Modelos de regresión, EJ2

Pérez Efremova, Daniel

Enero 2020

Índice

Índice de Códigos

Ejercicio 1

Apartado a

Para obtener inferencias a cerca de los parámetros debemos ajustar un modelo de regresión lineal múltiple (RLM) con y como variable respuesta y x_1, x_2 como variables regresoras, que son las dos últimas columnas de la matriz de datos X:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{U}$$

Esto está justificado por que el procedimiento para generar la variable respuesta es observando $y \sim N(\boldsymbol{X\beta}, \sigma^2\boldsymbol{I})$, lo cual hace que se cumplan las especificaciones de un modelo RLM.

Vamos a realizar tres tipos de inferencias:

1. *Estimación puntual*: estimaremos los coeficientes mediante el método de máxima verosimilitud[1].

   $$\hat{\boldsymbol{\beta}} = (\boldsymbol{X'X})^{-1}\boldsymbol{X'y}$$

   Además estimaremos los residuos mediante

   $$\boldsymbol{e} = \boldsymbol{Y} - \hat{\boldsymbol{Y}}$$

   para construir la desviación típica residual

   $$s_R = \frac{\boldsymbol{e'e}}{n-k-1}$$

   que utilizaremos más adelante.

2. *Estimación por intervalos*: estimaremos un intervalo de confianza para los coeficientes del modelo ya que por la definición de $\boldsymbol{y}$ se tiene que

   $$\hat{\beta}_i \sim N\left(\beta_i, \frac{\sigma}{VNE_{i.R}}\right) \Rightarrow t = \frac{\hat{\beta}_i}{\hat{s}_R/VNE_{i.R}} \sim t_{n-k-1} \tag{1}$$

   y podemos definir un intervalo de confianza para β_i basado en la t de Student:

   $$I = \left[\hat{\beta}_i \pm t_{n-k-1;\alpha/2}\hat{s}_R/VNE_{i.R}\right]$$

3. *Test de hipótesis*: Realizaremos distintos test de hipótesis acerca de los parámetros, sobre todo estamos interesados en los más usuales: el test de regresión individual sobre cada coeficiente basado en el estadístico t de (1) y el test general sobre todos los coeficientes basado en la F de Snedecor.

[1] que al cumplirse las especificaciones del modelo RLM coincide con el método de mínimos cuadrados

Vemos en el Código 1 los cálculos necesarios para obtener las estimaciones puntuales.

Código 1: Estimaciones puntuales

```
#definicion de los datos del enunciado
X = matrix(c(1,7,8,
1,3,10,
1,5,14,
1,4,12,
1,6,15,
1,2,18,
1,5,17), nrow=7, ncol=3, byrow=TRUE)
y = matrix(c(4,6,5,3,7,8,6), ncol=1)

# definicion de los datos en desviaciones
X_desv = t(apply(X[1:7, 2:3], 1, function(x) x - colMeans(X[1:7, 2:3])))
y_desv = y - mean(y)

#defincion de las dimensiones de los datos
n = dim(X)[1]
k = dim(X)[2]-1

#estimacion de los coeficientes en desviaciones y usual (con los 1)
beta = inv(t(X)%*%X)%*%t(X)%*%y
beta_desv = inv(t(X_desv)%*%X_desv)%*%t(X_desv)%*%y_desv

y_est = X%*%beta  # valores de la respuesta ajustados
e = y-y_est  #residuos

s_R = sqrt(t(e)%*%e/(n-k-1))[1,1] # desviacion tipica residual

print(c('beta0:', beta[1]))
print(c('beta1:', beta[2]))
print(c('beta2:', beta[3]))

[1] "beta0:"    "2.542595190000001"
[2] "beta1:"    "-0.17457588"
[3] "beta2:"    "0.28498206"
```

Ahora en el Código 2 vemos los intervalos de confianza para los coeficientes.

Código 2: Estimaciones por intervalos

```
# Ajuste de un modelo de regresion de x_1 sobre x_2: x_1 = alpha0 +
    alpha1*x_2
alpha1 = cov(X[1:7, 2], X[1:7, 3])/var(X[1:7, 3])
alpha0 =  mean(X[1:7, 2])-alpha1*mean(X[1:7, 3])
X_1_ajust = alpha0 + alpha1*X[1:7, 3]

# Ajuste de un modelo de regresion de x_2 sobre x_1: x_2 = ro0 + ro1*x_1
ro1 = cov(X[1:7, 2], X[1:7, 3])/var(X[1:7, 2])
ro0 =  mean(X[1:7, 3])-ro1*mean(X[1:7, 2])
X_2_ajust = ro0 + ro1*X[1:7, 2]

# Calculamos la varianza no explicada de los modelos
VNE_12 = sum((X[1:7, 2] - X_1_ajust)**2)
VNE_21 = sum((X[1:7, 3] - X_2_ajust)**2)

print('Intervalo de confianza para beta1:')
print(c(beta[2] - abs(qt(.025, df=n-k-1))*s_R/VNE_12, beta[2] + abs(qt
    (.025, df=n-k-1))*s_R/VNE_12))

print('Intervalo de confianza para beta2:')
print(c(beta[3] - abs(qt(.025, df=n-k-1))*s_R/VNE_21, beta[3] + abs(qt
    (.025, df=n-k-1))*s_R/VNE_21))
```

```
[1] "Intervalo de confianza para beta1:"
[2] -0.4558778  0.1067260
[3] "Intervalo de confianza para beta2:"
[4] 0.2224705 0.3474936
```

Finalmente vemos el resultado de los contrastes de significación individuales y el contraste de significación global.

El contraste significación individual tiene la forma:

$$H_0 : \beta_i = 0$$

$$H_1 : \beta_i \neq 0$$

Y su estadístico de contraste es (1) y el p-valor del test es:

$$p = 2P\{|t| < t_{n-k-1}\}$$

Por otro lado el contraste de significación global tiene la forma:

$$H_0 : \beta_i = 0 \, \forall$$

$$H_1 : \beta_i \neq 0 \, \forall$$

con estadístico de contraste

$$F = \frac{\hat{\boldsymbol{b}}'(\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{X}})\hat{\boldsymbol{b}}}{k\hat{s}_R} \sim F_{k,n-k-1}$$

con $\tilde{\boldsymbol{X}}$ la matriz de datos centrados y $\hat{\boldsymbol{b}}$ la estimación de los coeficientes con desviaciones (sin el 1). El p-valor del test es:

$$p = P\{F < F_{k,n-k-1}\}$$

Vemos estos cálculos y su resultado en el Código 3.

Código 3: Contrastes sobre los coeficientes

```
#contraste individual
t1 = beta[2]/(s_R/sqrt(VNE_12))
t2 = beta[3]/(s_R/sqrt(VNE_21))

pvalor1 = 2*pt(abs(t2), n-k-1, lower.tail = FALSE, log.p = FALSE)
pvalor2 = 2*pt(abs(t1), n-k-1, lower.tail = FALSE, log.p = FALSE)

print(c('pvalor del test beta1:', pvalor1))
print(c('pvalor del test beta2:', pvalor2))

# contraste general
F = (t(beta_desv) %*% (t(X_desv) %*% X_desv) %*% beta_desv/(k*(s_R**2)))
    [1,1]  # estadistico de contraste
pvalor_cgeneral = pf(F, k, n-k-1, lower.tail=FALSE)

print(c('pvalor del contraste general de regresion:', pvalor_cgeneral))

[1] "pvalor del test beta1:" "0.198236476187164"
[2] "pvalor del test beta2:" "0.679393085573287"
[3] "pvalor del contraste general de regresion:" "0.271871757531989"
```

A modo de comprobación podemos ajustar un modelo lineal con la función interna de R y comparar con nuestros resultados en los Códigos 1,2 y 3.

```
summary(lm(y ~ X[1:7, 2:3]))

Residuals:
1       2       3       4       5       6       7
0.3996  1.1313 -0.6595 -2.2641  1.2301  0.6769 -0.5144

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    2.5426     3.6340   0.700    0.523
X[1:7, 2:3]1  -0.1746     0.3924  -0.445    0.679
X[1:7, 2:3]2   0.2850     0.1850   1.541    0.198

Residual standard error: 1.52 on 4 degrees of freedom
Multiple R-squared:  0.4786,    Adjusted R-squared:  0.2179
F-statistic: 1.836 on 2 and 4 DF,  p-value: 0.2719
```

Apartado b

La nueva observación $(3, 15)$ se encuentra dentro del rango de datos observados, por lo que la estimación del valor de la respuesta puede hacerse mediante el valor medio estimado por el modelo (interpolación).

El modelo estimará el valor promedio de la respuesta y condicionada a la observación (x_1, x_2) mediante:

$$\hat{y} = E[y|x_1, x_2] = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$

esta estimación es centrada en el verdadero valor de la respuesta. Vemos en el Código 4 la estimación de la respuesta.

Código 4: Predicción del valor de la respuesta para la nueva observación (3, 15)

```
y = c(beta) %*% c(1, 3, 15)
print(y[1,1])

[1] 6.293598
```

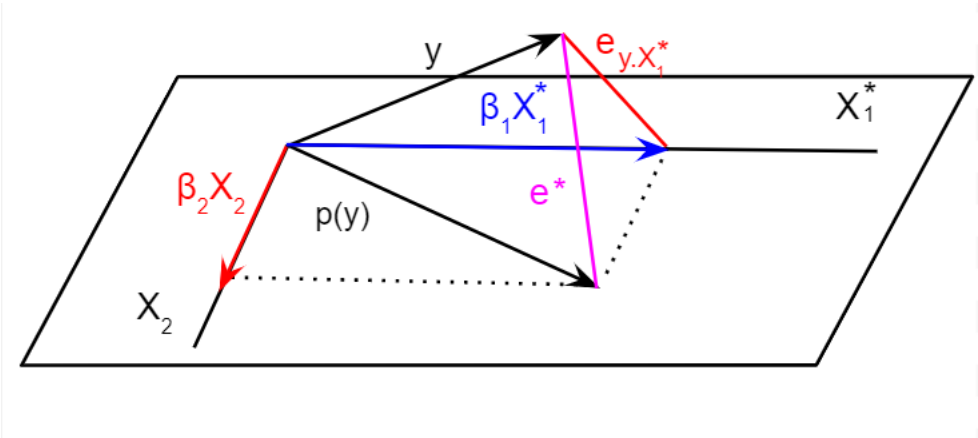
Ejercicio 2

Para justificar que las dos estimaciones son iguales, vemos un argumento geométrico en la Figura 1.

Al estimar por mínimos cuadrados el coeficiente de X_2 en la regresión $y = \beta_2 X_2 + e_{y,2}$, lo que hacemos es proyectar ortogonalmente el vector y sobre la dirección de X_2, de manera que $\beta_2 X_2 \perp e_{y,2}$.

Veamos ahora como obtener el vector $\beta_2 X_2$ por otro camino haciendo uso de la regresión múltiple. Como X_1^* es el residuo de la regresión de X_1 sobre X_2, no puede contener información acerca de X_2 y por tanto, $X_1^* \perp X_2$. Consideremos

Figura 1: Interpretación geométrica de la regresión propuesta



entonces el plano $P = \{X_2, X_1^*\}$. Al ajustar la regresión de y sobre X_1^*, tenemos que el residuo $e_{y.x_1^*}$ no contiene información de X_1^*, luego $e_{y.x_1^*} \perp X_1^*$, y también $e_{y.x_1^*} \perp X_2$, con $e_{y.X_1^*} \notin P$, por tanto, denotando $P_{X_2}^{\perp}$ a la proyección ortogonal sobre la dirección de X_2, tenemos:

$$\beta_2 X_2 = P_{X_2}^{\perp}(y) = P_{X_2}^{\perp}(y - \beta_1 X_1^*) + P_{X_2}^{\perp}(\beta_1 X_1^*) = P_{X_2}^{\perp}(e_{y.X_1^*}) + \vec{0} = P_{X_2}^{\perp}(e_{y.X_1^*}) = \beta_2 X_2$$

donde $P_{X_2}^{\perp}(y) = P_{X_2}^{\perp}(e_{y.X_1^*})$ es el resultado que se quiere probar. La proyección ortogonal de $\beta_1 X_1^*$ en la dirección de X_2 es nula, puesto que los vectores son ortogonales y están en el mismo plano. Además, la última igualdad se debe a la regresión múltiple por etapas.

Ejercicio 3

El intervalo de confianza para el promedio de la variable respuesta se deduce directamente de que:

$$\frac{\hat{y}_h - m_h}{\sigma\sqrt{h_{hh}}} \sim N(0,1)$$

por lo que, estimando σ con la varianza residual $\hat{s}_R$, obtenemos:

$$\frac{\hat{y}_h - m_h}{\hat{s}_R\sqrt{h_{hh}}} \sim t_{n-k-1}$$

siendo n la cantidad de datos y k la cantidad de regresores. La expresión del intervalo de confianza es:

$$I = \left[\hat{y}_h \pm t_{(n-k-1;\,\alpha/2)}\frac{\hat{s}_R}{\sqrt{\hat{n}_h}}\right]$$

si además tenemos s observaciones con $x = x_h$, entonces $\hat{n}_h = s\hat{n}_x$, siendo $\hat{n}_x$ el número de observaciones equivalente del punto x, y el intervalo queda:

$$I = \left[\hat{y}_h \pm t_{(n-k-1;\,\alpha/2)} \frac{\hat{s}_R}{\sqrt{s\hat{n}_x}} \right]$$

de lo que podemos deducir algo interesante, y es que, a medida que crece el número de observaciones s con $\hat{s}_R$ constante, la longitud del intervalo de confianza:

$$l(I) = 2t_{(n-k-1;\,\alpha/2)} \frac{\hat{s}_R}{\sqrt{s\hat{n}_x}}$$

tiende a disminuir, algo bastante razonable si pensamos en que, al crecer s, tenemos mucha información acerca del comportamiento de la variable respuesta y en el punto x, y por tanto, la recta de regresión pasará muy cerca del verdadero valor de la media de la respuesta condicionada a x, haciendo que $\hat{y}_h - m_h \longrightarrow 0$.

En el caso extremo en que $\hat{s}_R$ se mantenga constante, para un α fijo y aumente s, tendremos que:

$$\lim_s l(I) = \lim_s 2t_{(n-k-1;\,\alpha/2)} \frac{\hat{s}_R}{\sqrt{s\hat{n}_x}} = \lim_s C \cdot \frac{1}{\sqrt{s}} = 0$$

obteniendo un intervalo de confianza para la respuesta extremadamente preciso, tanto, que hablaríamos ya de una estimación puntual.