Regresión múltiple y otras técnicas multivariadas

Tarea 04

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

Marzo 06, 2019

Ejercicio 1

Mostrar que, para el modelo RLS, se cumple $SCR = S_{xx}\hat{\beta}_1^2$.

Demostración. Recordemos las siguientes identidades:

$$\hat{y}_i \overset{(i)}{=} \hat{\beta}_0 + \hat{\beta}_1 x_i \qquad \hat{\beta}_1 \overset{(ii)}{=} \frac{S_{xy}}{S_{xx}} \qquad \hat{\beta}_0 \overset{(iii)}{=} \bar{y} - \frac{S_{xy}}{S_{xx}}\bar{x}$$

Entonces tenemos que:

$$\begin{aligned} SCR &= \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \overset{(i)}{=} \sum_{i=1}^{n} \left(\hat{\beta}_0 + \hat{\beta}_1 x_i - \bar{y}\right)^2 \overset{(ii),(iii)}{=} \sum_{i=1}^{n} \left(\bar{y} - \frac{S_{xy}}{S_{xx}}\bar{x} + \frac{S_{xy}}{S_{xx}} x_i - \bar{y}\right)^2 \\ &= \sum_{i=1}^{n} \left(\frac{S_{xy}}{S_{xx}}(x_i - \bar{x})\right)^2 = \sum_{i=1}^{n} \left[\left(\frac{S_{xy}}{S_{xx}}\right)^2 (x_i - \bar{x})^2\right] = \left(\frac{S_{xy}}{S_{xx}}\right)^2 \sum_{i=1}^{n} (x_i - \bar{x})^2 \\ &\overset{(ii)}{=} \hat{\beta}_1^2 S_{xx} \end{aligned}$$

□

Ejercicio 2

Inciso 2.a

Mostrar que, para el modelo RLS, se cumple

$$F = \frac{R^2\,(n-2)}{1 - R^2},$$

donde F es el estadístico del ANOVA y R^2 es el coeficiente de determinación del modelo.

Demostración. Sabemos que

$$R^2 = 1 - \frac{SCE}{SCT}, \quad F = \frac{SCR}{\frac{SCE}{(n-2)}} \quad \text{y} \quad SCT = SCR + SCE$$

Entonces

$$
\begin{aligned}
F &= \frac{SRE}{\frac{SCE}{(n-2)}} = \frac{SCR(n-2)}{SCE} = \frac{(SCT - SCE)(n-2)}{SCE}\left(\frac{SCT}{SCT}\right) \\
&= \frac{\left(\frac{SCT}{SCT} - \frac{SCE}{SCT}\right)(n-2)}{\frac{SCE}{SCT}} = \frac{\left(\frac{SCT}{SCT} - \frac{SCE}{SCT}\right)(n-2)}{1 - 1 + \frac{SCE}{SCT}} = \frac{\left(1 - \frac{SCE}{SCT}\right)(n-2)}{1 - \left(1 - \frac{SCE}{SCT}\right)} \\
&= \frac{R^2\,(n-2)}{1-R^2}
\end{aligned}
$$

Por lo tanto

$$
F = \frac{R^2\,(n-2)}{1-R^2},
$$

□

Inciso 2.b

Suponer que para un conjunto de 20 pares de observaciones $(x, y), r = -0.74$. Si se ajusta un modelo RLS a estos datos, ¿que se puede decir sobre la significancia del modelo?

Pdemos decir que el modelo tiene una asociación lineal inversa fuerte entre las observaciones (x, y), es decir, si una crece la otra decrece.

Ejercicio 3

Los siguientes datos representan la relación entre el número de errores de alineación y el número de remaches faltantes para 10 diferentes aeronaves

Los siguientes datos representan la relación entre el número de alineación y el número de remaches faltantes para 10 diferentes aeronaves.

Num. remaches (x)	13	15	10	22	30	7	25	15	20	15
Num. errores (x)	7	7	5	12	15	2	13	9	11	8

Inciso 3.a

Ajustar un modelo RLS e interpretar las estimaciones de los coeficientes del modelo.

```
num_remaches <- c(13, 15, 10, 22, 30, 7, 25, 16, 20, 15)
num_errores <- c(7, 7, 5, 12, 15, 2, 13, 9, 11, 8)
aeronaves <- data.frame(num_remaches, num_errores)
```

```
modelo <- lm(num_errores~num_remaches, data = aeronaves)
summary(modelo)$coefficients
```

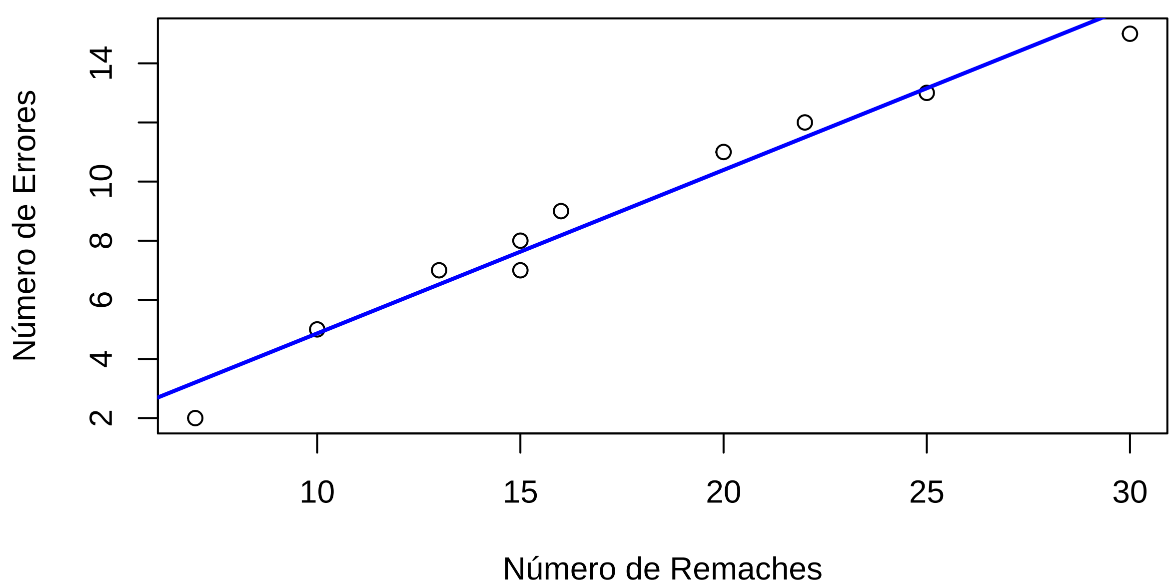
```
##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)  -0.6639400 0.65475832 -1.014023 3.402547e-01
## num_remaches  0.5528289 0.03533818 15.643957 2.780621e-07
```

Para el coeficiente β_0 se tiene un número negativo, lo que nos indica que tendriamos -0.6639 errores cuando no tenemos remaches y para el coeficiente β_1 podemos interpretar que tenemos aproximadamente la mitad de errores según la cantidad de remaches

Inciso 3.b

Graficar el diagrama de dispersión de los datos y sobreponer la recta de regresión ajustada.

```
plot(num_errores~num_remaches, data = aeronaves, xlab = "Número de Remaches",
     ylab = "Número de Errores")
abline(modelo, col = "blue", lwd = 2)
```



Inciso 3.c

Reportar el R^2 del modelo ajustado. ¿La interpretación del R^2 es consistente con la gráfica del inciso anterior?

```
summary(modelo)$r.squared
```

```
## [1] 0.9683461
```

Obtenemos una R^2 de 0.9683 lo que nos indica que sí es consistente puesto que el modelo se ajusta bien en la grafica de dispersión.

Inciso 3.d

Contrastar las hipótesis

$$H_0 : \beta_0 = 1 \quad \text{v.s.} \quad H_1 : \beta_0 \neq 1$$

Utilizar el tamaño de prueba $\alpha = 0.1$. Interpretar el resultado en el contexto del problema.

```
#calculamos el estadistico de prueba
b_0 <- 1
v.b_0.est <- vcov(modelo)[1,1]
b_0.est<-summary(modelo)$coefficients[1,1]
T <- (b_0.est - b_0)/v.b_0.est

#cuantil de t (n-2)
t <- qt(1-.1/2, df.residual(modelo))
T
```

```
## [1] -3.881286
```

```
t
```

```
## [1] 1.859548
```

Como $3.88 = |T| > t = 1.85$ rechazamos H_o, es decir que en el modelo β_0 debe ser diferente de 1 en el contexto del problema nos indica que no podemos obtener errores si no hay remaches.

Inciso 3.e

Estimar puntualmente el número esperado de errores de alineación de un avión con 24 remaches faltantes y construir un IC 90%.

Estimamos puntualmente:

```
b_1.est <- summary(modelo)$coefficients[2,1]
y.24 <- b_0.est+(b_1.est*24)
y.24
```

```
## [1] 12.60395
```

La estimación es de 12.6 errores.

Intervalos de confianza:

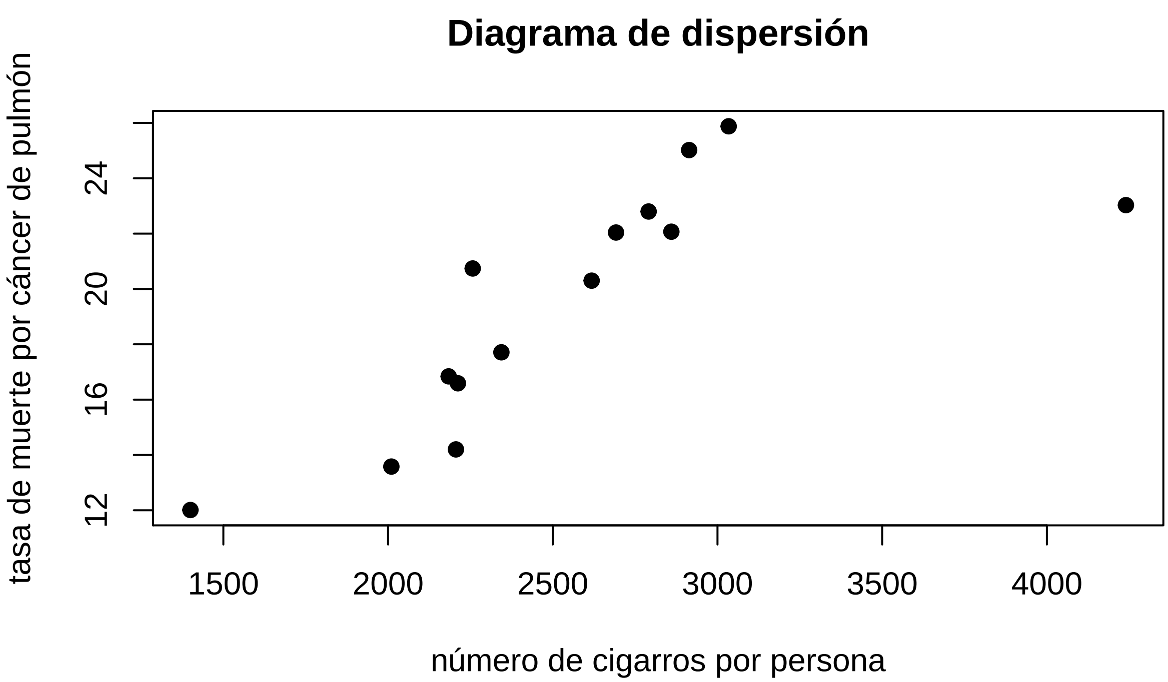
```
s2.est <- summary(modelo)$sigma^2
ta <- qt(0.95, nrow(aeronaves)-2)
IC.24 <- y.24+c(-1, 1)*ta*sqrt(s2.est)
IC.24

## [1] 11.22539 13.98252
```

Ejercicio 4

Utilizar el conjunto de datos que se envía adjunto para responder lo siguiente.

Inciso 4.a

Graficar un diagrama de dispersión del número de cigarros por persona (variable cigarrette) contra la tasa de muertes por cáncer de pulmón (por cada 100, 000 habitantes, variable lung). ¿La gráfica indica la posibilidad de una asociación lineal entre las variables?



La gráfica muestra que hay una asociación positiva entre el número de cigarros por persona y la tasa de muerte por cada mil habitantes causada por cáncer de pulmón.

Tabla 1: ANOVA

Fuentes.Variación	Grados.de.Libertad	Suma.de.Cuadrados	Cuadrados.Medios	Prueba.F
regresión	1	2434.72	2434.72	12.85353
error	12	2679.89	189.42	NA
total	13	248.00	NA	NA

Inciso 4.b

Ajustar un modelo RLS para explicar la distribución de lung como función de cigarrettes. Reportar las estimaciones de los coeficientes e interpretarlos en el contexto de los datos.

El valor del intercepto $\beta_0 = 6.1976283$ podría interpretarse como la tasa de muertes por cáncer de pulmón (por cada 100000 habitantes) dado que en el estado observado el promedio o número de cigarros por persona es cero. Esto puede tener sentido debido a que fumar tabaco puede no ser la única causa de cáncer de pulmón. El valor de $\beta_1 = 0.0052023$ por su parte indica que un incremento de un cigarro más a nivel estatal incrementaría la tasa de muerte por cáncer de pulmón en 0.0052.

Inciso 4.c

Construir la tabla ANOVA para la hipótesis de significancia del efecto del número de cigarros por persona en la tasa de muertes por cáncer de pulmón. `bquote(H[0] : b[0] == 1 vs. H[1]: != 1)`

La suma de cuadrados total es 247.7415214 La suma de cuadrados de regresión es 2434.7218625 La suma de cuadrados del error es 2679.8857526

Inciso 4.d

Concluir sobre la significancia del modelo si se utiliza un tamaño de prueba de 0.05. Reportar el p-value de la prueba.

Rechazamos la prueba $H_0 : \beta_1 = 0$ si el valor de $F > F^{(1-\alpha)}_{(1,n-2)}$

$F = 12.8535349 > 0.2443067$, por lo tanto con un tamaño de prueba $\alpha = 0.05$ rechazamos la hipótesis nula. En otras palabras, los datos aportan evidencia estadísticamente significativa sobre la relación entre tasa de muerte por cáncer de pulmón y número de cigarros por persona.

Ejercicio 5

En un estudio se midió la estatura (X, en cm) y el peso (Y , en kg) de 50 mujeres de entre 20 y 24 años, y se ajustó un modelo RLS para explicar el la distribución del peso en como función de la estatura. A continuación de muestra un resumen de la información obtenida.

$$\bar{x} = 164.9 \qquad \bar{y} = 59.3 \qquad S_{xx} = 2875.7 \qquad S_{yy} = 1423.5 \qquad S_{xy} = 1222.5$$

Responder lo siguiente:

Inciso 5.a

Reportar las estimaciones de los coeficientes del modelo e interpretarlas en el contexto de los datos.

Podemos estimar los coeficientes del modelo con las identidades vistas en clase, esto es:

$$\begin{aligned} \widehat{\beta_0} &= \overline{y_n} - \frac{S_{xy}}{S_{xx}} \overline{x_n} = -10.8 \\ \text{y } \widehat{\beta_1} &= \frac{S_{xy}}{S_{xx}} = 0.425 \\ \longrightarrow \widehat{\beta_0} = -10.8 \;\; &\wedge \;\; \widehat{\beta_1} = 0.425 \end{aligned}$$

De esta manera podemos concluir que el estimador $\widehat{\beta_0}$ carece de sentido en este contexto, pues no podemos hablar de mujeres que tengan una altura de cero centímetros. El estimador $\widehat{\beta_1}$, por otra parte, nos indica que hay una relación de crecimiento positiva por talla y el peso en las mujeres, más específico, que por cada centímetro extra en la talla (altura), habrá un incremento esperado de 0.425 kg en el peso.

Inciso 5.b

¿Hay evidencia de que la estatura tiene algún efecto en el peso esperado de una persona?

Para responder esta pregunta podemos plantear el siguiente contraste de hipótesis:

$$H_o : \widehat{\beta_1} = 0 \text{ vs } H_1 : \widehat{\beta_1} \neq 0$$

Rechazamos H_0 cada vez que $F > F_{n-2}^{1-\alpha}$ con $n = 50$.

Tenemos que $F = \frac{SCR}{\frac{SCE}{n-2}}$, para poder calcularlo haremos uso de algunas propiedades que ya hemos visto o demostrado con anterioridad, esto es:

$$\begin{aligned} SCR = S_{xx}\hat{\beta}_1^2 = (2875.7) \cdot (0.425)^2 &= 519.7 \\ SCE = \frac{S_{xx}S_{yy} - S_{xy}^2}{S_{xx}} &= 903.79 \\ \text{y } F = \frac{SCR}{\frac{SCE}{n-2}} = \frac{519.7}{\frac{903.79}{48}} &= 27.6 \end{aligned}$$

Y calculando F con ayuda de ”R´´ y con $\alpha = 0.5$ entonces

```
## [1] 4.042652
```

Tabla 2: Tabla ANOVA

FV	GL	SC	CM	F
Regresión	1	519.7	519.7	27.6
Error	48	903.79	18.82143.25	
Total	49	143.25		

De aquí podemos concluir que $F = 27.6 > 4.04$ por lo que hay evidencia suficiente para rechazar la hipótesis H_0. En el contexto del problema podemos interpretar este resultado como que hay aceptable evidencia para concluir que la estatura tiene efecto en el peso esperado para las mujeres al menos en el rango de 20 a 24 años.

Inciso 5.c

Construir la tabla ANOVA para este modelo

Donde:

$$SCT = SCR + SCE = 1423.5$$

$$CMR = \frac{SCR}{1} = 519.7$$
$$CME = \frac{CME}{n-2} = 18.82$$

y $F = 27.6$

Inciso 5.d

Calcular el R^2 e interpretar el resultado

$$R^2 = 1 - \frac{SCE}{SCT} = 1 - \frac{903.79}{1423.5} = 1 - 0.6349 = 0.3650$$

Este cociente representa la proporción de la variación muestral de y que es explicada por x, esto es, la variación explicada entre la variación total. Como podemos observar el valor es más cercano a cero que a 1, por lo que podemos concluir que el modelo no tiene un buen ajuste, dentro del problema que estamos trabajando.

Ejercicio 6

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

Tabla 3: Tabla ANOVA

FV	GL	SC	CM	F	p-value
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

Inciso 6.a

Completar la información de la tabla anterior (sólo las celdas señaladas con X).

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

Tabla 4: Tabla ANOVA

FV	GL	SC	CM	F	p-value
Regresión	1	SCR	SCR	$\frac{\text{CMR}}{\text{CME}}$	$P(F > F_{(1,n-2)})$
Error	n - 2	SCE	$\frac{\text{SCE}}{n-2}$		
Total	n - 1	SCT			

Comparando las dos tablas anteriores, obtenemos las siguiente igualdades:

$$n - 1 = 20 \qquad SCE = 92.62 \qquad SCT = 112.7 \qquad SCR = 20.11$$

De lo anterior obtenemos lo siguiente: Comparando las dos tablas anteriores, obtenemos las siguiente igualdades:

$$n = 21 \qquad n - 2 = 18 \qquad \frac{SCE}{n-2} = \frac{92.62}{18} = 5.15 \qquad F = \frac{20.11}{5.15} = 3.90$$

El p-value lo calculamos usando `R`.

```
p_value <- pf(3.90, df1 = 1, df2 = 18, lower.tail = FALSE)
```

Obteniendo así un resultado de $\text{p} - \text{value} = 0.0638283$.

Con la información anterior, procedemos a llenar la tabla de ANOVA

Tabla 5: Tabla ANOVA

FV	GL	SC	CM	F	p-value
Regresión	1	92.62	20.11	3.90	0.0638283
Error	18	92.62	5.15		
Total	20	112.7			

Inciso 6.b

Contrastar la significancia del modelo. Considerar un tamaño $\alpha = 0.1$.

Utilizando un $\alpha = 0.1$ se tiene que el cuantil de la distribución de referencia es

```
alpha <- 0.1
f.a <- qf(1 - alpha, df1 = 1, df2 = 18)
```

Es decir, el cuantil de la distribución de referencia es 3.0069766. El cual es un valor menor al $F = 3.90$. Esto implica que se debe rechazar H_0 con un nivel de significancia del 0.1. Esto implica que la variable no aleatoria X tiene algún efecto en la distribución de la variable aleatoría Y.

Inciso 6.c

Estimar $|\hat{\beta}_1|$ y estimar el error estándar del estimador.

En el ejrcicio 1 se demostró que $SCT = \hat{\beta}_1^2 S_{xx}$ y de la tabla ANOVA sabemos que $SCT = 112.7$. Además que nos dan la información de que $S_{xx} = 770.0$. Por lo tanto se tiene que:

$$|\hat{\beta}_1| = \sqrt{\left(\frac{SCT}{S_{xx}}\right)} = 0.382575$$

Inciso 6.d

Calcular el R^2 e interpretar el resultado.

Sabemos que el coeficiente de determinación del modelo RLS, se define como

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

De la tabla ANOVA construida en el inciso 6.a, sabemos que $SCE = 92.62$ y $SCT = 112.7$, por lo tanto

$$R^2 = 1 - \frac{SCE}{SCT} = 1 - \frac{92.62}{112.7} = 1 - 0.8218279 = 0.1781721$$

Esto nos indica que el ajuste del modelo es malo, ya que el modelo solamente logra explicar un 17.8% de la variación total de las observaciones de la variable independiente Y.