Hoja de Ejercicios 4
Análisis de regresión con información cualitativa

Nota: En aquellos ejercicios en los que se incluyen estimaciones y referencia al archivo de datos utilizado, el estudiante debería comprobar los resultados obtenidos en Gretl.

1. [Basado en ejecicios 4.2 y 7.13 del Wooldridge] Para explicar el salario de un director general, *salary*, se ha estimado la siguiente ecuación con los datos del archivo `CEOSAL1.GDT`:

$$\begin{aligned}\widehat{\log(salary)} &= \underset{(0.294)}{4.362} + \underset{(0.033)}{0.275}\log(sales) + \underset{(0.0039)}{0.0179}roe \\ n &= 209, \quad R^2 = 0.282\end{aligned}$$

donde *sales* son las ventas anuales y *roe* es el rendimiento sobre el valor nominal de una acción (return-on-equity).

   (a) Interprete el coeficiente de $\log(sales)$ y contraste si es significativamente positivo.

   (b) Se decide incluir una variable dummy, *rosneg*, que es igual a 1 cuando *ros* es negativa y 0 si *ros* es cero o positivo, donde *ros* es el rendimiento sobre el valor real de la acción, proponiendo la especificación

$$\begin{aligned}\log(salary) &= \beta_0 + \beta_1\log(sales) + \beta_2 roe + \beta_3 rosneg + \beta_4\log(sales)*rosneg \\ &\quad +\beta_5 roe*rosneg + \varepsilon\end{aligned}$$

   y obteniendo la siguiente estimación MCO,

$$\begin{aligned}\widehat{\log(salary)} &= \underset{(0.307)}{4.074} + \underset{(0.035)}{0.314}\log(sales) + \underset{(0.004)}{0.017}roe \\ &\quad +\underset{(1.009)}{2.094}rosneg \underset{(0.112)}{-0.258}\log(sales)*rosneg - \underset{(0.0178)}{0.00343}roe*rosneg \\ n &= 209, \qquad R^2 = 0.315,\end{aligned}$$

   obteniéndose las siguientes varianzas estimadas para los coeficientes de $\log(sales)$, *roe*, *rosneg*, $\log(sales)*rosneg$ y $roe*rosneg$, respectivamente,

$$\widehat{V}\begin{pmatrix}\hat\beta_1\\ \hat\beta_2\\ \hat\beta_3\\ \hat\beta_4\\ \hat\beta_5\end{pmatrix} = \begin{pmatrix} 0.001236 & 1.47E-05 & 0.010414 & -0.001236 & -1.47E-05 \\ 1.47E-05 & 1.64E-05 & 0.000410 & -1.47E-05 & -1.64E-05 \\ 0.010414 & 0.000410 & 1.018975 & -0.109247 & -0.003193 \\ -0.001236 & -1.47E-05 & -0.109247 & 0.012544 & -0.000116 \\ -1.47E-05 & -1.64E-05 & -0.003193 & -0.000116 & 0.000318 \end{pmatrix}.$$

   Contraste si es necesario distinguir en el modelo las empresas en función del signo de *ros*.

   (c) Contraste si, para las empresas que tienen *ros* negativo, un aumento de las ventas implica necesariamente un aumento del salario del director general, manteniéndose los demás factores constantes.

   (d) Explique cómo contrastaría la hipótesis de que para un director general de una empresa con *ros* negativo, $\log(sales) = 10$ y $roe = 20$, es igual de interesante *(a)* incrementar las ventas hasta $\log(sales) = 11$, o *(b)* conseguir que el *ros* pase a ser positivo (en ambos casos, sin cambiar las otras variables).

2. Considere los siguientes modelos para explicar el peso de un recién nacido, *bwght*, con los datos para EE.UU. del archivo `BWGHT.GDT`, donde *bwght* es el peso (en onzas) del niño al nacer, *cigs* es el número diario de cigarrillos fumados por la madre durante el embarazo, *faminc* es la renta familiar anual, en miles de dólares, *male* es una variable ficticia que indica si el recién nacido es niño ($male = 1$) o niña ($male = 0$), y *white* es otra variable ficticia que indica si es blanco ($= 1$) o no ($= 0$) y $nowhite = 1 - white$.

$$\begin{aligned}
\widehat{\log(bwght)} &= \underset{(.019)}{4.69} - \underset{(.00085)}{0.0042}cigs + \underset{(.0059)}{0.0084}\log(faminc) + \underset{(.01)}{0.026}male + \underset{(.014)}{0.053}white \\
R^2 &= 0.0416, \quad n = 1388 \\
\widehat{\log(bwght)} &= 4.687 - 0.0042cigs + 0.0083\log(faminc) + 0.028male + 0.054white \\
&\quad -0.002white * male \\
R^2 &= 0.0417, \quad n = 1388 \\
\widehat{\log(bwght)} &= 4.689 - 0.0042cigs + 0.0077\log(faminc) + 0.028male * nowhite \\
&\quad +0.0677white \\
R^2 &= 0.0381, \quad n = 1388
\end{aligned}$$

   (a) En la primera ecuación, interprete el coeficiente de la variable *cigs*. Proporcione un intervalo de confianza al 95% para el efecto sobre el peso del recién nacido de fumar 10 cigarrillos más, manteniendo todos los demás factores constantes.

   (b) Considere ahora las dos primeras ecuaciones ¿En cuánto predice cada modelo el peso adicional de un niño ($male = 1$) recién nacido blanco respecto a otro que no sea blanco, manteniendo los demás factores constantes? ¿Es la diferencia entre las dos predicciones significativa?

   (c) Usando el segundo modelo, estime la diferencia de peso entre una niña y un niño al nacer, ambos blancos, manteniéndose constantes todos los restantes factores. ¿Es dicha diferencia significativa?

3. Las siguientes ecuaciones de salarios han sido estimadas utilizando datos correspondientes a trabajadores en Bangladesh:

$$\widehat{\log(salario)} = \underset{(0.35)}{1.25} + \underset{(0.03)}{0.15}\,hombre + \underset{(0.004)}{0.02}\,experiencia, \tag{1}$$

$$\widehat{\log(salario)} = \underset{(0.48)}{1.55} + \underset{(0.05)}{0.10}\,hombre + \underset{(0.005)}{0.015}\,experiencia - \underset{(0.002)}{0.005}\,hombre{*}experiencia, \tag{2}$$

donde *salario* está medido en dólares de EEUU, y *hombre* es una variable binaria que toma el valor 1 si el trabajador es hombre y 0 si es mujer; *experiencia* mide los años de experiencia laboral. Los números entre paréntesis son los errores estándar.

   (a) ¿Cuál es la diferencia media estimada entre el salario de un hombre con 5 años de experiencia y el de una mujer con 10 años de experiencia utilizando la ecuación (1)?

   (b) ¿Cuál es la diferencia media estimada entre el salario de un hombre con 5 años de experiencia y el de una mujer con 10 años de experiencia utilizando la ecuación (2)?

   (c) Contraste que la diferencia salarial entre hombres y mujeres no depende de la experiencia.

4. Suponga que se reúne información sobre salarios, educación, experiencia laboral y sexo a partir de una encuesta. Además, se pregunta sobre el consumo de marihuana. La pregunta se formula así: "¿En cuántas ocasiones fumaste marihuana el mes pasado?"

   (a) Escriba una ecuación que permita estimar los efectos de su consumo en el salario, tomando en cuenta los efectos de otros factores. El objetivo es poder realizar afirmaciones del tipo "si consume cinco veces más marihuana al mes se prevé un cambio en el salario de x %".

   (b) Especifique un modelo que permita contrastar si el consumo de drogas tiene distintos efectos en los salarios de hombres y mujeres. ¿Cómo se contrastaría que no existen diferencias entre hombres y mujeres?

   (c) Suponga que se considera preferible medir el consumo de marihuana clasificando a la gente en cuatro categorías: no consumidor, consumidor ocasional (de una a cinco veces al mes), consumo moderado (de seis a diez) y consumidor habitual (más de diez veces al mes). Escriba un modelo que permita estimar los efectos de esta droga sobre el salario.

   (d) Usando el modelo propuesto en el apartado *(c)*, explique con detalle cómo contrastaría la hipótesis nula de que el consumo de marihuana no afecta al salario. La respuesta debe ser muy específica e incluir una lista detallada de los grados de libertad.

   (e) ¿Cuáles son los problemas potenciales para realizar inferencia causal con estos datos de encuesta?

5. Supongamos que estamos interesados en analizar las posibles diferencias en el consumo de cerveza según el sexo. Para ello especificamos el modelo de regresión lineal

$$Y_i = \beta_0 + \beta_1 D_i + \beta_2 X_i + \beta_3 (D_i X_i) + u_i$$

   donde Y_i es el gasto en cerveza del individuo i, X_i es su renta y D_i es una variable artificial que vale 1 si el individuo es mujer y 0 si es hombre. En base a una muestra de tamaño $n = 34$ se ha obtenido el siguiente resultado:

$$\begin{array}{rcll} \widehat{Y}_i & = & 186.47 - 126.00 D_i + 2.33 X_i - 1.29 (D_i X_i) & R^2 = 0.5055 \\ & & \quad (45.67) \quad (57.01) \qquad (0.86) \quad (1.02) & \end{array}$$

   Los números entre paréntesis son errores estándar. Además, utilizando la misma muestra se ha estimado el modelo $Y_i = \alpha_0 + \alpha_1 X_i + u_i$, obteniéndose un coeficiente de determinación de 0.1355.

   (a) ¿Cuál será la diferencia en el consumo de cerveza entre un hombre y una mujer con el mismo nivel de renta?

   (b) Contraste al 5% las siguientes afirmaciones

      (i) No existen diferencias en el consumo de cerveza según el sexo.

      (ii) No existen diferencias en la propensión marginal al consumo de cerveza según el sexo.

6. En algunas ocasiones, la observación de datos antes y después de un cambio exógeno de política económica que afecta sólo a un colectivo de la población proporciona un **experimento natural** que permite analizar el efecto de una determinada política en el comportamiento de los agentes. En otras palabras, aunque los datos no sean experimentales, pueden considerarse como tales si el cambio de política afecta sólo a un colectivo (que consideraríamos el grupo experimental) pero no al resto (que consideraríamos el grupo de control). La idea de un experimento natural es que la asignación de los individuos al grupo de tratamiento o de control es exógena a sus acciones, o dicho de otra forma, las razones por las que se encuentran en uno u otro grupo son exógenas o accidentales.
En el caso más simple, hay dos períodos y dos grupos, un grupo A de **control** compuesto por aquellos a los que no afecta el cambio de política y un grupo B de **tratamiento** (o experimental), que contiene a aquellos que sí se ven afectados.
Sean $\overline{Y}_{A1}$ e $\overline{Y}_{A2}$ las medias de Y para el grupo de control en los períodos 1 y 2 respectivamente, e $\overline{Y}_{B1}$ e $\overline{Y}_{B2}$ los análogos para el grupo de tratamiento.
Si los datos se pudieran generar verdaderamente a partir de un experimento puro, entonces podríamos medir el efecto del tratamiento ignorando el primer período (previo al cambio de política) y comparando a los individuos del grupo de tratamiento con los del grupo de control una vez producido el cambio de política. Es decir, evaluaríamos el efecto del cambio de política calculando la diferencia de medias entre el grupo de tratamiento y el grupo de control después del cambio de política,

$$\left(\overline{Y}_{B2} - \overline{Y}_{A2}\right)$$

En la práctica, con datos no experimentales, este estimador tiene el problema de que parte de la diferencia entre la media del grupo de tratamiento y la media del grupo de control después del cambio de política puede deberse a diferencias sistemáticas e inobservables entre ambos grupos que no tienen nada que ver con el cambio de política.
Una medida interesante podría ser el efecto del cambio de política sobre el grupo de tratamiento,

$$\left(\overline{Y}_{B2} - \overline{Y}_{B1}\right).$$

El problema de esta medida es que la media del grupo de tratamiento puede cambiar a lo largo del tiempo (entre ambos períodos) por razones diferentes al cambio exógeno de política económica.
La medida apropiada para capturar el efecto del tratamiento (esto es, del cambio de política económica) consiste en comparar los cambios acaecidos en los grupos de tratamiento y de control, respectivamente, de manera que podamos controlar tanto las diferencias ex ante existentes entre ambos grupos como cambios acaecidos ajenos al cambio de política, es decir:

$$\left(\overline{Y}_{B2} - \overline{Y}_{B1}\right) - \left(\overline{Y}_{A2} - \overline{Y}_{A1}\right).$$

Consideremos las variables binarias *dB*, que toma el valor 1 si el individuo pertenece al grupo de tratamiento y 0 en otro caso, y *d2*, que toma el valor 1 para el segundo período (después del cambio) y 0 para el primero (antes del cambio). En su versión más simple, la ecuación para analizar el impacto de un cambio de política es

$$Y = \beta_0 + \delta_0\ d2 + \beta_1\ dB + \delta_1\ \ (d2 \times dB) + u, \tag{3}$$

donde Y es la variable de interés, y:
d2 es una variable binaria que captura factores agregados que afectan a Y a lo largo del

tiempo de igual manera tanto al grupo de tratamiento como al grupo de control;
dB captura posibles diferencias entre los grupos de tratamiento y de control ex ante (antes del cambio de política).

Nótese que si no tuviéramos en cuenta las diferencias entre los grupos de tratamiento y de control antes del cambio de política, podríamos atribuir incorrectamente parte de dicha diferencia como efecto del tratamiento.

El coeficiente de interés, δ_1, está asociado a la interacción entre ambas variables binarias, *d2* y *dB* (cuyo producto no es más que una nueva variable binaria que es igual a 1 para los elementos del grupo de tratamiento en el segundo período).
Sean $\overline{Y}_{A1}$ e $\overline{Y}_{A2}$ las medias de Y para el grupo de control en los períodos 1 y 2 respectivamente, e $\overline{Y}_{B1}$ e $\overline{Y}_{B2}$ los análogos para el grupo de tratamiento. Entonces, es fácil comprobar que el estimador MCO de δ_1, d_1, se puede expresar como

$$d_1 = \left(\overline{Y}_{B2} - \overline{Y}_{B1}\right) - \left(\overline{Y}_{A2} - \overline{Y}_{A1}\right) \tag{4}$$

Este estimador se denomina estimador de **diferencias-en-diferencias** (DED).
Por supuesto, para que el estimador DED estimador evalúe consistentemente el efecto del cambio de política, es necesario que dicho cambio de política no esté sistemáticamente relacionado con otros factores inobservables (contenidos en u) que afecten a Y.

En julio de 1980, el estado de Kentucky (EE.UU.) aumentó el tope máximo del subsidio por accidente o enfermedad laboral. Dichos subsidios son un porcentaje del salario, con un límite superior (el tope máximo). Por tanto, el aumento del límite superior afectaba tan sólo a aquellos trabajadores con salarios altos, que llegaban al tope máximo del subsidio. Este cambio de política disminuyó el coste de oportunidad de estar de baja laboral para los trabajadores de sueldos altos. El cambio de política nos permite evaluar si un sistema público de subsidios por accidente o enfermedad laboral más generoso tiende a prolongar el período de baja laboral.
El archivo `KENTUCKY.GDT` incluye datos, para el estado de Kentucky, de trabajadores que han sufrido algún accidente o enfermedad laboral. La variable *d2* es igual a 1 para observaciones posteriores al cambio en el tope máximo del subsidio y 0 en otro caso, y *dB* es una variable binaria que es igual a 1 para trabajadores con salarios altos afectados por el tope máximo y 0 para los restantes.

(a) Evalúe el efecto sobre el logaritmo neperiano de la duración de la baja laboral (en días) *ldur* del cambio de política, aplicando el estimador de DED propuesto anteriormente. ¿En qué porcentaje aumentó (o disminuyó) la duración media de la baja laboral tras el cambio de política?

(b) En la mayoría de las aplicaciones, la ecuación (3) incluye factores observables que afectan a Y, permitiendo así la posibilidad de que haya diferencias sistemáticas en dichos factores dentro de cada grupo, y aislando en d_1 el efecto puro del cambio de política. (En ese caso, d_1 no tiene ya una representación tan simple como la de (4), aunque conceptualmente sigue siendo la misma idea).
Reestime dicho efecto controlando también por el sexo del trabajador (*sexo*), su estado civil (*casado*), así como las variables binarias del tipo de lesión o enfermedad (*cabeza*, *cuello*, *brazos*, *tronco*, *lumbares*, *piernas*, *enfocup* –esta última se refiere a dolencias derivadas de la ocupación laboral–) y el logaritmo de la edad (*ledad*).

¿Cómo cambian los resultados? ¿Qué estimación del efecto del cambio de política le parece mejor y por qué?

(c) A la vista del valor del R^2, ¿podemos concluir que los resultados son poco relevantes?