

Modelos no paramétricos y de regresión/Estadística II | Semestre 2018-2

Tareas 10

Fecha de entrega: 25 de mayo

1. Suponer un modelo RLM con matriz de diseño \mathbf{X} con p variables y n observaciones. Responder lo siguiente.

- a) Mostrar que la matriz $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ es simétrica e idempotente.
- b) Mostrar que la matriz $\mathbf{I}_n - \mathbf{H}$ es simétrica e idempotente.
- c) Mostrar que la matriz $\mathbf{H} - \frac{1}{n}\mathbf{J}$ es simétrica e idempotente.

2. Suponer un modelo RLM con p variables y n observaciones. Mostrar las siguientes igualdades.

- a) $SC_{TC} = \sum_{i=1}^n (Y_i - \bar{Y}_n)^2 = \mathbf{Y}'(\mathbf{I}_n - \frac{1}{n}\mathbf{J}_n)\mathbf{Y}$.
- b) $SC_{reg} = \sum_{i=1}^n (\hat{Y}_i - \bar{Y}_n)^2 = \mathbf{Y}'(\mathbf{H} - \frac{1}{n}\mathbf{J}_n)\mathbf{Y}$.
- c) A partir de los incisos anteriores, justificar que

$$SC_{TC} = SC_{reg} + SC_{error}.$$

- d) Mostrar que $(\mathbf{I}_n - \mathbf{H})(\mathbf{H} - \frac{1}{n}\mathbf{J}_n) = \mathbf{0}_{n \times n}$.
- e) A partir del resultado anterior, justificar que $SC_{reg} \perp SC_{error}$.

3. Suponer un modelo RLM con p variables y n observaciones.

- a) Mostrar que $\mathbf{X}'(\mathbf{I}_n - \mathbf{H}) = \mathbf{0}_{(p+1) \times n}$.
- b) A partir del inciso anterior, justificar que la suma de los residuos es 0, es decir,

$$\sum_{i=1}^n \hat{e}_i = 0$$

donde \hat{e}_i es la i -ésima componente del vector $\hat{\mathbf{e}} = \mathbf{y} - \hat{\mathbf{y}}$.

Hint: Expresar la suma de los residuos como $\mathbf{a}'\mathbf{e}$, para algún vector \mathbf{a} adecuado, ¿qué relación tienen \mathbf{a} y \mathbf{X} ?

4. En el siguiente cuadro se muestran los resultados de un análisis de varianza para un modelo RLM.

F.V.	G.L.	S.C.	C.M.	F
Reg.	3	X	1600.81	X
Error	36	146.9	X	
TC	X	X	X	

Responder lo siguiente.

- a) Completar la información de la tabla anterior. Únicamente las celdas marcadas con X.
- b) ¿Con cuántas variables explicativas y cuántas observaciones se ajustó el modelo?
- c) Si se toma $\alpha = 0.01$, ¿el modelo ajustado es significativo?
- d) Estimar puntual y por intervalo (de confianza 99%) la varianza del modelo.
- e) Calcular los coeficientes R^2 y R^2 -ajustado del modelo.
5. El conjunto de datos **Publicidad** contiene: cuatro variables **Ventas**, **TV**, **Radio** e **Impresos** que corresponden a las ventas semanales y gasto en publicidad de un determinado producto en 200 mercados diferentes. El objetivo es modelar las ventas semanales a partir del gasto en publicidad, ambas variables expresadas en miles de USD.
- a) Ajustar un modelo RLM para explicar las ventas semanales a partir del gasto en publicidad. Reportar las estimaciones de β , σ^2 y $V(\hat{\beta})$.
- b) Interpretar $\hat{\beta}_1$, $\hat{\beta}_2$ y $\hat{\beta}_3$ en el contexto de los datos.
- c) Construir la tabla ANOVA y probar la significancia del modelo. Interpretar los resultados. Utilizar $\alpha = 0.01$.
- d) Construir intervalos de confianza 99% para las componentes de β , individuales y simultáneos (Bonferroni y Hotelling-Scheffé). Comparar las longitudes de los intervalos.
- e) Probar la significancia del modelo utilizando los intervalos de confianza simultáneos del inciso anterior y comparar los resultados con el inciso 3. Utilizar $\alpha = 0.01$.
- f) Calcular el R^2 con y sin ajuste e interpretar.
- g) Contrastar si el efecto de **TV** es el doble que el efecto de **Radio**. Interpretar los resultados en el contexto de los datos.
6. Se tiene interés en explicar el Producto Interno Bruto per capita de las entidades del país a partir de algunas variables utilizadas en el Índice de Rezago Social del CONEVAL. Un análisis exploratorio de los datos sugiere utilizar el modelo

$$\log \text{PIBpp} = \beta_0 + \beta_1 \text{EBIN} + \beta_2 \text{NDRE} + \beta_3 \text{NLAV} + \epsilon \quad (1)$$

donde:

- $\log \text{PIBpp}$ es el logaritmo de PIB per capita.
- **EBIN** es el porcentaje de población de 15 y más años con educación básica incompleta.
- **NDRE** es el porcentaje de viviendas particulares habitadas que no disponen de drenaje.
- **NLAV** es el porcentaje de viviendas particulares habitadas que no disponen de lavadora.

Utilizar los datos en el archivo [rezago.csv](#) para ajustar el modelo (1).

- a) Reportar las estimaciones de β , σ^2 y $V(\hat{\beta})$.
- b) Interpretar $\hat{\beta}_1$, $\hat{\beta}_2$ y $\hat{\beta}_3$ en el contexto de los datos.
- c) Probar la significancia del modelo (1). Interpretar los resultados. Utilizar $\alpha = 0.05$.
- d) Construir intervalos de confianza 95 % para las componentes de β , individuales y simultáneos (Bonferroni y Hotelling-Scheffé). Comparar las longitudes de los intervalos.
- e) Probar la significancia del modelo (1) utilizando los intervalos de confianza simultáneos del inciso anterior y comparar los resultados con el inciso 3. Utilizar $\alpha = 0.05$.
- f) Calcular el R^2 con y sin ajuste e interpretar.
- g) Contrastar si el efecto de EBIN y NDRE es el mismo, es decir, contrastar la hipótesis $H_0 : \beta_1 = \beta_2$. Interpretar los resultados en el contexto de los datos.
Hint: esta es una prueba t para $\mathbf{a}'\beta$, con $\mathbf{a} = (0, 1, -1, 0)$.