Regresión múltiple y otras técnicas multivariadas

Tarea 05

Rivera Torres Francisco de Jesús
Rodríguez Maya Jorge Daniel
Samayoa Donado Víctor Augusto
Trujillo Bariios Georgina

Marzo 13, 2019

Ejercicio 1

En un estudio que consistió en explorar la relación entre la longitud de la mandíbula (en milímetros) y la edad gestacional (en semanas) de 158 fetos, se obtuvieron los siguientes resultados:

$$\bar{x} = 20.1, \qquad \bar{y} = 23.6, \qquad S_{xx} = 2473.8, \qquad S_{yy} = 8652.4, \qquad S_{xy} = 4385.4$$

Responder lo siguiente:

Inciso 1.a

Ajustar un modelo RLS para explicar la distribución de la longitud de la mandíbula de los fetos como función de la edad gestacional. Reportar las estimaciones de los parámetros. Interpretar los resultados en el contexto de los datos.

Sabemos que los estimadores de los coeficientes del modelo RLS, se pueden obtener de la siguiente manera:

$$\hat{\beta}_0 = \bar{y} - \frac{S_{xy}}{S_{xx}}\bar{x} \qquad\qquad \hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$$

por lo tanto, se tiene que:

$$\hat{\beta}_0 = 23.6 - \frac{4385.4}{2473.8}20.1 = -12.03204 \qquad \text{y} \qquad \hat{\beta}_1 = \frac{4385.4}{2473.8} = 1.772738$$

teniendo así el modelo RLS, $\hat{y} = -12.03204 + 1.772738x$.

De acuerdo con el contexto de los datos, el parámetro $\hat{\beta}_0$ no tiene una interpretación dentro del problema.

Por otro lado, el estimador del parámetro $\hat{\beta}_1$ nos indica que por cada semana que transcurre en la gestación del feto, la longitud de la mandibula crece, en promedio, 1.77 miímetros.

Inciso 1.b

Construir la tabla ANOVA y contrastar la significancia del modelo del inciso anterior. Usar un tamaño de prueba $\alpha = 0.05$.

Sabemos que, para el modelo RLS, la tabla ANOVA se define como sigue:

Tabla 1: Tabla ANOVA

FV	GL	SC	CM	F	$F^{(1-\alpha)}_{(1,n-2)}$
Regresión	1	SCR	SCR	$\frac{\text{CMR}}{\text{CME}}$	$F^{(1-\alpha)}_{(1,n-2)}$
Error	n - 2	SCE	$\frac{\text{SCE}}{n-2}$		
Total	n - 1	SCT			

donde $n = 158$ y

$$SCR = S_{xx}\hat{\beta}_1^2, \qquad SCE = \frac{S_{xx}S_{yy} - S_{xy}^2}{S_{xx}}, \qquad SCT = SCR + SCE$$

$$SCR = 2473.8(1.772738)^2, \qquad SCE = \frac{(2473.8)(8652.4) - (4385.4)^2}{2473.8}, \qquad SCT = SCR + SCE$$

$$SCR = 7774.164 \qquad SCE = 878.2335 \qquad SCT = 8652.397$$

Con estos valores se procede a calcular los campos restantes de la tabla ANOVA:

$$CMR = SCR \qquad CME = \frac{SCE}{n-2} \qquad F = \frac{CMR}{CME}$$

$$CMR = 7774.164 \qquad CME = 5.629702 \qquad F = 1380.919$$

Utilizando un $\alpha = 0.05$ se tiene que el cuantil de la distribución de referencia es

```
alpha <- 0.05
f.a <- qf(1 - alpha, df1 = 1, df2 = 156)
```

Es decir, el cuantil de la distribución de referencia es 3.9017607. El cual es un valor menor a $F = 1380.919$. Esto implica que se debe rechazar H_0 con un nivel de significancia del 0.05. Esto implica que la variable explicativa **edad gestacional** tiene algún efecto en la distribución de la variable objetivo **longitud de la mandíbula**.

Usando la información anterior, la tabla ANOVA queda como sigue:

Tabla 2: Tabla ANOVA

FV	GL	SC	CM	F	$F_{(1,n-2)}^{(1-\alpha)}$
Regresión	1	7774.164	7774.164	1380.919	3.9017607
Error	156	878.2335	5.629702		
Total	157	8652.397			

Inciso 1.c

Calcular un intervalo de predicción para la media de la longitud de la mandíbula de un feto con 23 semanas de gestación.

Sabemos que la media de la longitud de la mandíbula de un feto con x_0 semanas de gestación está dada por

$$\hat{\mu}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0 \qquad \text{entonces} \qquad \hat{\mu}_0 = -12.03204 + 1.772738(23) = 28.74093$$

Además, sabemos que el intervalo de confianza está dado por:

$$\hat{\mu}_0 \pm t_{n-2}^{1-\alpha/2} \hat{\sigma}_{\text{MCO}} \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}$$

donde podemos estimar la varianza mediante la igualdad $\hat{\sigma}_{\text{MCO}} = \dfrac{SCE}{n-2} = CME.$

Calculando el estadístico $t_{n-2}^{1-\alpha/2}$ se obtiene

```
alpha <- 0.05
n <- 158
ta <- qt(1 - alpha/2, n - 2)
```

$t_{n-2}^{1-\alpha/2} = 1.9752875$

Entonces tenemos que el intervalo de confianza está dado por:

$$\hat{\mu}_0 \pm t_{n-2}^{1-\alpha/2} \hat{\sigma}_{\text{MCO}} \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}$$

$$28.74093 \pm (1.975288)\sqrt{5.629702}\sqrt{\frac{1}{158} + \frac{(23 - 20.1)^2}{2473.8}}$$

$$28.74093 \pm (1.975288)(2.372699)\sqrt{0.006329114 + 0.003399628}$$

$$28.74093 \pm (1.975288)(2.372699)(0.09863439)$$

$$28.74093 \pm 0.4622761$$

Por lo tanto, el intervalo de predicción para la media de la longitud de la mandíbula de un feto con 23 semanas de gestación está dado por:

$$[28.27865, 29.20321]$$

esto es, la media de la longitud de la mandibula de un feto con 23 semanas de gestación está entre 28.27865 y 29.20321 milímetros con un 95% de confianza.

Inciso 1.d

Calcular el R^2 del modelo e interpretar el resultado.

Recordemos que

$$R^2 = 1 - \frac{SCE}{SCT} = 1 - \frac{878.2335}{8652.397} = 0.8984982$$

Esto nos indica que el ajuste del modelo es adecuado, ya que el modelo logra explicar un 89.84% de la variación total de la longitud de la mandíbula de los fetos.

Ejercicio 2

En un estudio sobre los 67 condados del estado de Florida (EUA), se obtuvieron los siguientes resultados sobre

Variable	**Promedio**	**Desv. est.**
Ingreso	24.51	4.69
Educación	69.49	8.86

Se ajustó un modelo RLS a los datos, para explicar la distribución del ingreso como función de la educación, y se obtuvieron los siguientes resultados:

$$\hat{\beta}_0 = -4.63, \qquad \hat{\beta}_1 = 0.42$$

Inciso 2.a

Construir intervalos simultáneos de confianza 95% para β_0 y β_1.

Sabemos que un intervalo de confianza simultáneo está dado por:

$$\beta_0 \in \left(\hat{\beta}_0 \pm t_{(n-2)}^{(1-\alpha/4)} \hat{\sigma}_{MCO} \sqrt{\hat{V}(\hat{\beta}_0)}\right)$$
$$\beta_1 \in \left(\hat{\beta}_1 \pm t_{(n-2)}^{(1-\alpha/4)} \hat{\sigma}_{MCO} \sqrt{\hat{V}(\hat{\beta}_1)}\right)$$

Y tomando en cuenta que

$$SCR = S_{xx}\hat{\beta}_1^2,\ SCE = \frac{S_{xx}S_{yy} - S_{xy}^2}{S_{xx}} \text{ y que } SCT = SCR + SCE$$

Tenemos que $n = 64$, así $S_{xx} = 67 \cdot (8.86)^2 = 5259.473$, luego

$S_{xy} = \hat{\beta}_1 \cdot S_{xx} = (0.42)(5259.473) = 2208.9787$,

$SCT = \sum(y_i - \overline{y}_n)^2 = 67 \cdot (4.69)^2 = 1473.73$,

$SCR = \sum(\hat{y}_i - \overline{y}_n)^2 = \hat{\beta}_1^2 S_{xx} = (0.42)^2 \cdot 5259.473 = 927.77$, así $SCR = SCT - SCR = 545.967$.

Además, $\hat{\sigma}_{MCO}^2 = 545.96/65 = 8.3995$,

$\hat{V}(\hat{\beta}_0) = \left(\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right) \cdot \hat{\sigma}_{MCO}^2 = (0.933) \cdot (8.3995) = 7.8367$ y

$\hat{V}(\hat{\beta}_1) = \hat{\sigma}_{MCO}^2 / S_{xx} = 0.001597$

Calculando el estadístico $t_{67-2}^{1-(0.05)/4}$

```
t <- qt(1-0.05/4, 65)
t
```

```
## [1] 2.294512
```

$t_{67-2}^{1-(0.05)/4} = 2.2945$

Por lo que los intervalos simultáneos son:

```
## [1] -23.24588  13.98588
```

```
## [1] 0.1542524 0.6857476
```

IC para β_0: $(-23.2458, 13.9858)$ y IC para β_1: $(0.1542, 0.6857)$

Inciso 2.b

Construir la tabla ANOVA y contrastar la significancia del modelo ajustado. Utilizar $\alpha = 0.05$.

Tabla 3: Tabla ANOVA

FV	GL	SC	CM	F
Regresión	1	927.77	927.77	110.454
Error	65	545.96	8.3995	
Total	66	1473.73		

Para $H_0: \ \beta_0 = 0$ vs $H_1: \ \beta_1 \neq 0$

$$T = \frac{\hat{\beta}_1 - b_1}{\hat{\sigma}_{MCO}/\sqrt{S_{xx}}} = 10.5097$$

y $t_{65}^{0.975} = 1.9971$

Como $T = 10.5097 > t = 1.9971$ se concluye rechazar H_0, esto quiere decir que existe evidencia en la releción entre la educación y el ingreso.

Por otra parte de la tabla ANOVA

$$F = 110.45 > \left(t_{65}^{0.975}\right)^2 = 3.9885$$
$$T^2 = (10.5097)^2 = 110.45 = F$$

Inciso 2.c

Reportar la estimación de σ y calcular un intervalo de confianza 95%.

Tenemos que

$$\hat{\sigma}^2_{MCO} = \frac{\sum(y_i - \hat{y}_n)^2}{n-2}$$

y además sabemos que

$$\sum(y_i - \hat{y}_n)^2 = \sum(y_i - \overline{y}_n)^2 - \sum(\hat{y}_i - \overline{y}_n)^2 =$$
$$= 1473.7387 - 5259.4732 = 545.9676$$
$$\Rightarrow \hat{\sigma}^2_{MCO} = 545.9676/65 = 8.3995$$

Para los intervalos de confianza utilizamos:

$$\left(\frac{(n-2)\cdot\hat{\sigma}^2_{MCO}}{\chi^2_{n-2}(1-\alpha/2)}, \frac{(n-2)\cdot\hat{\sigma}^2_{MCO}}{\chi^2_{n-2}(\alpha/2)}\right)$$

$$\Rightarrow \chi^2_{65}(0.975) = 89.177,$$

$$\text{y } \chi^2_{65}(0.025) = 44.6029$$

Así el intervalo de confianza para σ^2 es

$$(6.1222, 12.2406)$$

Inciso 2.d

Calcular el R^2 del modelo e interpretar el resultado.

Sabemos que $R^2 = 1\frac{SCE}{SCT}$ entonces

$$R^2 = 1 - \frac{545.9676}{1473.7387} = 0.6295$$

Por lo que podemos concluir que el modelo no describe adecuadamente el comportamiento de los datos.