Modelos no paramétricos y de regresión/Estadística II Semestre 2018-1

Tarea 3

Fecha de entrega: 11 de septiembre

1. En el modelo RLS con errores normales, encontrar la prueba más potente para contrastar las hipótesis
$$H_0 : \beta_0 = 0 \qquad \text{vs.} \qquad H_1 : \beta_0 \neq 0.$$
En este par de hipótesis, cuando H_0 es cierta, el modelo se reduce al de RLS *sin intercepto* y en ejercicios anteriores ya se ha calculado los EMV de este modelo. Bajo H_1 el modelo es el de RLS habitual y ya se ha visto cuáles son los EMV.

2. En el modelo RLS, mostrar la igualdad
$$R^2 = r_{xy}^2,$$
donde R^2 es el coeficiente de determinación del modelo y $r_x y$ es el coeficiente de correlación de las observaciones (x_i, y_i), $i = 1, \ldots, n$.

   El coeficiente de determinación R^2 se expresa como cociente de dos sumas de cuadrados, para mostrar la igualdad conviene recordar algunas expresiones equivalentes a alguna de éstas.

3. Suponer que se ajusta un modelo RLS con errores normales a las observaciones (x_i, y_i), $i = 1, \ldots, n$. y sea (x^*, y^*) una nueva observación.

   a) Mostrar que $\hat{\mu}^* = \hat{\beta}_0 + \hat{\beta}_1 x^*$ es un estimador insesgado de $E(Y|x^*)$.

   b) Calcular la varianza del estimador $\hat{\mu}^*$.

   Como se asume que x^* es fijo, basta aplicar las propiedades de la esperanza y la varianza y los resultados que ya se conocen acerca de $\hat{\beta}_0$ y $\hat{\beta}_1$

4. Sea $(\Omega, \mathcal{R}, P)$ un espacio de probabilidad y $E_1, \ldots, E_n \in \mathcal{F}$ eventos. Probar que se cumple la siguiente desigualdad
$$P\left(\bigcap_{k=1}^{n} E_k\right) \geq 1 - \sum_{k=1}^{n} P(E_k^c),$$
donde E_k^c es el complemento de $E_k$relativo a Ω, $k = 1, \ldots, n$.

   En esta desigualdad se basa el método de Bonferrioni para construir intervalos de confianza simultáneos.

5. Con los datos de desarrollo humano de las entidades del país (que pueden descargar aquí), considerar un modelo RLS para explicar la esperanza de vida con el logaritmo del ingreso.

   a) ¿Hay algún efecto del ingreso en la esperanza de vida? Plantear el juego de hipótesis pertinente y contrastar con $\alpha = 0.05$.

   b) ¿Se debe estimar un modelo con o sin intercepto? Plantear el juego de hipótesis pertinente y contrastar con $\alpha = 0.05$.

   c) ¿La varianza del modelo es menor a 2.5? Plantear el juego de hipótesis pertinente y contrastar con $\alpha = 0.05$.

   d) Estimar la media de los años de esperanza de vida para un ingreso anual de $35,000$ USD PPC.

   e) Construir intervalos de confianza para la estimación anterior e interpretar los resultados. Reportar los errores estándar estimados y los cuantiles utilizados.

f) Construir intervalos de predicción para una nueva observación dado un ingreso anual de $35,000$ USD PPC.

g) Calcular SC_{TC}, SC_{reg} y SC_{error}.

h) ¿Qué tan bueno es el ajuste del modelo? Calcular el coeficiente R^2 a partir de las sumas del inciso anterior.

Para hacer más sencillas las interpretaciones, utilizar el logaritmo base 2.

6. Utilizar los resultados del ejercicio anterior para responder lo siguiente.

   a) Calcular intervalos de simultáneos de confianza 90 % para β_0 y β_1, con el método de Bonferroni.

   b) Calcular intervalos simultáneos de confianza 90 % para β_0 y β_1, con el método de Hotelling-Scheffé.

   c) Comparar la longitud de los intervalos simultáneos, ¿cuáles son mejores?

7. Utilizar los resultados del ejercicio 5 para responder lo siguiente. Suponer que se quiere hacer inferencia sobre la media de la esperanza de vida $E(Y \mid x)$, para niveles de ingreso $x = 10000$, 15000, 20000, 25000, 30000 USD PPC anuales.

   a) Estimar puntualmente $E(Y \mid x)$ y su error estándar, para los valores de x que se indican en el enunciado.

   b) Construir intervalos simultáneos de confianza 90 % para las cinco medias del enunciado, utilizando el método de Bonferroni.

   c) Construir intervalos simultáneos de confianza 90 % para las cinco medias del enunciado, utilizando el método de Hotelling-Scheffé.

   d) Comparar las longitudes de los intervalos simultáneos, ¿cuáles son mejores? ¿la respuesta coincide con la del ejercicio anterior?

   El modelo ajustado se hizo con el logaritmo base 2 del ingreso, por lo que aún hace falta transformar los valores de X que se dan en el enunciado.

8. Se ajustó un modelo de regresión lineal simple a un conjunto de datos y se obtuvo la siguiente tabla ANOVA

FV	GL	SC	CM	F	$P(> F)$
Regresión	1	X	20.11	X	X
Error	X	92.62	X		
Total	20	112.7			

   Además se calculó $S_{xx} = 770.0$. Responda lo siguiente.

   a) Completar la información de la tabla anterior. (Sólo las celdas marcadas con X).

   b) ¿Cuántas observaciones se utilizaron en el ajuste?

   c) Hacer el contraste de $H_0 : \beta_1 = 0$. Considerar $\alpha = 0.1$.

   d) Estimar a σ^2 puntualmente y por intervalo. Considerar 90 % de confianza.

   e) Estimar $|\beta_1|$ y calcular estimar el error estándar del estimador.

   f) ¿Qué porcentaje de la variabilidad es explicada por el modelo?

   Este ejercicio permite entender la relación que hay entre las entradas de una tabla ANOVA. La última columna corresponde al *p-value*, que es la probabilidad de observar un valor del estadístico F más extremo que el calculado, bajo el supuesto que H_0 es cierta.