamos de 1 año a 2, de 2 a 3, etc, con independencia de que el año adicional suponga o no la obtención de algún título concreto como el titulo de Bachiller o un título universitario.
- Una alternativa a los años de educación, es considerar la variable cualitativa que refleja el grado educativo que posee cada individuo. Así podemos definir $J+1$ variables ficticias

$$
\begin{aligned}
Educ_{0t} &= \begin{cases} 1 \text{ si } t \text{ no ha finalizado el grado educativo 1} \\ 0 \text{ en caso contrario} \end{cases} \\
Educ_{1t} &= \begin{cases} 1 \text{ si } t \text{ ha finalizado el grado educativo 1} \\ \qquad \text{pero no ha finalizado el grado educativo 2} \\ 0 \text{ en caso contrario} \end{cases} \\
&\vdots \\
Educ_{Jt} &= \begin{cases} 1 \text{ si } t \text{ ha finalizado el grado educativo } J \\ 0 \text{ en caso contrario} \end{cases}
\end{aligned}
$$

- y considerar el modelo

$$salario_t = \alpha_1 + \delta_1 Educ_{1t} + \delta_2 Educ_{2t} + ... + \delta_J Educ_{Jt} + \alpha_2 exp_t + u_t$$

sin la dummy de no haber finalizado el grado educativo 1.

- En este modelo, la media del salario para individuos con los mismos años de experiencia pero con distintos grados educativos es:

$E(salario_t/t$ no ha finalizado el grado educativo 1)
$= E(salario_t/Educ_{1t} = 0, Educ_{2t} = 0, ...Educ_{Jt} = 0) = \alpha_1 + \alpha_2 exp$

$E(salario_t/t$ ha finalizado el grado educativo 1 pero no el 2)
$= E(salario_t/Educ_{1t} = 1, Educ_{2t} = 0, ...Educ_{Jt} = 0) = \alpha_1 + \delta_1 + \alpha_2 exp$

$E(salario_t/t$ ha finalizado el grado educativo 2 pero no el 3)
$= E(salario_t/Educ_{1t} = 0, Educ_{2t} = 1, ...Educ_{Jt} = 0) = \alpha_1 + \delta_2 + \alpha_2 exp$

$\vdots$

$E(salario_t/t$ ha finalizado el grado educativo J)
$= E(salario_t/Educ_{1t} = 0, Educ_{2t} = 0, ...Educ_{Jt} = 1) = \alpha_1 + \delta_J + \alpha_2 exp$

- Al incorporar las variables ficticias de educación estamos permitiendo que el término independiente del modelo pueda ser distinto para individuos con distintos niveles educativos, ya que para los individuos que no han completado el primer nivel educativo el término constante es α_1, para aquellos que han completado el primero pero no el segundo es $\alpha_1 + \delta_1, ..$ y por tanto, $\delta_1, .., \delta_J$ reflejan las posibles diferencias en el término constante entre individuos que han completado el nivel educativo 1,...J e individuos que no han completado el primer nivel educativo.
- Además

$$\begin{aligned}
\delta_1 &= E(salario_t / t \text{ ha finalizado el grado educativo 1 pero no el 2}) \\
&\quad -E(salario_t / t \text{ no ha finalizado el grado educativo 1}) \\
&\vdots \\
\delta_J &= E(salario_t / t \text{ ha finalizado el grado educativo } J) \\
&\quad -E(salario_t / t \text{ no ha finalizado el grado educativo 1})
\end{aligned}$$

- Por tanto δ_1 mide la diferencia en el salario medio entre los individuos que han completado el nivel educativo 1 pero no el 2 y los que no han completado el 1 y tienen la misma experiencia laboral, δ_2 mide la diferencia en el salario medio entre los individuos que han completado el nivel 2 pero no el 3 y los que no han completado el 1 con la misma experiencia laboral,.., y δ_J mide la diferencia en el salario medio entre los individuos que han completado el nivel J y los que no han completado el 1 con la misma experiencia laboral. Por tanto, en este modelo las diferencias entre individuos con distintos niveles educativos se reflejan simplemente en los términos constante y la hipótesis de ausencia de diferencias en el salario medio entre niveles educativos sería $\delta_1 = \delta_2 = ... = \delta_J = 0$.
- También podríamos escribir el modelo incluyendo la dummy $Educ_0$ y excluyendo cualquiera de las otras. Los modelos definidos excluyendo cualquiera de las dummies de educación son equivalentes y podriamos obtener los estimadores MCO y las varianzas estimadas de cualquiera de los modelos utilizando los resultados para uno de los otros análogamente al ejemplo 1.

Ejemplo 4:

- En base a la misma muestra del ejemplo 1 se ha estimado ahora el siguiente modelo para el salario por hora

$$salario_t = \alpha_1 + \delta_1 Bachiller_t + \delta_2 Univ_t + u_t$$

  donde $Bachiller_t$ toma el valor 1 si el inviduo t ha completado el nivel educativo que permite acceder a la universidad pero no tiene un título universitario y $Univ_t$ toma el valor 1 si el individuo t tiene un título universitario. La categoría omitida es no haber finalizado el "bachillerato".
- Los resultados de esta estimación son:

$$\widehat{salario}_t = \underset{(0{,}3797)}{2{,}862} + \underset{(0{,}3628)}{1{,}708}\, Bachiller_t + \underset{(0{,}4373)}{4{,}917}\, Univ_t + \underset{(0{,}0108)}{0{,}0570} exp_t$$

- El salario medio por hora para individuos con la misma experiencia laboral es 1,708 dolares mayor para los que han completado el "bachillerato"que para los que no lo han finalizado y 4,917 dólares mayor para los que tienen un título universitario.

- Para comparar el salario medio entre universitarios y estudiantes con el "bachillerato"terminado con la misma experiencia, tenemos que comparar los coeficientes de $Univ_t$ y $Bachiller_t$. La diferencia entre los coeficientes de las variables *Univ* y *Bachiller* es la diferencia de salarios medio que buscamos para un mismo nivel de experiencia, como indica la siguiente expresión:

$$\begin{aligned}
&\delta_2 - \delta_1 = \\
&= [E(salario_t / t \text{ universitario}) - E(salario_t / t \text{ no bachillerato})] \\
&- [E(salario_t / t \text{ con bachillerato}) - E(salario_t / t \text{ no bachillerato})] \\
&= E(salario_t / t \text{ universitario}) - E(salario_t / t \text{ con bachillerato})
\end{aligned}$$

- Las estimaciones obtenidas indican que los individuos con título universitario ganan un salario medio 3.209 dólares mayor que el salario medio de los individuos que han acabado "bachillerato".
- Para contrastar la hipótesis nula de que las diferencias entre universitarios y estudiantes con bachillerato no son significativas (es decir, $H_0 : \delta_2 - \delta_1 = 0$ frente $H_1 : \delta_2 - \delta_1 \neq 0$) es necesario conocer la covarianza estimada entre $\hat{\delta}_1$ y $\hat{\delta}_2$.

- Sin embargo, es más fácil realizar este contraste si el modelo se estima especificando como categoría omitida haber finalizado el "bachillerato", es decir

$$salario_t = \gamma_1 + \lambda_1 NoBach_t + \lambda_2 Univ_t + u_t$$

donde $NoBach_t$ toma el valor 1 si el individuo no ha completado el "bachillerato" y donde λ_2 es la diferencia entre el salario medio de los individuos con título universitario y el salario medio de los individuos con "bachillerato".
- Las estimaciones correspondientes a este modelo son

$$\widehat{salario}_t = \underset{(0{,}2688)}{4{,}5702} - \underset{(0{,}3628)}{1{,}708}\, NoBach_t + \underset{(0{,}3603)}{3{,}209}\, Univ_t + \underset{(0{,}0108)}{0{,}0570} exp_t$$

- La estimación del coeficiente de *Univ* coincide con la diferencia salarial media entre universitarios e individuos que han finalizado sólo el "bachillerato", puesto que este es el grupo de referencia. Por lo tanto, el contraste anterior es equivalente a $H_0 : \lambda_2 = 0$ frente a la alternativa $H_1 : \lambda_2 \neq 0$, para el cuál es únicamente necesario conocer el error estándar de $\widehat{\lambda}_2$

Modelos con varios factores cualitativos:

- Consideremos ahora la variable ficticia

$$Casado_t = \begin{cases} 1 \text{ si el individuo } t \text{ está casado} \\ 0 \text{ si el individuo } t \text{ no está casado} \end{cases}$$

y el modelo

$$salario_t = \beta_1 + \beta_2 Hombre_t + \beta_3 Casado_t + \beta_4 exp_t + u_t \qquad (4)$$

- Si calculamos en este modelo la media del salario para hombres y mujeres, casados y no casados, y que tengan los mismos años de experiencia laboral tenemos

$$E(salario_t / t \text{ es mujer no casada}) = \beta_1 + \beta_4 exp$$

$$E(salario_t / t \text{ es hombre no casado}) = \beta_1 + \beta_2 + \beta_4 exp$$

$$E(salario_t / t \text{ es mujer casada}) = \beta_1 + \beta_3 + \beta_4 exp$$

$$E(salario_t / t \text{ es hombre casado}) = \beta_1 + \beta_2 + \beta_3 + \beta_4 exp$$

- Por tanto

$$\begin{aligned}
\beta_2 &= E(salario_t/t \text{ hombre no casado})\text{-}E(salario_t/t \text{ mujer no casada}) \\
&= E(salario_t/t \text{ hombre casado})\text{-}E(salario_t/t \text{ mujer casada}) \\
\beta_3 &= E(salario_t/t \text{ mujer casada})\text{-}E(salario_t/t \text{ mujer no casada}) \\
&= E(salario_t/t \text{ hombre casado})\text{-}E(salario_t/t \text{ hombre no casado})
\end{aligned}$$

  de forma que en este modelo β_2 mide las diferencias en el salario medio por hora entre hombres y mujeres con la misma experiencia laboral y el mismo estado civil, mientras que β_3 mide las diferencias en el salario medio por hora entre individuos casados y no casados con la misma experiencia laboral y el mismo sexo.
- También podríamos escribir el modelo utilizando la dummy de ser mujer en lugar de la de ser hombre y/o la de no estar casado en lugar de la de estar casado y los modelos serían equivalentes, y podríamos obtener los estimadores MCO y las varianzas estimadas de cualquiera de los modelos utilizando los resultados para uno de los otros análogamente al ejemplo 1.

Ejemplo 5:

- En base a la misma muestra del ejemplo 1 se ha estimado ahora el siguiente modelo para el salario por hora

$$\widehat{salario}_t = 3{,}73 + \underset{(0{,}3025)}{2{,}298}\, Hombre_t + \underset{(0{,}3262)}{1{,}222}\, Casado_t + \underset{(0{,}0116)}{0{,}0133} exp_t$$

- Según estos resultados el salario medio de los hombres es 2,298 dólares mayor que el de las mujeres con la misma experiencia laboral y el mismo estado civil, mientras que el salario medio de los casados es 1,222 dólares mayor que el de los solteros con la misma experiencia laboral y el mismo sexo.

Interacciones entre dos variables ficticias:

- El modelo

$$salario_t = \beta_1 + \beta_2 Hombre_t + \beta_3 Casado_t + \beta_4 exp_t + u_t$$

  no permite que las diferencias por sexo para un nivel de experiencia y estado civil dados, difieran dependiendo del estado civil del individuo.

- Para permitir que las diferencias entre hombres y mujeres sean diferentes para los distintos estados civiles, y que las diferencias salariales medias entre casados y solteros sean distintas entre hombres y mujeres, debemos añadir al modelo una variable que sea una interación entre la variable ficticia de *Hombre* y la variable ficticia *Casado*.

- Así pues, la variable ficticia interacción $HomCasado_t = Hombre_t * Casado_t$ toma el valor 1 si el individuo t es hombre y está casado, y 0 en caso contrario.

- Consideremos el siguiente modelo:

$$salario_t = \gamma_1 + \gamma_2 Hombre_t + \gamma_3 Casado_t + \gamma_4 HomCasado_t + \gamma_5 exp_t + u_t$$

- Si calculamos en este modelo la media del salario para hombres y mujeres, casados y no casados, y que tengan los mismos años de experiencia laboral obtenemos

$$E(salario_t / t \text{ es mujer no casada}) = \gamma_1 + \gamma_5 exp$$

$$E(salario_t / t \text{ es hombre no casado}) = \gamma_1 + \gamma_2 + \gamma_5 exp$$

$$E(salario_t / t \text{ es mujer casada}) = \gamma_1 + \gamma_3 + \gamma_5 exp$$

$$E(salario_t / t \text{ es hombre casado}) = \gamma_1 + \gamma_2 + \gamma_3 + \gamma_4 + \gamma_5 exp$$

- Por lo tanto,

$$
\begin{aligned}
\gamma_2 &= E(salario_t/\text{hombre no casado})\text{-}E(salario_t/\text{mujer no casa}\\
\gamma_2+\gamma_4 &= E(salario_t/\text{hombre casado})\text{-}E(salario_t/\text{mujer casada})\\
\gamma_3 &= E(salario_t/\text{mujer casada})\text{-}E(salario_t/\text{mujer no casada})\\
\gamma_3+\gamma_4 &= E(salario_t/\text{hombre casado})\text{-}E(salario_t/\text{hombre no casad}
\end{aligned}
$$

- Introducir en el modelo la interación entre las dos variables ficticias nos permite que las diferencias entre hombres y mujeres sean distintas entre solteros y casados (la diferencia por sexos para individuos casados es γ_4 unidades mayor que para los no casados). Asimismo, las diferencias entre solteros y casados es distinta entre hombres y mujeres (la diferencia entre casados y solteros es γ_4 unidades mayor para los hombres que para las mujeres).

Ejemplo 5 (Continuación):

- En base a la misma muestra del ejemplo 1 se ha estimado el modelo (5), y los resultados obtenidos son los siguientes:

$$\begin{aligned}\widehat{salario}_t &= 4{,}47 + \underset{(0{,}4739)}{0{,}575}\, Hombre_t - \underset{(0{,}4320)}{0{,}132}\, Casado_t \\ &\quad + \underset{(0{,}6081)}{2{,}8365} HomCasado_t + \underset{(0{,}0113)}{0{,}0110} exp_t\end{aligned}$$

- Para contrastar la hipótesis nula de que las diferencias salariales medias por sexo son iguales entre casados y no casados, debemos contrastar la significatividad individual del parámetro γ_4, es decir

$$\begin{aligned} H_0 &: \gamma_4 = 0 \\ H_1 &: \gamma_4 \neq 0 \end{aligned}$$

- En este caso,

$$t = \frac{\widehat{\gamma}_4}{SE(\widehat{\gamma}_4)} \sim t_{523} \simeq N(0,1) \qquad \text{Bajo } H_0$$

  En este ejemplo $t = 2{,}8365/0{,}6081 = 4{,}66$.
- Puesto que $|t| = 4{,}66 > z_{0{,}025} = 1{,}96$, podemos rechazar H_0 al 5 %, y por tanto concluimos las diferencias salariales medias entre sexos es diferente para los individuos casados que para los individuos no casados.

Interacciones en las que intervienen variables ficticias:

- El modelo

$$salario_t = \beta_1 + \beta_2 Hombre_t + \beta_3 exp_t + u_t \qquad (1)$$

permite analizar posibles diferencias en el salario medio entre hombres y mujeres pero en este modelo se supone que el efecto marginal de la experiencia es el mismo para hombres y mujeres.

- Para permitir que este efecto marginal sea distinto para hombres y mujeres podemos considerar un modelo que incluya la interacción de la variable experiencia con una de las dummies de sexo. Por ejemplo, podemos considerar el modelo

$$salario_t = \beta_1 + \beta_2 Hombre_t + \beta_3 exp_t + \beta_4 Hombreexp_t + u_t \qquad (6)$$

donde $Hombreexp_t = Hombre_t * exp_t$.

- En este modelo el efecto marginal de la experiencia puede ser distinto para hombres y mujeres ya que

$$\frac{\partial salario_t}{\partial exp_t} = \beta_3 + \beta_4 Hombre_t$$

y por tanto β_3 mide el efecto que tiene para las mujeres un año más de experiencia laboral mientras que $\beta_3 + \beta_4$ mide ese efecto para los hombres.
- La hipótesis de que el efecto marginal de la experiencia sobre los salarios es el mismo para hombres y mujeres es $\beta_4 = 0$.
- Por otra parte

$$E(salario_t / exp_t = 0) = \beta_1 + \beta_2 Hombre_t$$

β_1 es el salario medio de las mujeres sin experiencia laboral y $\beta_1 + \beta_2$ es el salario medio de los hombres sin experiencia laboral. Por tanto, la hipótesis de que los salarios son idénticos para hombres y mujeres con la misma experiencia laboral (hipótesis de ausencia de discriminación salarial por sexo) es $\beta_2 = \beta_4 = 0$.

- Alternativamente podríamos haber incluido la dummy de ser mujer en lugar de la de ser hombre

$$salario_t = \alpha_1 + \alpha_2 Mujer_t + \alpha_3 exp_t + \alpha_4 Mujerexp_t + u_t \qquad (7)$$

- En este modelo el efecto marginal de la experiencia es

$$\frac{\partial salario_t}{\partial exp} = \alpha_3 + \alpha_4 Mujer_t$$

  y por tanto α_3 mide el efecto que tiene para los hombres un año más de experiencia laboral mientras que $\alpha_3 + \alpha_4$ mide ese efecto para las mujeres. La hipótesis de que el efecto marginal de la experiencia sobre los salarios es el mismo para hombres y mujeres es $\alpha_4 = 0$.

- Por otra parte

$$E(salario_t / exp_t = 0) = \alpha_1 + \alpha_2 Mujer_t$$

  α_1 es el salario medio de los hombres sin experiencia laboral y $\alpha_1 + \alpha_2$ es el salario medio de las mujeres sin experiencia laboral.

- Por tanto, la hipótesis de que los salarios son idénticos para hombres y mujeres con la misma experiencia laboral (hipótesis de ausencia de discriminación salarial por sexo) es $\alpha_2 = \alpha_4 = 0$. La hipótesis nula de que el efecto marginal de la experiencia laboral es el mismo para hombres y para mujeres (permitiendo que los salarios medios para un mismo nivel de experiencia sean distintos entre hombres y mujeres) es $\alpha_4 = 0$.
- Los modelos (6) y (7) son equivalentes y si comparamos los dos modelos tenemos que

$$
\begin{aligned}
\alpha_1 &= \beta_1 + \beta_2 \\
\alpha_1 + \alpha_2 &= \beta_1 \\
\alpha_3 &= \beta_3 + \beta_4 \\
\alpha_3 + \alpha_4 &= \beta_3
\end{aligned}
$$

- Finalmente también podemos escribir el modelo como

$$salario_t = \gamma_1 Hombre_t + \gamma_2 Mujer_t + \gamma_3 Hombreexp_t + \gamma_4 Mujerexp_t + u_t$$

- En este caso el efecto marginal de la experiencia es

$$\frac{\partial salario_t}{\partial exp} = \gamma_3 Hombre_t + \gamma_4 Mujer_t$$

  y por tanto γ_3 mide el efecto que tiene para los hombres un año más de experiencia laboral mientras que γ_4 mide ese efecto para las mujeres.
- La hipótesis de que el efecto marginal de la experiencia sobre los salarios es el mismo para hombres y mujeres es $\gamma_3 = \gamma_4$.

- Por otra parte

$$E(salario_t/exp_t = 0) = \gamma_1 Hombre_t + \gamma_2 Mujer_t$$

γ_1 es el salario medio de los hombres sin experiencia laboral y γ_2 es el salario medio de las mujeres sin experiencia laboral. Por tanto, la hipótesis de que los salarios son idénticos para hombres y mujeres con la misma experiencia laboral (hipótesis de ausencia de discriminación salarial por sexo) es $\gamma_1 = \gamma_2$, $\gamma_3 = \gamma_4$. Los modelos (6), (7) y (8) son equivalentes y si comparamos los modelos tenemos que

$$\begin{aligned}
\gamma_1 &= \alpha_1 = \beta_1 + \beta_2 \\
\gamma_2 &= \alpha_1 + \alpha_2 = \beta_1 \\
\gamma_3 &= \alpha_3 = \beta_3 + \beta_4 \\
\gamma_4 &= \alpha_3 + \alpha_4 = \beta_3
\end{aligned}$$

Ejemplo 6:

- En base a la misma muestra del ejemplo 1 se ha estimado ahora el siguiente modelo para el salario por hora

$$\widehat{salario}_t = \underset{(0{,}3398)}{4{,}612} + \underset{(0{,}4819)}{1{,}547}\, Hombre_t + \underset{(0{,}0159)}{0{,}0015} exp_t + \underset{(0{,}0222)}{0{,}0551} Hombreexp_t$$

- Según estos resultados, un año más de experiencia supone un incremento del salario por hora de 0,0015 dólares para las mujeres (practicamente nulo, de hecho la variable *exper* no es significativa) y de $0{,}0015 + 0{,}0551 = 0{,}0566$ dólares para los hombres.
- El efecto marginal de la experiencia en el salario es 0,0551 dólares mayor para hombres que para mujeres.
- Por otra parte el salario medio estimado para individuos sin experiencia laboral es 4,612 dólares para las mujeres y $4{,}612 + 1{,}547 = 6{,}159$ para los hombres.

- Alternativamente podemos escribir el modelo:

$$\widehat{salario}_t = \underset{(0{,}3417)}{6{,}159} - \underset{(0{,}4819)}{1{,}547}\, Mujer_t + \underset{(0{,}0154)}{0{,}0566} exp_t - \underset{(0{,}0222)}{0{,}0551} Mujerexp_t$$

$$\widehat{salario}_t = \underset{(0{,}3417)}{6{,}159}\, Hombre_t + \underset{(0{,}3398)}{4{,}612}\, Mujer_t + \underset{(0{,}0154)}{0{,}0566} Hombreexp_t + \underset{(0{,}0159)}{+0{,}0015} Mujerexp_t$$

  Los efectos estimados son idénticos a los que obtuvimos antes.
- Nótese que en los modelos (6), (7) y (8) permitimos que tanto el término constante como la pendiente sean distintas para hombres y mujeres.
- Otra posibilidad sería estimar por separado para hombres y mujeres el modelo de regresión del salario sobre una constante y la experiencia laboral. La diferencia es que cuando estimamos por separado permitimos que la varianza de los errores sea distinta para hombres y mujeres, mientras que cuando estimamos el modelo (6) estámos imponiendo que la varianza de los errores es la misma a la hora de calcular los errores estándar.

Ejemplo 6 (Continuación):

- En base a la misma muestra del ejemplo 1 se ha estimado ahora el siguiente modelo para el salario por hora por separado para hombres y mujeres

$$\text{Para la muestra de mujeres:} \quad \widehat{salario_t} = \underset{(0{,}2501)}{4{,}612} + \underset{(0{,}0117)}{0{,}0015} exp_t$$

$$\text{Para la muestra de hombres:} \quad \widehat{salario_t} = \underset{(0{,}4073)}{6{,}159} + \underset{(0.{,}0184)}{0{,}0566} exp_t$$

- Si comparamos estos resultados vemos que los efectos estimados son idénticos. Los efectos marginales de la experiencia y los salarios medios para individuos sin experiencia laboral coinciden tanto para hombres como para mujeres con los que obteniamos anteriormente. Lo que cambia son los errores estándar, ya que ahora no estamos imponiendo que la varianza de los errores sea la misma para hombres y mujeres.

Contrastes de cambio estructural

- Cuando trabajamos con datos de series temporales podemos estar interesados en analizar si los parámetros del modelo son diferentes en dos sub-periodos temporales distintos, para eso consideraremos el modelo:

$$\begin{array}{ll} Y_t = \beta_1 + \beta_2 X_{2t} + .. + \beta_k X_{kt} + u_t, & t = 1, 2, ..., T_1 \\ Y_t = \alpha_1 + \alpha_2 X_{2t} + .. + \alpha_k X_{kt} + u_t, & t = T_1 + 1, T_1 + 2, ..., T \end{array} \tag{9}$$

- Una posibilidad es estimar por separado el modelo para los dos sub-periodos y posteriormente contrastar si los parámetros de los dos modelos son iguales, es decir contrastar $\beta_1 = \alpha_1, \beta_2 = \alpha_2, .., \beta_k = \alpha_k$.
- Cuando los parámetros no son iguales en los dos sub-periodos diremos que se ha producido un cambio estructural.

- Alternativamente, si la varianza del término de error es constante para todas las observaciones, es decir si $var(u_t) = \sigma^2$ para $t = 1, 2, ..., T_1, T_1 + 1, ..., T$, podemos escribir el modelo utilzando una variable ficticia que tome el valor 1 para uno de los dos sub-periodos y cero para el otro.
- Así, si definimos

$$\begin{cases} D_{1t} = 1, \text{ si } t \leq T_1 \\ D_{1t} = 0, \text{ en otro caso} \end{cases}$$

$$\begin{cases} D_{2t} = 1, \text{ si } t > T_1 \\ D_{2t} = 0, \text{ en otro caso} \end{cases}$$

- Podemos escribir el modelo como

$$Y_t = \gamma_1 + \delta_1 D_{1t} + \gamma_2 X_{2t} + \delta_2 D_{1t} X_{2t} + .. + \gamma_k X_{kt} + \delta_k D_{kt} X_{2t} + u_t \quad (10)$$

- En este modelo el efecto marginal de X_j sobre Y puede ser distinto en los dos sub-periodos

$$\begin{aligned}\frac{\partial Y_t}{\partial X_{jt}} &= \gamma_j + \delta_j \text{ si } t \leq T_1 \text{ (es decir, si } D_{1t} = 1) \\ \frac{\partial Y_t}{\partial X_{jt}} &= \gamma_j \text{ si } t > T_1 \text{ (es decir, si } D_{2t} = 1)\end{aligned}$$

- El término constante del modelo también puede ser distinto en los dos sub-periodos ya que para $t \leq T_1$ el término constante es $\gamma_1 + \delta_1$ mientras que para $t > T_1$ el término constante es γ_1.
- Puesto que los parámetros son iguales en los dos sub-periodos si $\delta_1 = \delta_2 = ... = \delta_k = 0$, el contraste de cambio estructural consiste en contrastar esta hipótesis frente a la alternativa de que al menos uno de los δ_j sea distinto de cero.

- Nótese que existe la siguiente equivalencia entre los parámetros de los modelos (9) y (10) y también entre los estimadores MCO de los dos modelos

$$\begin{aligned} \gamma_j + \delta_j &= \beta_j \\ \gamma_j &= \alpha_j \end{aligned} \qquad j = 1, 2, .., k$$

- Además, los residuos MCO del modelo (10) coinciden con los residuos MCO de la primera ecuación del modelo (9) para $t = 1, 2, .., T_1$ y con los residuos MCO de la segunda ecuación del modelo (9) para $t = T_1 + 1, T_1 + 2, .., T$.
- Consecuentemente, la SCR del modelo (10) es la suma de la SCR del modelo en la primera ecuación del modelo (9) y la SCR del modelo en la segunda ecuación del modelo (9).

- Teniendo en cuenta la hipótesis nula del contraste del cambio estructural, el modelo sin restringir es el modelo (10) mientras que el modelo restringido es

$$Y_t = \gamma_1 + \gamma_2 X_{2t} + .. + \gamma_k X_{kt} + u_t \qquad (11)$$

- Así pues, el estadístico de contraste del cambio estructural es el estadístico F que utilizamos en el Tema anterior:

$$F = \frac{[(SCR_{11} - SCR_{10})/k]}{[(SCR_{11})/(T-2k)]} \sim F_{k,T-2k} \qquad \text{bajo } H_0$$

  puesto que el modelo sin restringir contiene $2k$ variables explicativas y donde SCR_{11} es la suma cuadrática residual obtenida de la estimación por MCO del modelo (11) y SCR_{10} es la suma cuadrática residual obtenida de la estimación por MCO del modelo (10).

- Alternativamente, igual que en el caso de los modelos (6), (7) y (8), podríamos haber escrito el modelo interaccionando todas las variables con D_{2t} en lugar de con D_{1t}, o bien incluyendo las interaciones de todas las variables con las dos dummies y sin incluir las variables explicativas sin interaccionar con las dummies.

- Como señalabamos anteriormente, para que los modelos (9) y (10) sean equivalentes tiene que verificarse que la varianza del error coincida en los dos sub-periodos.
- Vamos a ver cómo contrastar esta hipótesis: Sea σ_1^2 la varianza de los errores de las observaciones correspondientes al primer sub-periodo y sea σ_2^2 la varianza de los errores de las observaciones correspondientes al segundo sub-periodo, es decir

$$\begin{aligned} var(u_t) &= \sigma_1^2, \quad t = 1, 2, .., T_1 \\ var(u_t) &= \sigma_2^2, \quad t = T_1 + 1, T_1 + 2, .., T \end{aligned}$$

- La varianza de los errores será la misma para todas las observaciones si $\sigma_1^2 = \sigma_2^2$, y por tanto, vamos a proponer un contraste para

$$H_0 : \sigma_1^2 = \sigma_2^2$$
$$H_1 : \sigma_1^2 \neq \sigma_2^2$$

- Si los errores son normales podemos utilizar el estadístico de contraste

$$F = \frac{\widehat{\sigma}_1^2}{\widehat{\sigma}_2^2} \sim F_{T_1-k, T_2-k} \quad \text{Bajo } H_0$$

donde T_2 es el número de observaciones en el segundo sub-periodo y

$$\widehat{\sigma}_1^2 = \frac{\sum_{t=1}^{T_1} e_t}{T_1 - k} \qquad \text{y} \qquad \widehat{\sigma}_2^2 = \frac{\sum_{t=T_1+1}^{T} e_t}{T_2 - k}$$

y e_t son los residuos MCO del modelo (9) y (10). Puesto que se trata de un contraste de dos colas, rechazaremos H_0 a nivel α si $F > F_{T_1-k, T_2-k, \alpha/2}$ o bien $F < F_{T_1-k, T_2-k, 1-\alpha/2}$.