Universidad de Chile
Facultad de Ciencias Físicas y Matemáticas
Departamento de Ingenieria industrial
IN-3242 – Estadística

Auxiliar pre examen

Repaso de contenidos escenciales

Profesor: Raimundo Undurraga
Auxiliares: Brandon Galarza, Camila Jáuregui Leonardo Meneses, Francisca Monetta, Matias Reyes, Bastian Urzua, Antonia Villegas

Comentes

1. El teorema de Gauss-Markov nos dice que siempre el estimador de Mínimos Cuadrados Ordinarios será el mejor estimador lineal insesgado
2. El método EMV sirve para estimar cualquier parámetro de cualquier distribución, ya que encuentra un máximo en la curva de verosimilitud.
3. Si la variable dependiente es binaria [0, 1], MCO siempre entregará predicciones dentro de ese rango.
4. La estimación de máxima verosimilitud siempre produce estimaciones insesgadas.
5. Se tiene el siguiente modelo $Y = \beta_0 + \beta_1 X_1 + ... + \beta_k X_k + \mu$, 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 βk.
6. La eficiencia del estimador de máxima verosimilitud siempre es la mejor posible entre los estimadores insesgados.

P1.- OLS

Dado el contexto nacional, usted como estudiante está interesado en aportar en el debate de las pensiones. Para tener una opinión informada, usted se consigue los datos de la Superintendencia de Pensiones para analizar qué variables de los pensionados tiene un mayor efecto en el monto de su pensión. Para entender lo que está ocurriendo, usted propone el siguiente modelo:

$$ln(Monto\_Pension) = \beta_0 + \beta_1 Sexo + \beta_2 AñosCotizados + \beta_3 EdadInicio + \epsilon$$

Donde cada variable significa: $\ln(Monto\_Pension)$ = logaritmo natural del monto de la primera pensión en pesos considerando el aporte del pilar solidario, Sexo = el sexo del cotizante (1 es hombre y 0 es mujer), AñosCotizados = número de años cotizados, EdadInicio = edad en la que empezó a cotizar.

Los resultados que obtiene son:

Source	SS	df	MS	Number of obs	=	464,249
				F(3, 464245)	=	71883.98
Model	64986.7865	3	21662.2622	Prob > F	=	0.0000
Residual	139900.399	464,245	.301350363	R-squared	=	0.3172
				Adj R-squared	=	0.3172
Total	204887.186	464,248	.441331327	Root MSE	=	.54895

MontoPension	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
Sex	.2686708	.0017014	157.91	0.000	.2653361	.2720054
AnosCotizados	.0286848	.0000926	309.63	0.000	.0285033	.0288664
EdadInicio	-.0077055	.000076	-101.44	0.000	-.0078544	-.0075566
_cons	16.22094	.0036753	4413.47	0.000	16.21373	16.22814

Figura 1: Resultados de la regresión

1. Testee la significancia individual de al menos 2 parámetros del modelo. Explicite el test y el método utilizado.
2. Para cada variable, interprete el valor obtenido. ¿Qué puede deducir al respecto?, ¿Qué variables aportan a mejorar las pensiones?
3. [Propuesto] Un compañero de trabajo le dice que existe alta correlación en el modelo. Identifique al menos un par de variables independientes que pudiesen estar correlacionadas. ¿Cómo afecta esto a la estimación del modelo?
4. [Propuesto] Otro compañero le dice que hay variables relevantes que no están incorporadas al modelo. ¿Qué variable podría ser? ¿Qué implicancias tiene esto en las estimaciones que hizo? Explicite matemáticamente el efecto.
5. Finalmente, usted quiere saber si el modelo efectivamente tiene relevancia. Para ello, fíjese en el R^2 y además realice un test de significancia global. ¿Qué concluye?

P2.- Inferencia Causal

Un grupo de estudiantes del DII deses aumentar su productividad, teorizando que el mayor problema de la baja productividad es el uso de RRSS.

Se ofrece a los 1.800 industriales acceder al estudio y NO usar RRSS, recibiendo como recompensa por la participación 1 décima en cualquier exámen. En la práctica, alrededor de un 10 % de los estudiantes decide participar.

Se accede a una base de datos donde se tiene la variable de productividad se mide en una escala continua de 0 a 100, y un indicador dicotómico que indica si NO usó RRSS. El modelo general que se construye es el siguiente:

$$Productividad_i = \beta_0 + \beta_1 NoUsaRRSS_i + \beta_2 VaAclases_i + \beta_3 HorasDeSueño_i + \epsilon_i$$

Los resultados de las estimaciones fueron las siguientes:

Table 1: Modelos lineales de Productividad

VARIABLES	**Modelo 1**	**Modelo 2**	**Modelo 3**	**Modelo 4**
NO usa RRSS = 1	13.98***	13.87***	8.310***	8.349***
	(1.073)	(1.095)	(1.308)	(1.315)
Va a Clases = 1		2.395		-0.012
		(0.809)		(0.809)
Horas de sueño			4.905***	5.026***
			(0.391)	(0.393)
Constante	58.64***	58.35***	49.38***	49.50***
	(0.361)	(0.620)	(1.275)	(1.374)
Observaciones	1,800	1,800	1,800	1,800
R cuadrado	0.019	0.019	0.025	0.025

Robust standard errors in parentheses
***p<0.001, **p<0.05, *p<0.1

1. Describa el mecanismo por el cual se explica que al agregar el regresor de Horas de sueño cambie el coeficiente de NoUsaRRSS, y la dirección de ese efecto. De una posible explicación de por qué esto no ocurre para el caso de la variable de Va a Clases.

2. De forma general, describa detalladamente por qué estos estimadores no deberían tomarse como los efectos causales del experimento. Explique para este contexto específico cuál es el problema que impide obtener una estimación limpia del impacto de RRSS sobre su productividad.

3. Explique formalmente, cuál es el rol del término de error en el problema que anteriormente describió y qué información podría contener ε_i

4. Te encomiendan rediseñar el experimento de tus colegas DII, ¿cuál sería la clave, en términos estadísticos, para poder obtener su efecto causal respecto de la productividad de los estudiantes?

P3.- **MLE- Caso continuo**

Sea Y_1,..., Y_n una muestra aleatoria cuya función de densidad está dada por:

$$f(x|y,\alpha) = \begin{cases} \frac{1}{\Gamma(\alpha)\theta^{\alpha}} y^{\alpha-1} e^{\frac{-y}{\theta}} \quad si \quad y > 0 \\ \\ 0 \sim \end{cases}$$

y que cumple $\mathbb{E}(Y) = \alpha\theta$ y $\mathbb{V} = \alpha\theta^2$, con $\alpha > 0$ conocido.

1. Encuentre el estimador de máxima verosimilitud para θ.
2. Calcule $\mathbb{E}(\hat{\theta})$.

Ejercicios propuestos

MLE - Caso discreto

1. Suponga que tiene dos monedas en su bolsillo. Una de ellas es equilibrada, mientras que la otra no lo es. La probabilidad de observar cara con la moneda equilibrada es 0,5 mientras que con la no equilibrada la probabilidad de observar cara es 0,8. Identificando cara y sello con 1 y 0, respectivamente. Suponga ahora que saca una de las monedas de su bolsillo y deberá determinar si se trata de la moneda equilibrada o de la otra, pudiendo tirarla muchas veces en forma independiente. Suponga que en n = 100 lanzamientos observa la muestra. Plantee la función de verosimilitud para ambas monedas.

$$\underbrace{1,\ldots,1}_{\text{12 veces}}\underbrace{0,\ldots,0}_{\text{5veces}}\underbrace{1,\ldots,1}_{\text{23 veces}}\underbrace{0,\ldots,0}_{\text{8veces}}\underbrace{1,\ldots,1}_{\text{15 veces}}\underbrace{0,\ldots,0}_{\text{3 veces}}\underbrace{1,\ldots,1}_{\text{11 veces}}\underbrace{0,\ldots,0}_{\text{4 veces}}\underbrace{1,\ldots,1}_{\text{13 veces}}\underbrace{0,\ldots,0}_{\text{6 veces}}$$

2. Suponga ahora que desconoce por completo qué tipo de monedas tiene en su bolsillo. Siendo que en la muestra de n = 100 lanzamientos observó 74 caras. Calcule $\hat{\theta}$ el estimador de máxima verosimilitud para la obtención de caras y siga todos los pasos.

Más comentes

1. Sean dos muestras de variables aleatorias de tamaño n: A = (A_1, A_2, .., A_n) y B = (B_1, B_2, .., B_n), ambas independientes e idénticamente distribuidas en su respectivo

conjunto. Si se asume que tanto las variables de A y B distribuyen Exponencial de parámetro λ, el cual es desconocido pero se puede estimar por el método de máxima verosimilitud. Se aplica este procedimiento sobre cada muestra de manera separada, es decir, se maximiza $L(\lambda|\text{A})$ y $L(\lambda|\text{B})$. Y resulta que usando los datos de A se tiene un EMV equivalente λ_A=1 y si se usan las variables B se tiene que el EMV es igual a λ_B=0.5. ¿Cómo explica esta diferencia desde el punto de vista del EMV?

2. El teorema de Gauss-Markov indica que bajo los supuestos de MCO, entonces el estimador OLS es, dentro de todos los estimadores, aquel con mínima varianza.

3. El método EMV sirve para estimar cualquier parámetro de cualquier distribución, ya que encuentra un máximo en la curva de verosimilitud.

4. El modelo Logit se utiliza porque no podemos estimar decisiones discretas usando el modelo de regresión lineal.

5. Si se desea diseñar un modelo con OLS para evaluar el efecto de un tratamiento, ¿Qué elementos estadísticos son clave para obtener el efecto causal de este?

6. Se estima un modelo Probit para encontrar la probabilidad de que un individuo n escoja usar transporte público durante la pandemia ($Y_n = 1$) versus usar transporte particular ($Y_n = 0$). La función de probabilidad estimada tiene la siguiente forma: P(Y_n) = -0.07 - 2.798 $\cdot$ Wagen + ... Este modelo nos indica que si aumenta el salario de una persona en una unidad entonces disminuye en un 2.7 % la probabilidad de que utilice el transporte público.

Modelos de elección discreta

Una cadena de supermercados dispone de un panel de clientes que registran rutinariamente sus compras en la tienda. En lo que sigue estudiaremos el comportamiento de compra en la categoría ketchup donde hay dos marcas principales: Heinz, Kroger y Hunts. El panel posee información del comportamiento de 447 hogares que realizan 4.887 compras en esta categoría durante 138 semanas de actividad. Más específicamente, el panel describe para cada ocasión de compra la marca elegida, los precios actividad promocional de cada una de las marcas e información demográfica de los panelistas.

Teniendo en cuenta los datos, responda:

1. Formule un modelo de elección discreta que permita obtener la probabilidad de que una persona elija la marca Kroger bajo los atributos disponibles en los datos.

2. Exprese las probabilidades de acuerdo los modelos Logit y Probit.

Resumen

OLS (MCO): Método para obtener los coeficientes de una regresión lineal, que consiste en minimizar la suma de los errores al cuadrado. Ello entrega las expresiones conocidas de cada coeficiente:

- $\hat{\beta}_0 = \bar{y_i} - \beta_1 \bar{X}_{1i} - ... - \beta_k \bar{X}_{ki}$
- $\hat{\beta}_1 = \beta_1 + \beta_2 \frac{Cov(X_{1i}, X_{2i})}{Var(X_{1i})} + ... + \beta_k \frac{Cov(X_{1i}, X_{ki})}{Var(X_{1i})} + \frac{Cov(X_{1i}, \epsilon_i)}{Var(X_{1i})}$
- ...
- $\hat{\beta}_k = \beta_k + \beta_1 \frac{Cov(X_{ki}, X_{1i})}{Var(X_{ki})} + ... + \beta_{k-1} \frac{Cov(X_{ki}, X_{(k-1)i})}{Var(X_{ki})} + \frac{Cov(X_{ki}, \epsilon_i)}{Var(X_{ki})}$

Supuestos del MCO:

1. Linealidad: la relación entre la variable dependiente y los regresores debe ser lineal.
2. Aleatoriedad: cada dato i debe ser elegido de manera aleatoria.
3. No multicolinealidad: los regresores NO pueden estar relacionados de forma lineal, esto quiere decir que su correlación no debe ser perfecta.
4. Independencia: los valores de los errores para una observación i particular, no depende del valor del error de ninguna otra observación. Además los errores deben ser independientes a todos los regresores,es decir: $E[\epsilon | X_1, X_2, ..., X_k] = 0$
5. Homocedasticidad: quiere decir que la varianza siempre es igual, es decir, para cualquier combinación de regresores, los datos se dispersan de igual manera (homo=igual, cedasticidad=peso).
6. Normalidad del residuo: se espera que los residuos se comporten como una distribución normal $N(0, \sigma^2)$

Teorema de Gauss-Markov: Bajo los supuestos (A1)-(A5) y dentro de la familia de estimadores lineales e insesgados, asegura que el estimador que entrega MCO es aquel con la mínima varianza posible.

Inferencia causal: Estudia el porcentaje de efecto deseado, atribuible a un tratamiento. Para encontrarlo se debe excluir cualquier factor que no sea este tratamiento.

MLE: método estadístico utilizado para estimar los parámetros desconocidos de un modelo probabilístico. La idea fundamental detrás del EMV es encontrar aquellos valores de los parámetros que maximizan la verosimilitud de observar los datos que se han recopilado.

La verosimilitud representa la probabilidad de observar los datos que se han obtenido, dados ciertos valores de los parámetros en el modelo. El EMV busca encontrar los valores de los parámetros que hacen que los datos observados sean más probables bajo el modelo en consideración.

El proceso básico para utilizar el estimador de máxima verosimilitud implica:

1. **Definir el modelo probabilístico:** Esto implica establecer la distribución de probabilidad que describe los datos y depende de uno o varios parámetros desconocidos.
2. **Formular la función de verosimilitud:** Esta función mide la probabilidad de observar los datos dados los parámetros del modelo. Se expresa como el producto (o la suma en el logaritmo) de las densidades de probabilidad de cada observación bajo el modelo.
3. **Maximizar la función de verosimilitud:** Se encuentran los valores de los parámetros que maximizan esta función, típicamente utilizando técnicas de optimización matemática.
4. **Obtener los estimadores de los parámetros:** Los valores resultantes de la maximización de la función de verosimilitud se utilizan como estimaciones de los parámetros del modelo.

Modelos de elección discreta: modelos que se utilizan para predecir el resultado de una **variable categórica** que se asume dependiente de **variables explicativas** de su comportamiento.

Los modelos de elección discreta Logit y Probit son dos enfoques comunes en la modelización de elección discreta en estadística y econometría. Ambos modelos se utilizan para analizar y predecir decisiones discretas donde una persona o unidad debe elegir entre dos o más alternativas mutuamente excluyentes.

Modelo Logit

El modelo Logit se basa en la función logística y se usa para modelar la probabilidad de que un individuo elija una de las alternativas en función de las características observadas. En particular, asume que la probabilidad de elegir una opción sobre otra se relaciona con un conjunto de variables independientes a través de la función logística.

La función de probabilidad en un modelo Logit se expresa de la siguiente manera:

$$P(Y_i = 1|X_i) = \frac{e^{\beta \cdot x_i}}{1 + e^{\beta \cdot x_i}}$$

Donde:

- $P(\mathcal{Y}_i = 1|X_i)$ es la probabilidad de que la alternativa Y_i sea elegida dadas las variables explicativas X_i.
- β son los coeficientes del modelo.
- X_i son las variables explicativas.

Modelo Probit:

El modelo Probit, por otro lado, se basa en la función de distribución normal acumulativa (la función probit) y modela la probabilidad de elección en función de las variables explicativas utilizando esta función.

La función de probabilidad en un modelo Probit se expresa de la siguiente manera:

$$P(Y_i = 1|X_i) = \Phi(\beta \cdot X_i)$$

Donde:

- $P(Y_i = 1|X_i)$ es la probabilidad de que la alternativa Y_i sea elegida dadas las variables explicativas X_i.
- Φ es la función de distribución acumulativa de la distribución normal estándar.
- β son los coeficientes del modelo.
- X_i son las variables explicativas.