Examen Parcial - Soluciones

Parte 1: Interpretación de coeficientes.

A continuación se muestra una lista de variables disponibles de una hipotética base de datos con individuos en edad laboral que podrían utilizarse para llevar a cabo estimaciones:

- *cig*: número de cigarros que una persona fuma en promedio al día
- *educ*: años de escolaridad completada
- *trab*: horas al día que una persona trabaja (nota: los desempleados tienen trab=0)
- *region*: variable categórica que indica en qué región del país habita el individuo. Sus posibles valores son: sur (=1), centro (=2), norte (=3), costa (=4)
- *seguro*: dummy (=1) si es que el individuo tiene seguro de gastos médicos

En las siguientes preguntas se describe la interpretación de los coeficientes de algunas estimaciones de MCO hechas con estas variables. Partiendo de estas interpretaciones deberás indicar con formato de ecuación la estimación que se está llevando a cabo. Deberás indicar el valor de todos los coeficientes que puedas. Aquellos coeficientes cuyo valor no puedas determinar solo déjalos indicados. Tal vez necesitarás generar nuevas variables partiendo de las variables que están en la lista anterior. De ser asi, deberás indicar la definición de las variables que generaste. En estas preguntas no debes preocuparte por los errores estándar.

Ejemplo: *Caeteris paribus*, un aumento de una hora al día de trabajo esta relacionada a un incremento de 0.2 cigarros que una persona fuma al día.

Respuesta:

$$cig_i = \beta_0 + 0.2\ trab_i$$

1. *(7 puntos) Caeteris paribus*, un aumento de un año de escolaridad está relacionado a una disminución de 8% cigarros que una persona fuma al día. Además, una persona sin escolaridad en promedio fuma 2.7 cigarros al día.

   R: Sabemos que:

   $$\frac{\partial E[\ln(cig_i)|educ_i]}{\partial educ} = -0.08$$
   $$E[\ln(cig_i|educ_i = 0)] = \ln(2.7) = 0.99$$

   Por lo tanto:

   $$\ln(cig_i) = 0.99 - 0.08\ educ_i \tag{1}$$

2. *(7 puntos) Caeteris paribus*, las personas que viven en el centro del país fuman en promedio 3.6 cigarros menos al día que las personas del norte. Las personas del norte fuman en promedio 2.8 cigarros al día menos que las personas de sur. Las personas del sur fuman en promedio 5 cigarros al día más que las personas de la costa.

   R: Hay varias formas de plantear una respuesta correcta dependiendo de qué elijan como grupo de referencia. La respuesta a continuación utiliza *costa* como grupo de referencia:

   Para empezar usamos la variable región para generar las siguientes variables dummy:

$$\begin{aligned} Sur_i &= 1\{region_i = 1\} \\ Centro_i &= 1\{region_i = 2\} \\ Norte_i &= 1\{region_i = 3\} \end{aligned}$$

   Planteo la siguiente ecuación:

$$cig_i = \beta_0 + \beta_1\ Sur_i + \beta_2\ Norte_i + \beta_3\ Centro_i$$

   El párrafo de la respuesta sugiere las siguientes condiciones:

$$\begin{aligned} E[cig_i|Centro] - E[cig_i|Norte] &= \beta_3 - \beta_2 = -3.6 \\ E[cig_i|Norte] - E[cig_i|Sur] &= \beta_2 - \beta_1 = -2.8 \\ E[cig_i|Sur] - E[cig_i|Costa] &= \beta_1 = 5 \end{aligned}$$

   Con esto, podemos deducir que la especificación con valores estimados será:

$$cig_i = \beta_0 + 5\ Sur_i + 2.2\ Norte_i - 1.4\ Centro_i \tag{2}$$

   No podemos deducir el valor estimado de β_0, pero era de esperar porque tenemos 3 condiciones y 4 incógnitas.

3. *(7 puntos) Caeteris paribus*, un aumento de una hora de trabajo está relacionada con una disminución de 0.4 cigarros más para las personas que tienen seguro que para las que no lo tienen. Además, las personas desempleadas sin seguro fuman en promedio 2.5 cigarros menos al día que las personas que si tienen seguro.

   R: Dado que tenemos heterogeneidad del efecto de aumentar una hora de trabajo entre dos grupos (los que sí tienen versus los que no tienen seguro), necesitamos hacer una especificación con interacciones:

$$cig_i = \beta_0 + \beta_1\ trab_i + \beta_2\ seguro + \beta_3\ trab_i * seguro_i$$

Las condiciones descritas establecen las siguientes condiciones:

$$\frac{\partial E[cig_i|seguro\_1, trab_i]}{\partial trab_i} - \frac{\partial E[cig_i|seguro\_0, trab_i]}{\partial trab_i} = -0.4 = \beta_3$$
$$E[cig_i|seguro\_1, trab_i = 0] - E[cig_i|seguro\_0, trab_i = 0] = 2.5 = \beta_2$$

Es importante indicar que en la descripción, la segunda condición establecida arriba no aclara que la comparación es respecto a personas sin seguro que también están desempleados. Sin dicha condición no podría deducirse el valor estimado de β_2, ya que la condición hubiera indicado:

$$E[cig_i|seguro\_1, trab_i = 0] - E[cig_i|seguro\_0, trab_i] = 2.5 = \beta_2 - \beta_1 trab_i$$

La anterior diferencia quiere decir que para deducir el valor 2.5 se habría tenido que definir un valor para $trab_i$.

Suponiendo que es una comparación entre desempleados, lo que sabríamos de nuestra especificación es:

$$cig_i = \beta_0 + \beta_1\ trab_i + 2.5\ seguro - 0.4\ trab_i * seguro_i \tag{3}$$

4. *(7 puntos) Caeteris paribus*, sabes que las personas que más fuman son las que no tienen empleo y las que más horas al día trabajan. Además sabes que, condicional en trabajar tiempo completo (8 horas al dia), un aumento de una hora de trabajo al día conlleva un aumento promedio de 1 cigarro fumado al día. Por último, sabes que de acuerdo a tu estimación, las personas que menos fuman son las que trabajan 4 horas al día.

   R: La descripción de la respuesta coincide con una forma cuadrática convexa con valor mínimo en $trab_i = 4$. Por lo tanto proponemos la siguiente especificación:

$$cig_i = \beta_0 + \beta_1\ trab_i + \beta_2\ trab_i^2$$

La descripción de las respuestas establece las siguientes condiciones:

$$\frac{\partial E[cig_i|trab_i = 4]}{\partial trab_i} = \beta_1 + 2\beta_2(4) = 0$$
$$\frac{\partial E[cig_i|trab_i = 8]}{\partial trab_i} = \beta_1 + 2\beta_2(8) = 1$$

Basado en estas dos ecuaciones podemos deducir los valores estimados de β_1 y β_2 y obtener la siguiente especificación con valores estimados:

$$cig_i = \beta_0 - 1\ trab_i + \frac{1}{8}\ trab_i^2 \tag{4}$$

5. *(7 puntos) Caeteris paribus*, la elasticidad entre cigarros fumados al día y horas trabajadas al día es de -0.4. En esta pregunta en particular, además de indicar la ecuación, expresa este resultado en términos de cambios porcentuales de cigarros fumados y horas trabajadas.

   R: Al ser una elasticidad sabemos que es una relación en la cual tenemos logaritmo en la variable dependiente y explicativa. Con ello tenemos:

$$\ln(cig_i) = \beta_0 - 0.4 \ \ln(trab_i) \tag{5}$$

   Para interpretar a la elasticidad como cambios porcentuales tendríamos que: *caeteris paribus*, un aumento de 1% en las horas trabajadas al día está relacionada con una disminución de 0.4% en los cigarros fumados al día.

Parte 2: Beneficios de trabajar en la misma área para la que estudiaste.

Cristina está interesada en estudiar si existe algun beneficio en trabajar en la misma área que estudiaste, ceteris paribus. Por ejemplo, ella está interesada en la comparación entre dos personas idénticas (misma edad, sexo, experiencia, estado de la República donde vive), trabajando como guías en un museo, solo que una de ellas estudió arte mientras que la otra estudió psicología. A Cristina le interesa ver si la persona que estudió arte tiene una ventaja salarial por ser más productiva que la que no estudió arte. Para esto, Cristina generó una variable dummy llamada match (=1) si es que la persona trabaja en un empleo relacionado con sus estudios. Cristina está interesada, por lo tanto, en estimar la siguiente relación:

$$\ln(ingreso_i) = \beta_0 + \beta_1 \ match_i + \beta_2 \ edad_i + \beta_3 mujer_i + \beta_4 \ exper_i + U_i \tag{6}$$

6. *(4 puntos)* En una primera estimación, Cristina obtuvo que las personas que tienen match=1 ganan 15% mas. ¿Este 15% es un parámetro, un estimador o un valor estimado?

   R: Siendo un número específico y particular a la muestra en cuenstión, es un valor estimado.

7. *(4 puntos)* El 15% indicado en la pregunta anteiror: ¿es una variable aleatoria?

   R: Siendo un número específico, no es una variable aleatoria. La que sí es una variable es el estimador, ya que su valor dependerá de la muestra.

8. *(12 puntos)* Para el resultado de las dos preguntas anteriores, Cristina obtuvo un valor-p igual a 0.02 haciendo un test de significancia (i.e. la hipotesis nula es bilateral). Dado esto, sabríamos que el error estándar de $\widehat{\beta}_1$ es igual a XXX.

Si en cambio su hipótesis nula hubiese sido unilateral: $H_1 : \beta_1 > 0$. Usando el error estándar que acabas de calcular y el resultado de que las personas con match=1 ganan 15% más, el valor-p hubiera sido igual a: XXX.

R: Para deducir el error estándar tendríamos que primero calcular el estadístico-t que deriva el valor-p de 0.02 (usando la distribución de la normal estándar) y a partir de ahí despejar el error estándar correspondiente:

$$\begin{aligned} t &= \frac{\beta_1}{se(\beta_1)} = \frac{0.15}{se(\beta_1)} \\ \Phi(-|t|) &= 0.02 \\ |t| &= -\Phi^{-1}(0.02) = 2.33 \\ se(\beta_1) &= \frac{0.15}{2.33} = 0.06438 \end{aligned} \tag{7}$$

Finalmente, para el valor-p de una prueba unilateral, dado que sabemos el valor-p de una bilateral, que el estadístico t fue positivo (definido por el valor estimado de β_1) y que la distribución normal es simétrica, este simplemente es la mitad del valor-p original: 0.01.

9. *(10 puntos)* Indica por qué no podemos concluir que el resultado indicado es causal. Sugiere una variable omitida que podría estar generando sesgo y explica la lógica de si dicho sesgo es positivo o negativo. Deberás tener cuidado de explicar todas las condiciones que deben cumplirse para que haya sesgo y detallar qué tipo de relaciones deben existir con esta variable omitida para explicar el signo del sesgo (positivo o negativo).

   R: Hay dos condiciones que una variable omitida debería tener para generar sesgo: (a) tendría que estar correlacionada con la variable *match* y (b) tendría que ser un determinante importante del *ingreso* en la ecuación (6).

   Por dar un ejemplo, podríamos pensar en los *networks laborales* de un individuo. Tendríamos un sesgo positivo si las personas que están trabajando en la misma área en la que estudiaron ($match = 1$) construyeron redes laborales más fuertes que aquellos individuos que trabajan en un area distinta a la que estudiaron; y las personas que tienen redes laborales más fuertes tienen un mayor ingreso como resultado de dichos networks. las redes sociales que tienen los individuos.

   Podríamos pensar en otras variables omitidas con las condiciones antes mencionadas, por ello no podemos pensar que una estimación de la especificación (6) vía OLS podría darnos una estimación causal de β_1.

10. *(10 puntos)* Cristina supone que conforme aumenta la experiencia, la ventaja del match se diluye. ¿Cómo modificarías la estimación para reflejar esto? ¿Qué prueba de hipótesis tendrías

que plantear para evaluar si existe ventaja salarial para los que tienen $match = 1$ condicional en tener 10 años de experiencia?

R: Primero debes pensar si tu especificación con respecto a experiencia debería ser un polinomio de primer o segundo grado. En el caso mas sencillo, podemos empezar utilizando uno de primer grado como en la especificación (6), modificándola para que el efecto de experiencia sea distinto entre aquellos que si hacen y aquellos que no hacen match:

$$\ln(ingreso_i) = \beta_0 + \beta_1\ match_i + \beta_2\ edad_i + \beta_3 mujer_i + \beta_4\ exper_i + \beta_5\ match_i * exper_i + U_i$$

En el caso de la especificación, la diferencia salarial entre los que hacen y los que no hacen match dependerá ahora de los años de experiencia:

$$E[\ln(ingreso_i)|match = 1, X_i] - E[\ln(ingreso_i)|match = 0, X_i] = \beta_1 + \beta_5\ exper_i$$

Si hay una desventaja inicial que se diluye a mayor experiencia esperaríamos que $\beta_1 > 0$ y $\beta_5 < 0$. Para evaluar si hay ventaja salarial condicional en tener 10 años de experiencia tendríamos que evaluar:

$$\begin{aligned} H_0 &: \beta_1 + 10\ \beta_5 = 0 \\ H_1 &: \beta_1 + 10\ \beta_5 \neq 0 \end{aligned} \tag{8}$$

11. *(7 puntos)* Cristina está también interesada en complementar con variables de índole política su estimación. Quiere agregar a su estimación la información proveniente de dos variables: (a) el estado de origen del individuo y (b) el partido político que gobierna dicho estado.

    Empecemos SOLO por la variable de partido político. ¿Cómo le sugerirías modificar su estimación para agregar la información de partido político? Deberás mostrar la nueva especificación que le sugieres estimar y definir claramente cualquier variable que agregues.

    NOTA: supón que solo existen 3 partidos políticos: MORENA, PAN y PRI.

    R: Asumiendo que existen 3 partidos políticos, podríamos generar 3 variables dummy que indiquen el partido que gobierna el estado donde se ubica el individuo i. Estas variables dummy serían:

$$\begin{aligned} Morena_i &= 1\{\text{si Morena gobierna el estado del individuo } i\} \\ PAN_i &= 1\{\text{si PAN gobierna el estado del individuo } i\} \\ PRI_i &= 1\{\text{si PRI gobierna el estado del individuo } i\} \end{aligned}$$

Si son mutuamente excluyentes (i.e. no hay alianzas), tendríamos que $Morena_i + PAN_i = PRI_i = 1$. Por lo tanto para evitar multicolinearidad perfecta omitimos una de estas variables en nuestra nueva especificación:

$$\begin{aligned}\ln(ingreso_i) =& \beta_0 + \beta_1\ match_i + \beta_2\ edad_i + \beta_3 mujer_i + \beta_4\ exper_i \\ &+ \beta_5\ Morena_i + \beta_6\ PAN_i + U_i\end{aligned} \tag{9}$$

12. *(10 puntos)* Imagina que ahora quisieras agregar la información de estado de la República de origen del individuo. Una compañera de clase te advierte que eso provocaría un problema de multicolinearidad. Cuando le preguntas si multicolinearidad perfecta medio se hace bolas para contestarte. Explica si agregar estado de la republica provocaría multicolinearidad perfecta o no. Explica claramente por qué y qué le sugerirías hacer para resolver este problema en la práctica.

    R: Hacer esto sí involucraría un problema de multicolinearidad perfecta. Agregar la variable estado corresponde a crear 32*dummies*, una para cada estado, y agregar a la especificación 31 de ellas para evitar multicolinearidad con la constante. Pero si las variables de partido político ya fueron agregadas, esto provocaría la multicolinearidad. Por ejemplo, imaginemos que PAN gobierna en 4 estados: el estado 21, 22, 23 y 24. La multicolinearidad perfecta surge de que:

$$PAN_i = Estado(21)_i + Estado(22)_i + Estado(23)_i + Estado(24)_i \tag{10}$$

    De forma similar se puede mostrar la multicolinearidad perfecta para PRI_i y $Morena_i$. De forma intuitiva, la variación de estado que es más desagregada que la de partido tiene toda la variación que partido pretende agregar, asi que no se pueden utilizar ambas variables.

13. *(8 puntos)* Regresando a su especificación inicial, a Cristina le interesa ver si la variable *match* se comporta de forma distinta dependiendo de la distribución del ingreso. En particular, quiere saber si para los individuos más pobres –aquellos con un ingreso por debajo del percentil 10– la variable match tiene un efecto menor que para los más ricos –aquellos con ingreso mayor al percentil 90–. Un compañero de clase le sugiere utilizar regresiones cuantílicas. ¿Debe seguir este consejo, dado lo que le interesa medir?

    R: Regresiones cuantilicas no te darían dicha respuesta. Esto se debe a que la pregunta que tiene Cristina (si la variable match tiene un efecto menor para los individuos debajo del percentil 10 de ingreso que para los individuos con un ingreso mayor al percentil 90) no establece al ingreso como condicional en las variables explicativas, que es lo que hacen las regresiones cuantílicas.