Regresión múltiple y otras técnicas multivariadas

Tarea 03

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

Febrero 27, 2019

Ejercicio 1

Suponer que se ajusta un modelo RLS a las observaciones (x_i, y_i) con $i = 1, \cdots, n$. Mostrar que

$$SCE = \frac{S_{xx}S_{yy} - S_{xy}^2}{S_{xx}}$$

Donde:

- $SCE = \sum_i^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$
- $S_{yy} = \sum_i^n (y_i - \bar{y_n})^2$

Ejercicio 2

Mostrar la desigualdad de Bonferroni. Si $E_1, \ldots, E_k$ son eventos en un espacio de probabilidad (Ω, A, P) entonces:

$$P\left(\bigcap_{i=1}^{k} E_i\right) \geqslant 1 - \sum_{i=1}^{k} P(\Omega \setminus E_i)$$

Demostración. La demostración se realizará por inducción sobre el número de eventos en un espacio de probabilidad.

Base: k = 1

$$\begin{aligned}
P(\Omega) = 1, &\qquad \text{pero } \Omega = E \cup (\Omega \setminus E_1)\text{, entonces} \\
P(E \cup (\Omega \setminus E_1)) = P(E_1) + P(\Omega \setminus E_1) = 1, &\qquad \text{ya que son probabilidades mutuamente excluyentes} \\
P(E_1) = 1 - P(\Omega \setminus E_1), &\qquad \text{como se da la igualdad entonces tambien se satisface que} \\
P(E_1) \geqslant 1 - P(\Omega \setminus E_1)
\end{aligned}$$

Ahora, por hipótesis de inducción, suponemos que se vale para n eventos en el espacio de probabilidad, por lo que se satisface la desigualdad

$$P\left(\bigcap_{i=1}^{n} E_i\right) \geqslant 1 - \sum_{i=1}^{n} P(\Omega \setminus E_i)$$

Y procedemos a demostrar que siempre que se cumpla para n eventos, se debe de cumplir para $n+1$ eventos.

$$\begin{aligned} P\left(\bigcap_{i=1}^{n+1} E_i\right) &= P\left(\left(\bigcap_{i=1}^{n} E_i)\right) \bigcap E_{n+1}\right) \\ &= P\left(\bigcap_{i=1}^{n} E_i)\right) + P\left(E_{n+1}\right) - P\left(\left(\bigcap_{i=1}^{n} E_i)\right) \bigcup E_{n+1}\right) \end{aligned}$$

Pero notemos que $P\left(\left(\bigcap_{i=1}^{n} E_i\right)\right) \bigcup E_{n+1}) \leqslant 1$, por lo que $-P\left(\left(\bigcap_{i=1}^{n} E_i\right)\right) \bigcup E_{n+1}) \geqslant -1$, entonces

$$P\left(\bigcap_{i=1}^{n+1} E_i\right) \geqslant P\left(\bigcap_{i=1}^{n} E_i)\right) + P\left(E_{n+1}\right) - 1$$

aplicando la hipótesis de inducción, se tiene

$$\begin{aligned} &\geqslant 1 - \sum_{i=1}^{n} P(\Omega \setminus E_i) + P\left(E_{n+1}\right) - 1 \\ &\geqslant 1 - \sum_{i=1}^{n} P(\Omega \setminus E_i) + \left(1 - P\left(\Omega \setminus E_{n+1}\right)\right) - 1 \\ &\geqslant 1 - \sum_{i=1}^{n} P(\Omega \setminus E_i) + P\left(\Omega \setminus E_{n+1}\right) \\ &\geqslant 1 - \sum_{i=1}^{n+1} P(\Omega \setminus E_i) \end{aligned}$$

□

Ejercicio 3

Considerar los datos de ingreso y escolaridad utilizados en los ejemplos de intervalos de confianza de las notas. Reportar intervalos simultáneos de confianza 95% para las medias del ingreso por hora para 9, 15 y 19 años de escolaridad a) con el método de Bonferroni y b) con el método de Hotelling−Scheffé

Ejercicio 4

El conjunto de datos `airquality`, de paquete datasets de R contiene información sobre la calidad del aire en Nueva York registrada de Mayo a Septiembre de 1973 (se pude consultar más información con el comando `help("airquality")`. Para responder este ejercicio, descartar las observaciones con valores perdidos.

Inciso 4.a

Ajustar un modelo RLS para explicar el nivel de ozono como función del log_2 de la velocidad del viento. Reportar las estimaciones de los parámetros.

```
library(tidyverse)

modelo_air <- airquality %>%
              as_tibble() %>%
              filter(!is.na(Ozone), !is.na(Wind)) %>%
              mutate(log2_Wind = log2(Wind)) %>%
              lm(formula = Ozone ~ log2_Wind)

coefficients(modelo_air)
```

```
## (Intercept)   log2_Wind
##   166.64640   -38.92431
```

Inciso 4.b

Mostrar una gráfica de dispersión de los datos utilizados para ajustar el modelo del inciso anterior, la recta de regresión ajustada y bandas de confianza 95%. Anexar el código relacionado con el cómputo de las bandas de confinza.

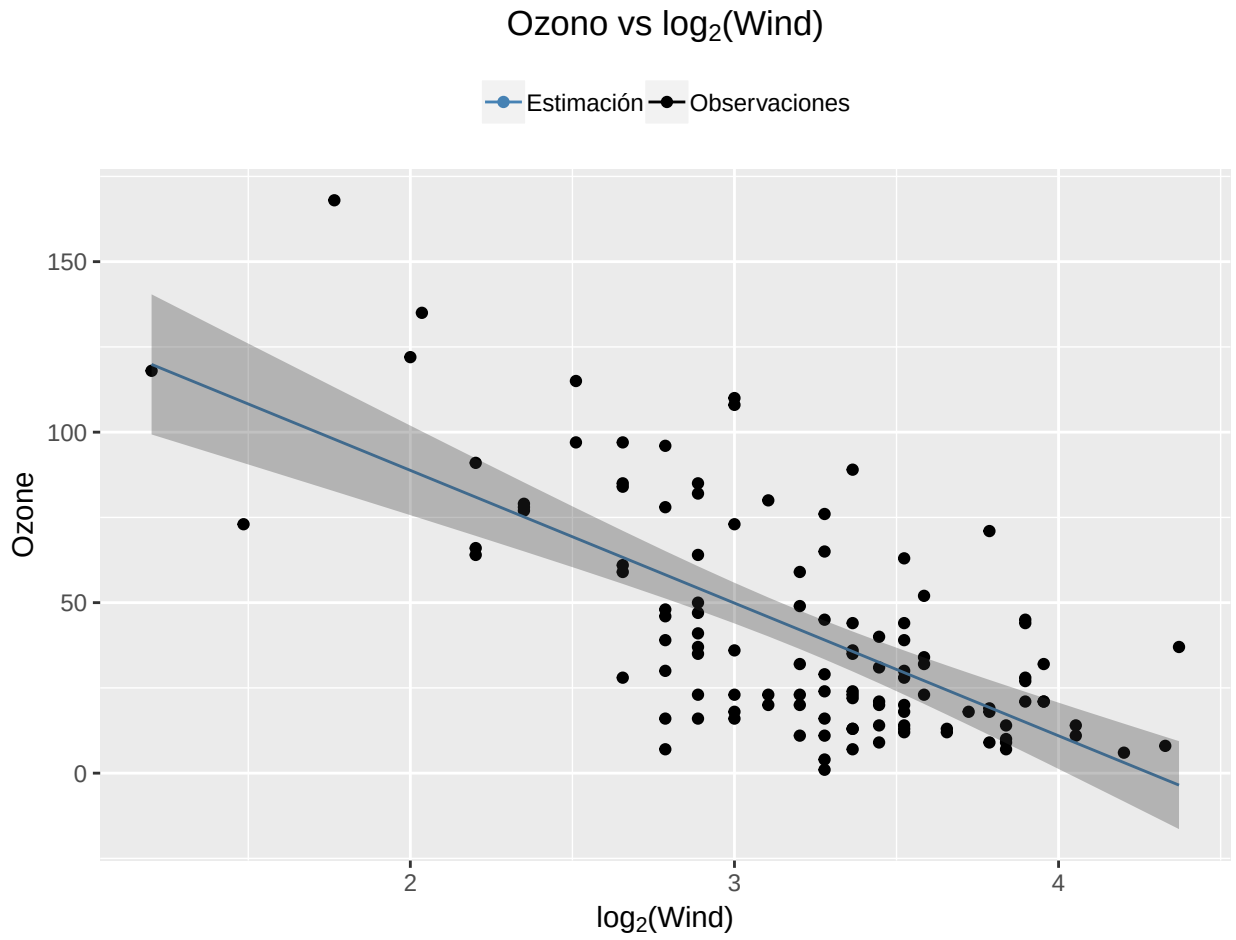
```
# Código para el cálculo de los intervalos de confianza para el módelo RLS
bandas_confianza_rls <- function(datos, variable_x, variable_y,
                                 formula_rls, alpha = 0.95) {
  # Función que calcula las bandas de confianza para el modelo RLS
  # usando el método de Hotelling-Scheffé

  datos <- datos %>%
           as_tibble()

  variable_x <- enquo(variable_x)
  variable_y <- enquo(variable_y)

  # Se calcula el modelo RLS
  modelo <- datos %>%
            lm(formula = formula_rls)

  # número de observaciones
  n <- nrow(datos)

  # estimador de beta0
  b0.hat <- coefficients(modelo)[1]

  # estimador de beta1
  b1.hat <- coefficients(modelo)[2]
```

```
  # estimador sigma2
  s2.hat <- (summary(modelo)$sigma)^2

  # Se calculan la media y la varianza de X
  datos_summary <- datos %>%
                   summarise(x.bar = mean(!! variable_x),
                             S.xx = sum((!! variable_x - mean(!! variable_x))^2))

  # Se extrae la media x
  x.bar <- datos_summary %>%
          pull(x.bar)

  # Se extrae la varianza de x
  S.xx <- datos_summary %>%
          pull(S.xx)

  # Se obtiene el quantil de la distribución F(2, n - 2),
  # con un nivel de confianza alpha
  fa <- qf(alpha, 2, n - 2)

  # Se calculan las bandas de confianza
  resultado <- datos %>%
               mutate(y.hat = b0.hat + b1.hat * !! variable_x,
                      banda_superior = y.hat + sqrt(2*fa*s2.hat)*
                                       sqrt(1/n + (!! variable_x-x.bar)^2/S.xx),
                      banda_inferior = y.hat - sqrt(2*fa*s2.hat)*
                                       sqrt(1/n + (!! variable_x-x.bar)^2/S.xx))
  return(resultado)
}

airquality %>%
  as_tibble() %>%
  select(Ozone, Wind) %>%
  filter(!is.na(Ozone), !is.na(Wind)) %>%
  mutate(log2_Wind = log2(Wind)) %>%
  bandas_confianza_rls(variable_x = log2_Wind, variable_y = Ozone,
                       formula_rls = Ozone ~ log2_Wind) %>%
  head(10) %>%
  knitr::kable(format = "latex", booktabs = TRUE, longtable = TRUE, linesep = "",
               caption = "Estimaciones y bandas de confianza al 95\\%")
```

Tabla 1: Estimaciones y bandas de confianza al 95%

Ozone	Wind	log2_Wind	y.hat	banda_superior	banda_inferior
41	7.4	2.887525	54.251482	60.67174	47.831228
36	8.0	3.000000	49.873481	55.84000	43.906964

12	12.6	3.655352	24.364365	31.58362	17.145113
18	11.5	3.523562	29.494195	35.97804	23.010348
28	14.9	3.897240	14.949019	23.86323	6.034805
23	8.6	3.104337	45.812249	51.52193	40.102570
19	13.8	3.786596	19.255763	27.35102	11.160510
8	20.1	4.329124	-1.861734	10.65935	-14.382819
7	6.9	2.786596	58.180070	65.13505	51.225086
16	9.7	3.277985	39.053117	44.73954	33.366693



Ejercicio 5

(Sheater) Un estadístico colaboró en un proyecto de investigación con dos entomólogos. El análisis involucró el ajuste de modelos de regresión a grandes conjuntos de datos. Entre los tres escribieron y sometieron un manuscrito a una revista de entomología. El escrito contenía varias gráficas de dispersión mostrando la recta de regresión ajustada y las bandas de confianza 95% para la verdadera recta de regresión calculadas con los IC individuales, así como los datos observados. Uno de los revisores del manuscrito hizo la siguiente observación:

x	5	6	7	10	12	15	18	20
y	7.4	9.3	10.6	15.4	18.1	22.2	24.1	24.8

No puedo entender cómo el 95% de las observaciones cae fuera de las bandas de confianza 95% que se muestran en las figuras

Ejercicio 6

(Ross) Suponer que se tiene el siguiente conjunto de datos donde x representa la humedad de una mezcla fresca de un determinado producto y y la densidad del producto terminado.

Ajustar un modelo RLS a los datos anteriores y responder lo siguiente.

a

Reportar la estimación puntual de σ^2 e interpretar el resultado en cuanto a la utilidad del modelo RLS ajustado.

b

Reportar el IC 90% para σ^2 con los cuantiles simétricos y su longitud.

c

Indicar cuáles son los cuantiles que proporcionan el IC 90% para σ^2 de menor longitud.

d

Reportar el IC 90% para σ^2 de menor longitud y compararlo con el intervalo del inciso a).