PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE
EPG4506 - Aprendizaje Estadístico y Computacional
Profesor: Jonathan Acosta

Clase en vivo 1
Solución de ejercicios

1. Se dispone de datos anuales durante un período de 14 años para las variables:

- Consumo (C),
- Exportaciones (Ex),
- Oferta Monetaria (OM),

de la economía de un determinado país. Al ajustar un modelo de regresión a las variables Consumo y Oferta Monetaria, se obtuvo:

$$\begin{array}{rcccc} C_i & = & 851.3 & + & 0.7945 \cdot OM_i \\ & & (15.52) & & (11.60) \end{array}$$

Además, se obtuvo $R^2 = 0.9182$ y $\text{SCError}_1 = 143868.69$.

Posteriormente, se ajustó el modelo de regresión con todas las variables, obteniendo:

$$\begin{array}{rcccccc} C_i & = & 655.8 & + & 3.359 \cdot Ex_i & + & 0.0556 \cdot OM_i \\ & & (11.39) & & (4.282) & & (0.312) \end{array}$$

Además, se obtuvo $R^2 = 0.969$ y $\text{SCError}_2 = 53938.8$.

En ambos modelos las cifras entre paréntesis corresponden a los valores de los estadísticos t-Student.

Si, además, se sabe que la inversa de la matriz $\boldsymbol{X}^\top \boldsymbol{X}$, donde el orden corresponde al modelo 2, es

$$(\boldsymbol{X}^\top \boldsymbol{X})^{-1} = \begin{pmatrix} 0.6758016 & -0.007303 & 0.001341 \\ -0.007303 & 0.000125 & -0.0000276 \\ 0.001341 & -0.0000276 & 0.0000064 \end{pmatrix}$$

a) Realice un análisis de significancia, para determinar si se justifica la incorporación de la variable Exportaciones.

b) En ambos modelos estime σ^2.

c) Suponga que $Ex = 275$, y $OM = 1000$. Utilizando ambos modelos y ésta información, estime el consumo C. Compare los intervalos de confianza de las predicciones realizadas por ambos modelos.

d) ¿Usted propondría alguno de estos dos modelos? Justifique su respuesta.

a) Asumiendo que residuos son iid normales de media 0 y varianza σ^2. Además, como $n = 14$, el modelo 1 tiene una covariable y el modelo 2 tiene 2 covariables, elentonces

$$F = \frac{(\text{SCError}_1 - \text{SCError}_2)/(2-1)}{\text{SCError}_2/(14-2-1)} \sim F_{1,11}$$

Luego el valor observado de F es $F_{obs} = 18.34$, luego el p-valor de la prueba es 0.0013, por lo tanto se rechaza H_0, la cual para esta prueba significa que se rechaza que ambos modelos explican lo mismo. Es decir, existe suficiente evidencia que justifica la incorporación de la variable Exportaciones.

b) Como en general $\frac{\text{SCError}}{\sigma^2} \sim \chi^2_{n-p-1}$, donde p es la cantidad de covariables. Entonces

Para el modelo 1: SCError = 143868.69 con 12 grados de libertad, por tanto,

$$\widehat{\sigma}^2 = \frac{143868.69}{12} = 11989.06.$$

Para el modelo 2: SCError = 53938.8 con 11 grados de libertad, por tanto,

$$\widehat{\sigma}^2 = \frac{53938.8}{11} = 4903.53.$$

c) Ya que $Ex = 275$ y $OM = 1000$ se tiene que

$$C = 851.3 + 0.7945 * 1000 = 1645.8 \quad \wedge \quad C = 655.8 + 3.359 \cdot 275 + 0.0556 \cdot 1000 = 1635.125$$

son los estimadores de consumo usando modelo 1 (reducido) y 2, respectivamente. Para los intervalos de confianza del valor medio (otra opción construir el IC del valor individual) se tiene:

- Modelo 1: $S^2_{01} = 11989.06 \begin{pmatrix} 1 \\ 1000 \end{pmatrix}^\top \begin{pmatrix} 0.2491311 & -0.0002715 \\ -0.0002715 & 0.0000003 \end{pmatrix} \begin{pmatrix} 1 \\ 1000 \end{pmatrix} = 144.4241.$

  Por lo tanto, por ejemplo, un IC del 95% de confianza es

  $$IC_{95\%}(C_0) = 1645.8 \mp t_{0.975,12} \cdot \sqrt{144.4241} = 1645.8 \mp 26.184 = [1619.62,\ 1671.98]$$

- Modelo 2: $S^2_{02} = 4903.53 \begin{pmatrix} 1 \\ 275 \\ 1000 \end{pmatrix}^\top \begin{pmatrix} 0.6758016 & -0.007303 & 0.001341 \\ -0.007303 & 0.000125 & -0.0000276 \\ 0.001341 & -0.0000276 & 0.0000064 \end{pmatrix} \begin{pmatrix} 1 \\ 275 \\ 1000 \end{pmatrix} = 70.006.$

  Por lo tanto, por ejemplo, un IC del 95% de confianza es

  $$IC_{95\%}(C_0) = 1635.125 \mp t_{0.975,11} \cdot \sqrt{70.006} = 1635.125 \mp 18.416 = [1616.71,\ 1653.54]$$

d) Ninguno de los dos modelos es completamente adecuado, ya que si bien entre el modelo 1 y el modelo 2, el segundo es mejor que el primero (ver respuesta item (a)), al observar el estadístico de prueba de la variable OM en el modelo 2, se aprecia que su valor es muy pequeño (0.312), es decir, dicha variable no es significativa (p-valor=0.7609), por lo tanto, sería conveniente comparar el modelo 2 con un tercer modelo que considere solo a las Exportaciones como variable explicativa del consumo.

2. En un estudio se examinaron las relaciones entre las condiciones meteorológicas durante los primeros 21 días después de la eclosión de las crías de codorniz escalada y su supervivencia hasta los 21 días de edad. Sea p la probabilidad de que las crías sobrevivan más de 21 días. Se utilizó un total de 54 crías en el estudio donde se ajustó un modelo logístico, cuyos resultados se muestran en la siguiente Tabla, donde $\widehat{\beta}_j$ son estimaciones de los parámetros, $se(\widehat{\beta}_j)$ sus respectivos errores estándar, mientras, que C el estadístico de prueba de razón de verosimilitud (delta de Deviance), cuando la variable indicada se excluyó del modelo completo que contenía las tres variables explicativas.

Variable explicativa	$\widehat{\beta}_j$	$se(\widehat{\beta}_j)$	C
Temperatura mínima durante los primeros 12 días	0,143	0,19	0,602
Temperatura máxima durante los primeros 7 días	1,247	0,45	14,83
Número de días con precipitaciones durante los primeros 7 días	-0,706	0,45	2,83

a) Con la información entregada, proponga un modelo basado en los estadísticos de la razón de verosimilitud (delta de Deviance).

b) Utilice las pruebas de Wald para determinar qué variables explicativas son significativas.

c) Suponga que el estimador del intercepto es $\widehat{\beta}_0 = -20.45$. Obtenga la estimación de la probabilidad que una cría sobreviva más de 21 días si se tiene una temperatura mínima durante los primeros 12 días fue de 2^oC, una temperatura máxima de 18^oC durante los primeros 7 días y solo un día de precipitaciones durante los primeros 7 días.

SOLUCIÓN:

a) Se definen las variables

- X_1: Temperatura mínima durante los primeros 12 días
- X_2: Temperatura máxima durante los primeros 7 días
- X_3: Número de días con precipitaciones durante los primeros 7 días

El test que se pone a prueba es

$$\mathrm{H}_0: \text{ Modelo Reducido} \quad \text{vs.} \quad \mathrm{H}_1: \text{ Modelo Completo}$$

En cada caso $\Delta D = C \sim \chi^2_{(1)}$, ya que el modelo mayor tienen 4 grados de libertad (3 covariables y el intercepto), mientras que los modelos reducidos tienen 3 grados de libertad (a cada uno se le quita una covariable). Luego, los p-valores de los test son:

- Si se omite X_1, se tiene; $p\text{-valor} = \mathbb{P}[\chi^2_{(1)} > 0,602] = 0,4378.$

  Concluyendo que, no se rechaza H_0, al nivel de significancia del 5%.
- Si se omite X_2, se tiene; $p\text{-valor} = \mathbb{P}[\chi^2_{(1)} > 14,83] = 0,0001.$

  Concluyendo que, se rechaza H_0, al nivel de significancia del 5%.
- Si se omite X_3, se tiene; $p\text{-valor} = \mathbb{P}[\chi^2_{(1)} > 2,83] = 0,09258.$

  Concluyendo que, no se rechaza H_0, al nivel de significancia del 5%.

Por lo tanto, aplicando el principio de parsimonia, el único modelo significativo al 5%, es aquel que incluye la variable X_2. Es decir, quitar del modelo dicha variable provoca un aumento significativo de la Deviance. Para proponer un modelo final, falta comparar el modelo completo, con aquel donde se quitan las variables X_1 y X_3 al mismo tiempo y analizar si dichos modelos tienen Deviance similar.

b) El test de Wald establece

$$\mathrm{H}_0: \ \beta_j = 0 \quad \text{vs.} \quad \mathrm{H}_1: \ \beta_j \neq 0$$

donde el estadístico de prueba es

$$T = \frac{\widehat{\beta}_j}{s.e(\widehat{\beta}_j)} \sim t_{(54-3-1)} \quad \text{aproximadamente}$$

Así se tiene

- Para β_1, se tiene que

$$t_{obs} = \frac{0,143}{0,19} = 0,753 \quad \Rightarrow \quad p - \text{valor} = 2 \cdot \mathbb{P}[T > 0,753] = 0,4552$$

  En conclusión, no se rechaza H_0, al nivel de significancia del 5%.

- Para β_2, se tiene que

$$t_{obs} = \frac{1,247}{0,45} = 2,771 \quad \Rightarrow \quad p - \text{valor} = 2 \cdot \mathbb{P}[T > 2,771] = 0,0078$$

  En conclusión, se rechaza H_0, al nivel de significancia del 5%.

- Para β_3, se tiene que

$$t_{obs} = \frac{-0,706}{0,45} = -1,569 \quad \Rightarrow \quad p - \text{valor} = 2 \cdot \mathbb{P}[T > 1,569] = 0,1230$$

  En conclusión, no se rechaza H_0, al nivel de significancia del 5%.

Por lo tanto, la única variable significativa es X_2.

c) Como

$$\widehat{p} = \frac{e^{\boldsymbol{X}\widehat{\boldsymbol{\beta}}}}{1 + e^{\boldsymbol{X}\widehat{\boldsymbol{\beta}}}} = \frac{e^{\widehat{\beta}_0 + \widehat{\beta}_1 \cdot X_1 + \widehat{\beta}_2 \cdot X_2 + \widehat{\beta}_3 \cdot X_3}}{1 + e^{\widehat{\beta}_0 + \widehat{\beta}_1 \cdot X_1 + \widehat{\beta}_2 \cdot X_2 + \widehat{\beta}_3 \cdot X_3}}$$

Para $X = (2, \ 18, \ 1)$, se tiene

$$\widehat{p} = \frac{e^{-20,45+0,143 \cdot 2+1,247 \cdot 18-0,706 \cdot 1}}{1 + e^{-20,45+0,143 \cdot 2+1,247 \cdot 18-0,706 \cdot 1}} = \frac{e^{1,576}}{1 + e^{1,576}} = 0,8286$$

Es decir, la probabilidad que una cría sobreviva más de 21 días, bajo las condiciones establecidas, es de 0,8286.