

# Resolución Examen de Calificación Econometría

Programa en Economía Aplicada  
Universidad Católica del Norte

Instructor: Yasna Cortes  
Office Hours: F 9:55-11:25am, or by appointment

E-mail: ycortes01@ucn.cl  
Class Hours: M-W 9:55 - 11:25am

---

PUNTAJE TOTAL: 100

1. Asuma que tiene una muestra de 1000 observaciones en la forma de  $(y, x_1, x_2)$  del modelo poblacional definido por:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon \quad (1)$$

Para cda uno de los siguientes problemas, discuta si (1) el estimador de  $\beta_1$  insesgado; (2) Si el estimador de  $\beta_1$  es insesgado, indique el efecto de este problema sobre la eficiencia del estimador. Use palabras o demostraciones matemáticas

- (a) Una correlación muestral entre  $x_{i1}$  y  $x_{i2}$  de 0.9. Aquí, tenemos un claro problema de multicolinealidad. En el caso de multicolinealidad perfecta, no podríamos estimar el modelo, ya que el regresor  $x_{i1}$  sería una combinación lineal de  $x_{i2}$ . En el caso de una multicolinealidad imperfecta, aún podríamos estimar los estimados del modelo vía OLS. De esta forma, cualquier estimador insesgado puede verse afectado por problemas de multicolinealidad. En este sentido, los estimadores aún son BLUE con perturbaciones normales. Sin embargo, cuando la multicolinealidad se vuelve extrema, los coeficientes estimados de OLS se vuelven indeterminados, ya que es difícil estimar  $(X'X)^{-1}$ . En adición, los coeficientes de OLS se vuelven infinitamente variables cuando se agudiza la multicolinealidad (las varianzas se vuelven más grandes). Por tanto, el estimador es insesgado, pero el estimador no será eficiente.
- (b) Omitiendo  $x_2$  cuando  $x_{i1}$  y  $x_{i2}$  no están correlacionados en la muestra. Asumiendo que  $x_{i1}$  y  $x_{i2}$  no están correlacionados, la omisión de  $x_2$  puede gatillar en un sesgo por omisión de variable (omitted variable bias). De esta forma, el regresor  $x_2$  está correlacionado con el término omitido, el cual es un determinante de la variable dependiente. En otras palabras, se excluye una variable relevante, subestimando el modelo verdadero presentado en la ecuación (1). Por tanto,  $\hat{\beta}_1$  no se acercará a su valor verdadero. Por tanto,  $\beta_1$  estaría sesgado.
- (c) Un problema en un computador quien aleatoriamente borró la mitad de las observaciones antes de la estimación. En este caso, aún el estimador es insesgado debido a que la selección de los casos es exógena al investigador/a. (proviene de un proceso aleatorio).

- (d) Un ayudante de investigación borró todas las observaciones con  $x_{i1} < 100$  antes de la estimación. En este caso, aún el estimador es insesgado y consistente. Solo estamos reduciendo el tamaño muestral. Por tanto, el estimador es insesgado pero produce estimadores menos eficientes.
- (e) Un ayudante de investigación borró todas las observaciones con  $y_i > 100$  antes de la estimación. En este caso, estamos en presencia del problema de sesgo de selección. Los estimadores serán sesgados. Tome en consideración el siguiente ejemplo: la estimación de un modelo que busca determinar si la educación influye en el salario que un individuo que puede potencialmente obtener en el mercado laboral. Sin embargo, en su muestra usted tiene una alta proporción de desempleados, y por ende, no hay información sobre las ganancias de los individuos. Entonces, estas observaciones no pueden ser usadas en la estimación de la ecuación de salario. De esta forma, usted está estimando la ecuación basado en una muestra no aleatoria. Esta selección no aleatoria puede sufrir del problema de sesgo de selección. Esto puede ser visto como un sesgo de endogeneidad, porque el proceso de selección genera endogeneidad en la sub-muestra seleccionada. Como consecuencia, los coeficientes estimados serán sesgados.
2. Usted dispone de datos para el año 2010 sobre el gasto del gobierno (G), inversión (I), y producto interno bruto (Y) para 30 países. Con estos datos se ha realizado tres regresiones diferentes obteniendo los siguientes resultados, donde entre paréntesis está expresada la desviación estándar de los coeficientes de regresión:

$$\hat{I} = -0.03 - 0.69G + 0.34Y, \quad R^2 = 0.78 \quad (2)$$

(0.28)
(0.16)
(0.03)

$$\hat{I} = 0.39 - 0.03\left(\frac{1}{Y}\right) - 0.45G, \quad R^2 = 0.98 \quad (3)$$

(0.04)
(0.42)
(0.22)

$$\ln(\hat{I}) = -2.44 - 0.63 \ln(G) + 1.6 \ln(Y), \quad R^2 = 0.96 \quad (4)$$

(0.26)
(0.12)
(0.11)

Con esta información, comente la calidad de cada una de las especificaciones. Decida cuál de estos tres modelos representa mejor la relación econométrica que se desea establecer. Justifique claramente con los conceptos técnicos aprendidos en el curso de econometría. <sup>1</sup>

Usando la intuición y lo aprendido en su curso de econometría, debemos buscar un modelo que presente un alto coeficiente de determinación, y coeficientes de regresión que sean significativos, de acuerdo a la relación que se desea estudiar a través del modelo econométrico planteado.

Por tanto, según la información:

- Tenemos el coeficiente  $R^2$ .
- Tenemos información que nos permite obtener los test t y determinar la significancia de los coeficientes de regresión.
- No tenemos información sobre el test F.

<sup>1</sup>Use un nivel de confianza del 95%

Procedemos a obtener el test t. Seguimos la estructura del contraste de hipótesis para determinar si los coeficientes son distintos de cero:

$$H_0 : \beta_i = 0$$

$$H_1 : \beta_i \neq 0$$

El estadístico t se obtiene de la siguiente forma:

$$t = \frac{\beta_i - 0}{\sqrt{\text{var}(\beta)}} \sim t_{(n-k, \alpha)}$$

Lo contrastamos a cero para determinar si esta diferencia es significativa estadísticamente. El denominador aparece en el ejercicio en paréntesis (desviación estándar de cada coeficiente).

De acuerdo a la cantidad de observaciones (30), el estadístico t se compara con un valor  $t = 2.05$ , ya que tenemos 27 grados de libertad.

De esta forma, podemos verificar que el modelo 1 y 2 se descartan, ya que poseen al menos una variable que no podemos determinar que sea estadísticamente significativa, de acuerdo al nivel de significancia aceptado. El modelo 3 si bien poseen un coeficiente de determinación menor al modelo 1 y 2, sus coeficientes son significativos permitiendo representar de mejor forma la relación que se desea establecer en el modelo.

3. (20 puntos) Un/a investigador/a busca explicar los factores que determinarían el comportamiento de votación de los residentes de Antofagasta, de cara a la elección del nuevo/a Gobernador/a Regional. El/la investigador/a ha determinado que la mejor especificación econométrica de esta relación debe ser la siguiente (matricialmente):

$$y = X\beta + \epsilon \quad (5)$$

donde la variable dependiente que refleja el comportamiento de votación del un/a residente de la ciudad está determinada por la situación de la persona, como la educación, ingresos, posición política, etc., cuyas variables se encuentran contenidas en la matriz X.

Sin embargo, según el periódico local, las personas dado su desconocimiento de la votación que se realizará en abril, han justamente comenzado a interactuar en mayor cantidad con sus vecinos, buscando clarificar ciertos elementos claves en la elección del Gobernador/a Regional, como por ejemplo, cómo reemplazarán a los intendentes en sus funciones., y así decidir por quién según ellos, tiene la mejor capacidad para tomar los desafíos del cargo.

De acuerdo al planteamiento anterior, responda las siguientes preguntas.

- Explique porqué un modelo de contagio es apropiado para reflejar esta modelación econométrica. ¿Cómo incorporaría este contagio en la especificación econométrica planteada inicialmente por el/la investigador/a?

Primero, debemos considerar según el caso descrito, las/los residentes están sujetos a las acciones y comportamientos de otras personas. Por tanto, al tomar en consideración opiniones y comportamientos de otros, las personas establecen y modifican su propio comportamiento. Claramente, esto implica un efecto de contagio, ya que su votación está influenciado por su vecindario, amigos, etc. Por tanto, un modelo de contagio es apropiado considerando:

- (a) Situaciones de incertidumbre. Los individuos esperan ver el comportamiento de otros para lidiar con la incertidumbre e imitan su actitud.
- (b) No hay información relevante. Las personas se adaptan a la falta de información tomando cuán significativa es para otras personas es esta falta de información para otros.
- (c) Cuando los actores conocen todos los factores relevantes, pero no están seguros de su importancia relativa. Pueden tomar lo que hacen otras personas.

En econometría espacial, podemos incorporar esta interacción a través de la matriz  $W$ . La matriz  $W$  representa el proceso de influencia que se asume en la red y que puede ser operacionalizado en diferentes formas. Esta especificación depende del investigador/a acerca de la estructura de influencia (proceso) en la red establecido por teoría. De esta forma, de forma sencilla se representa el efecto de contagio, o en otras palabras, como la opinión de otros es una versión ponderada de las opiniones de otros.

Otros requerimientos de la respuesta:

- Definición matriz  $W$ .
  - Interpretación (intuitiva) del modelo.
  - Interpretación del parámetro  $\rho$ .
- ¿Por qué este modelo es inconsistente cuando es estimado vía OLS? ¿Por qué este modelo es sesgado cuando se estima a través de OLS? Explique intuitivamente y matemáticamente la lógica detrás de estos problemas, considerando los problemas que traería incorporar el efecto de contagio a la especificación planteada al inicio del ejercicio.
    - Se viola el supuesto de independencia de las observaciones, ya que los residuos se encuentran espacialmente correlacionados. Si los residuos se encuentran rezagados espacialmente, el estimador será ineficiente y producirá intervalos de confianza incorrectos.
    - En adición, los contrastes de hipótesis asumen que las observaciones son independientes y los residuos no están correlacionados. De esta forma, al violar el supuesto de independencia, los contrastes de hipótesis entregarán conclusiones inadecuadas.
    - Si los residuos están espacialmente correlacionados, pueden subestimar la varianza residual y los contrastes de hipótesis podrían estar sobreestimados.
    - Al introducir endogeneidad al modelo, los estimadores vía OLS no son consistentes y serán sesgados al incluir la matriz  $W$ . A través de la matriz  $W$ , se representa las relaciones de dependencia espacial y su proceso de retroalimentación

simultánea entre observaciones, donde el proceso generador de datos (DGP) indica que el valor esperado de cada observación de la variable dependiente dependerá de la media de las variables explicativas más la combinación lineal de los valores tomados desde las observaciones vecinas, ponderado por el parámetro de dependencia espacial.

(Para simplificar el ejercicio, podemos omitir las variables explicativas.)

Considering that:

$$y = \rho W y + \epsilon \quad (6)$$

The OLS estimate for  $\rho$  is equal to:

$$\rho_{OLS} = [(Wy)'(Wy)]^{-1}(Wy)'y \quad (7)$$

Substituting (1) in (2) and assuming that  $Y_L = Wy$ .

$$\rho_{OLS} = \rho + (Y_L'Y_L)^{-1}Y_L'\epsilon \quad (8)$$

Also, the last expression can be written as:

$$\rho_{OLS} = \rho + \left( \frac{1}{n} \sum_{i=1}^n y_{L_i}^2 \right)^{-1} \left( \frac{1}{n} \sum_{i=1}^n y_{L_i} \right) \epsilon \quad (9)$$

Where  $\frac{1}{n} \sum_{i=1}^n y_{L_i}^2$  is a finite scalar. The second term will be equal to:

$$\frac{1}{n} \sum_{i=1}^n y_{L_i} \epsilon = E(y_{L_i} \epsilon) \quad (10)$$

$$= \text{tr}(C)E[\epsilon\epsilon'] \quad (11)$$

$$\neq 0 \quad (12)$$

Then, this term induces endogeneity and  $\rho_{OLS}$  is inconsistent.

Considering the spatial lag model:

$$y = \rho W y + \epsilon \quad (13)$$

The OLS estimate for  $\rho$  is equal to:

$$\rho_{OLS} = [(Wy)'(Wy)]^{-1}(Wy)'y \quad (14)$$

Substituting (1) in (2) and assuming that  $Y_L = Wy$ .

$$\rho_{OLS} = \rho + (Y_L'Y_L)^{-1}Y_L'\epsilon \quad (15)$$

Assuming that  $W$  is nonstochastic, the mathematical expectation of  $\rho_{OLS}$  is equal to:

$$E(\rho_{OLS}|W) = \rho + (Y_L'Y_L)^{-1}E(Y_L'\epsilon|W) \quad (16)$$

Clearly, the second part of the previous expression is not zero. Then,

$$\begin{aligned} E(Y_L'\epsilon|W) &= E[\epsilon'(I - \rho W)^{-1}W'\epsilon|W] \\ &= E[\epsilon'C'\epsilon|W] \\ &= E[\text{tr}(\epsilon'C'\epsilon)|W] \\ &= E[\text{tr}(C'\epsilon\epsilon')|W] \\ &= \text{tr}(C)E[\epsilon\epsilon'|W] \\ &\neq 0 \end{aligned}$$

where  $C = W(I - \rho W)^{-1}$ .

Here, we need to observe the behaviour of  $C$ . If  $\rho$  is equal to zero,  $C = W$  and  $\text{tr}(C) = \text{tr}(W) = 0$  because the diagonal elements of  $W$  are zeros. If  $\rho$  is different to zero, the OLS estimate of  $\rho$  is biased.