PAUTA EXAMEN

IN3242 - Estadística – Secciones I y II
Departamento de Ingeniería Civil Industrial, Universidad de Chile
Primavera 2022

Profesor Raimundo Undurraga
Auxiliares: Camila Pinto, Felipe Jorquera, Felipe Toloza, Guillermo Morales, Joaquín Cisternas, Rocío Figueroa y Rubén Ortega

Puntaje Total: 100 puntos

Pregunta 1 (40 puntos)

Indique Verdadero (V) o Falso (F). Si es Falso (F), justifique.

(a) (10 puntos) Uno de los supuestos fundamentales para estimar un modelo de regresión lineal vía OLS es que debe ser lineal en los parámetros. Por ejemplo, el modelo $Y = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 \frac{X_2}{X_1} + u$ no puede ser estimado por OLS, porque no es lineal en los parámetros.

Falso. El supuesto de linealidad indica que el moedlo debe ser lineal en los parámetros. De hecho, $Y = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 \frac{X_2}{X_1} + u$ es un modelo lineal en los parámetros y si puede ser estimado vía OLS.

(b) (10 puntos) Suponga el siguiente modelo lineal: $Y_i = \beta Z_i + u_i$, donde (Y, Z) son variables observables, mientras que u_i es el error (no observado). Si $cov[u, Z] = 0$, entonces $E[u|Z] = 0$.

Falso. u y Z son independientes si $E[u|Z] = E[u]$, lo cual implica que u y Z no están correlacionados ($cov[u, Z] = 0$). Sin embargo, que u y Z no estén correlacionados ($cov[u, Z] = 0$) no implica que u y Z sean independientes.

(c) (10 puntos) El teorema de Gauss Markov indica que bajo los supuestos de linealidad, aleatoridad de la muestra, no multicolinealidad, independencia entre el error y las variables independientes, y homocedasticidad, entonces el estimador OLS es, dentro de todos los estimadores insesgados, aquel con mínima varianza (lea bien antes de contestar).

Falso.Gauss Markov indica que OLS efectivamente es el estimador de mínima varianza, pero sólo dentro de la clase de estimadores **lineales e insesgados**.

(d) (10 puntos) Suponga un modelo de regresión lineal $Y = \beta_0 + \beta_1 X_1 + ... + \beta_k X_k + u$ que se estima vía OLS. Mientras menor sea la correlación entre X_k y el resto de las variables independientes del modelo, entonces mayor será la varianza muestral del estimador de β_k.

Falso. Por definición, la varianza muestral del estimador de β_k es $SE(\hat{\beta}_k) = \frac{\sigma}{\sqrt{SST_k(1-R_k^2)}}$.

Sabemos que $R_k^2 = 1 - \frac{\sum \tilde{r}_k^2}{SST_k}$ donde $\tilde{r}_k$ es el residuo obtenido al correr una regresión de X_k sobre el resto de variables independientes del modelo. En efecto, mientras mayor sea la correlación entre X_k y el resto de las variables independientes del modelo, entonces mayor

será el residuo $\tilde{r}_k$. Mientras mayor sea el residuo $\tilde{r}_k$, menor es R_k^2. R_k^2 está entre 0 y 1, así mientras menor es R_k^2, menor es $SE(\hat{\beta}_k)$.

Pregunta 2 (30 puntos).

Considere el siguiente modelo de regresión:

$$Salud_i = \beta_0 + \beta_1 SiFueAlHospital_i + \beta_2 Ingresos_i + \beta_3 SaludPadres_i + e_i$$

Suponga que usted sólo observa las variables Salud, Si el individuo fue al hospital o no, e Ingresos, pero no observa la Salud de los Padres. Suponga además que padres más saludables hacen más conscientes a sus hijos de la importancia de cuidar su salud, y por tanto más conscientes de ir al hospital cuando tengan algún problema de salud. Al mismo tiempo, padres con buenos niveles de salud probablemente tienen altos ingresos, y esos altos ingresos permiten financiar altos niveles de educación a sus hijos. En consecuencia, padres con buenos niveles de salud tienen hijos con altos niveles de ingresos.

(a) (5 puntos) Derive una expresión estadística que denote el problema de sesgo por variable omitida para β_2

Dado que se omite la variable SaludPadres, tendremos lo siguiente:

$$\mathbb{E}(\hat{\beta}_2) = \beta_2 + \beta_3 \frac{Cov(Ingresos, SaludPadres)}{Var(Ingresos)}$$

Al omitir la variable "SaludPadres", el efecto de ésta es capturada por las variables incluidas en el modelo que correlacionan con "SaludPadres"(e.g., "Ingresos"), haciendo que el estimador de β_2 sea sesgado e inconsistente.

(b) (5 puntos) De qué elementos depende la dirección del sesgo? ¿Cuál es la dirección esperada del sesgo? Explique intuitivamente.

El sesgo vendría dado por:

$$\beta_3 \frac{Cov(Ingresos, SaludPadres)}{Var(Ingresos)}$$

Del enunciado sabemos que β_3 sería positivo dado que a mejor salud de los padres, la persona hereda mejor salud. La covarianza entre el ingreso y la salud de los padres sería positiva a su vez ya que padres con buenos niveles de salud probablemente tienen altos ingresos y estos altos ingresos permiten financiar altos niveles de educación de sus hijos, por lo que en consecuencia padres con buenos niveles de salud tienen hijos con altos niveles de ingresos. Finalmente, por construcción Var(Ingresos) es positiva. Con todo lo anterior, el sesgo será positivo (se sobreestimaría β_2).

En términos generales, el sesgo depende de la covarianza entre la variable omitida y la variable que corresponde al parámetro a estimar, de su varianza (la cual siempre es positiva) y el parámetro asociado a la variable omitida.

(c) (10 puntos) Individuos con altos niveles de ingresos tienen menos restricciones presupuestarias para ir al hospital, de forma tal que se puede llegar a tener una fuerte correlación entre ambos aspectos. Qué problema puede generar esto? Explique matemáticamente. Indique en detalle un plan de acción para mitigar ese posible problema.

Una alta correlación entre estas variables generará problemas de multicolinealidad, por lo que en la estimación del parámetro $\beta = (X^\top X)^{-1} X^\top Y$ la matriz $X^\top X$ si bien será invertible, sus valores serán próximos a cero por lo que su inversa será un valor alto. En efecto, la varianza de los estimadores puede ser alta. A su vez, los coeficientes estimados serán sensibles ante cambios pequeños en los datos, entre otros problemas.

Una posible solución para este problema es hacer una transformación a la variable ingreso (como por ejemplo aplicarle logaritmo) o bien acotando su rango.

(d) (10 puntos) Ahora suponga que gracias a la ayuda de una base de datos que le provee un colega usted logra observar la variable Salud de los Padres y decide incluirla dentro del modelo. Otro colega (uno bastante escéptico) le dice que aún teniendo acceso a esos nuevos datos no va a lograr resolver el problema de sesgo por variable omitida. Describa un ejemplo que podría darle la razón a su colega escéptico y explique en detalle la dirección del sesgo en cada caso.

Dado lo anterior, algunos ejemplos serían:

- EducacionPadres: En este caso, dado que padres con mayor educación harían a las y los hijos más conscientes respecto de la prevención y cuidado de ésta, así, las variables ya incluidas en el modelo tendrían un sesgo de la forma:

$$\mathbb{E}(\beta_1) = \beta_1 + \beta_{EducPadres} \frac{Cov(SiFueHospital, EducacionPadres)}{Var(EducacionPadres)}$$

$$\mathbb{E}(\beta_2) = \beta_2 + \beta_{EducPadres} \frac{Cov(Ingresos, EducacionPadres)}{Var(EducacionPadres)}$$

$$\mathbb{E}(\beta_3) = \beta_3 + \beta_{EducPadres} \frac{Cov(SaludPadres, EducacionPadres)}{Var(EducacionPadres)}$$

  De lo anterior observamos que en los tres casos el sesgo sería positivo.

- Edad: En este caso se espera que a mayor edad la salud se vaya deteriorando, teniendo una relación negativa, por lo que tendremos:

$$\mathbb{E}(\beta_1) = \beta_1 + \beta_{Edad} \frac{Cov(SiFueHospital, Edad)}{Var(Edad)}$$

$$\mathbb{E}(\beta_2) = \beta_2 + \beta_{Edad} \frac{Cov(Ingresos, Edad)}{Var(Edad)}$$

$$\mathbb{E}(\beta_3) = \beta_3 + \beta_{Edad}\frac{Cov(SaludPadres, Edad)}{Var(Edad)}$$

De lo anterior observamos que $\hat{\beta}_1$ tendrá un sesgo negativo, en el caso de $\hat{\beta}_2$ tendrá un sesgo negativo y en $\hat{\beta}_3$ tendrá un sesgo positivo.

- IMC: Esta variable puede ser tomada como que a mayor índice de masa corporal hay más riesgos de enfermedades, siendo una dummy igual a cero si se está en el rango de 18.5 a 24.9 mientras que será 1 si no se está en este rango por lo que tendremos:

$$\mathbb{E}(\beta_1) = \beta_1 + \beta_{IMC}\frac{Cov(SiFueHospital, IMC)}{Var(IMC)}$$

$$\mathbb{E}(\beta_2) = \beta_2 + \beta_{IMC}\frac{Cov(Ingresos, IMC)}{Var(IMC)}$$

$$\mathbb{E}(\beta_3) = \beta_3 + \beta_{IMC}\frac{Cov(SaludPadres, IMC)}{Var(IMC)}$$

De lo anterior observamos que $\hat{\beta}_1$ tendrá un sesgo negativo, en el caso de $\hat{\beta}_2$ tendrá un sesgo positivo y en $\hat{\beta}_3$ tendrá un sesgo positivo.

Pregunta 3 (30 puntos).

(a) (15 puntos) Explique en qué consiste MLE. Discuta al menos 2 aspectos que lo diferencian de OLS.

MLE selecciona un set de valores para el conjunto de parámetros θ que caracterizan a un "modelo" dado tal que dicho set de valores maximiza la probabilidad de observar la muestra de estudio. El "modelo" asume que la muestra de estudio es generada a partir de una distribución poblacional determinada. Asumida dicha distribución y el conjunto de parámetros que la caracterizan, la probabilidad de observar la muestra de estudio viene dada por la función de verosimilitud de la muestra, es decir, la función de probabilidad conjunta de cada una de las observaciones contenidas en la muestra. Formalmente, $L(\theta; X_1, ...X_N) = \prod_{i=1}^{N} f_X(X_i; \theta)$, donde $f_X(X_i; \theta)$ representa la función de densidad de X dada la distribución poblacional asumida. En efecto, MLE lo que hace es encontrar los valores del vector θ que maximizan $L(\theta; X_1, ..., X_N)$. [5 puntos]

Dos diferencias fundamentales entre MLE y OLS son, entre otras, las siguientes:

(i) MLE asume una distribución poblacional desde la cual proviene la muestra de estudio. OLS es agnóstico al respecto y no impone restricción alguna a dicha forma funcional. [5 puntos]
(ii) MLE permite estimar modelos no lineales en los parámetros de interés. OLS, en cambio, es un método que tiene entre sus supuestos que el modelo estimado es lineal en los parámetros de interés. [5 puntos]

(b) (15 puntos) Una de las ventajas del estimador de Máxima Verosimilitud (MLE) es que al asumir que los errores del modelo se distribuyen normal, entonces no es necesario asumir

que la distribución poblacional desde la cual proviene la muestra de estudio tiene una forma funcional definida. Comente.

Falso. Cuando asumimos que los errores del modelo se distribuyen normal, lo que estamos asumiendo es que la distribución poblacional desde la cual proviene la muestra de estudio es normal.