

Guía ejercicios 2

Problema 1

Consideremos el siguiente modelo $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i$. para una muestra $n=8$ observaciones. Suponga que se han obtenido los siguientes resultados.

| variable | Estimación |
|----------|------------|
| Cte | 9.05 |
| X1 | 0.52 |
| X2 | 0.24 |

$$(X'X)^{-1} = \begin{vmatrix} 6.35 & -0.03 & -0.02 \\ -0.03 & 1.52 \times 10^{-3} & -6.5 \times 10^{-4} \\ -0.02 & -6.5 \times 10^{-4} & 5.38 \times 10^{-4} \end{vmatrix}$$

$$(X'X) = \begin{vmatrix} 8 & 672 & 1176 \\ 672 & 57822 & 100453 \\ 1176 & 100453 & 176758 \end{vmatrix}$$

F de Fisher = 15.05 y $\hat{S}^2 = 33.82$

Al respecto. Calcule

1.1) El coeficiente de correlación múltiple.

1.2) Construya intervalos de confianza individuales para β_1 y β_2 . ¿Aporta la variable x_1 al modelo?. ¿Aporta la variable x_2 al modelo?.

Solución

1.1)F-Fisher: $F = 15,05$

$$Se \text{ sabe } F = \frac{R^2/(r-1)}{(1-R^2)/(n-r)}$$

$$r = p + 1 = Rg(x)$$

$$r = 3$$

$$n = 8$$

$$\begin{aligned}
(r-1)(1-R^2)F &= R^2(n-r) \\
(r-1)F - (r-1)FR^2 &= R^2(n-r) \\
\Rightarrow R^2 &= \frac{(r-1)F}{(n-r) + (r-1)F} \\
R^2 &= \frac{2 \cdot 15,05}{5 + 2 \cdot 15,05} = \frac{30,1}{35,1} = \frac{6,02}{7,02} \\
R^2 &= 0,8575
\end{aligned}$$

1.2

$$\begin{aligned}
\frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}_{\hat{\beta}_j}} &\rightarrow t_{n-r} \\
P(a \leq \beta_j \leq b) &= 1 - \alpha \quad \alpha = 0,05 \\
P\left(\underbrace{\frac{\hat{\beta}_j - b}{\hat{\sigma}_{\hat{\beta}_j}}}_{V_1} \leq t_{n-r} \leq \underbrace{\frac{\hat{\beta}_j - a}{\hat{\sigma}_{\hat{\beta}_j}}}_{V_2}\right) &= 1 - \alpha
\end{aligned}$$

Simetría: $u_1 = -u_2$

$$\begin{aligned}
b &= \hat{\beta}_j + \hat{\sigma}_{\hat{\beta}_j} \cdot 2,571 \\
\hat{\sigma}_{\hat{\beta}_j}^2 &= \hat{s}^2 \cdot (x^t x)^{-1}_{jj} \quad \hat{s}^2 = 33,82
\end{aligned}$$

$$\text{Para } \hat{\beta}_1 = 0,52 \quad \hat{\sigma}_{\hat{\beta}_1}^2 = 33,82 \cdot 1,52 \cdot 10^{-3} = 0,0514$$

$$\begin{aligned}
a &= 0,52 - 2,571 \cdot 0,2267 = -0,0628 \\
b &= 0,52 + 2,571 \cdot 0,2267 = 1,1028 \\
\beta_1 &\in [-0,0628; 1,1028] \quad \hat{\sigma}_{\hat{\beta}_2}^2 = 33,82 \cdot 5,38 \cdot 10^{-4} = 0,018 \\
a &= 0,24 - 2,571 \cdot 0,1349 = -0,1068 \\
b &= 0,24 + 2,571 \cdot 0,1349 = 0,5868 \\
\beta_2 &\in [-0,1068; 0,5868]
\end{aligned}$$

Para verificar validez de cada variable se realiza test t-Student.

Para x_1 :

$$H_0: \beta_1 = 0$$

$$\text{Bajo } H_0: q_1 = \frac{\hat{\beta}_1}{\hat{\sigma}_{\hat{\beta}_1}} = \frac{0,52}{0,2267} = 2,294$$

$$\begin{aligned} P(|t_5| \geq q_1) &= P(t_5 \leq -q_1) + P(t_5 \geq q_1) \\ &= 2P(t_5 \geq q_1) = 2P(t_5 \geq 2,294) \\ &= 2 \cdot 0,0375 \cong 0,075 > 0,05 \end{aligned}$$

\Rightarrow No se rechaza H_0

\Rightarrow La variable x_1 no es significativa para el modelo.

Para x_2 :

$$H_0: \beta_2 = 0$$

$$\text{Bajo: } H_0: q_2 = \frac{\hat{\beta}_2}{\hat{\sigma}_{\hat{\beta}_2}} = \frac{0,24}{0,1349} = 1,779$$

$$\begin{aligned} P(|t_s| \geq q) &= 2P(t_s \geq q_2) = 2P(t_5 \geq 1,779) \\ &\approx 2 \cdot 0,075 = 0,15 > 0,05 \end{aligned}$$

\Rightarrow No se rechaza $H_0 \Rightarrow$ La variable x_2 no es significativa para el modelo.

Las conclusiones anteriores pueden no ser muy correctas, debido a la poca cantidad de datos en la muestra ($n = 8$).

Según la parte anterior, la región de confianza de t_q :

$$P(F_{p+1, n-p-1} \leq F_c) = 1 - \alpha \quad \alpha = 0,05$$

$$\Leftrightarrow P(F_{2,5} \leq F_c) = 0,95$$

$$P(F_{2,5} > F_c) = 0,05$$

$$P(F_{2,5} > F_c) = 0,05$$

$$\Rightarrow F_c = 5,78$$

Explicación:

Debido a que se trabaja con una submatriz de $(x^b x)$ real se debe reducir el rango de la matriz a 2, en vez de tres, lo que podría pensarse (se le quita la primera fila y la primera columna a $x^b x$).

Así la región de confianza para (β_1, β_2) es la solución de:

$$(\beta_1 - 0,52; \beta_2 - 0,24) \overbrace{\begin{bmatrix} 57822 & 100453 \\ 100453 & 176758 \end{bmatrix}}^{\text{Nueva matriz}(x^t x)} \begin{pmatrix} \beta_1 - 0,52 \\ \beta_2 - 0,24 \end{pmatrix} \\ \leq \underbrace{33,83}_{\bar{s}^2} \cdot \underbrace{5,78}_{F_c} \cdot \underbrace{2}_{p+1}$$

$$57822(\beta_1 - 0,52)^2 + 200906(\beta_1 - 0,52)(\beta_2 - 0,24) + 176758(\beta_2 - 0,24)^2 \leq 391,075$$

lo que equivale a una región elíptica en el plano. (β_1, β_2) , centrada en el punto $(0,52; 0,24)$.

Problema 2

Sea una muestra bivariada $\{(x_i, y_i) / i = 1, \dots, n\}$

Si x toma solamente el valor 1 o 0. (x define dos poblaciones diferentes). Se llaman \bar{y}_1 la media de la población definida por $x = 1$ e \bar{y}_0 para $x = 0$. Se plantea el modelo lineal: $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$. Dé el estimador de mínimos cuadrados y deduzca que el modelo equivale a:

$$E(y) = \bar{y}_1 \text{ si } x = 1 \text{ y } E(y) = \bar{y}_0 \text{ si } x = 0.$$

Solución

$$X^t X = \begin{pmatrix} 1 & \dots & 1 \\ x_1 & \dots & x_n \end{pmatrix} \begin{pmatrix} 1 & x_1 \\ \cdot & \cdot \\ 1 & x_n \end{pmatrix} = \begin{pmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{pmatrix}$$

$$\sum x_i = n_1 \quad \sum x_i^2 = n_1 \\ \Rightarrow (X^t X)^{-1} = \frac{1}{n_1 n - n_1^2} \begin{pmatrix} n_1 & -n_1 \\ -n_1 & n \end{pmatrix} = \frac{1}{n_1(n - n_1)} \begin{pmatrix} n_1 & -n_1 \\ -n_1 & n \end{pmatrix} = \frac{1}{n_1 n_0} \begin{pmatrix} n_1 & -n_1 \\ -n_1 & n \end{pmatrix}$$

$$X^t Y = \begin{pmatrix} 1 & \cdot & \cdot & 1 \\ x_1 & \cdot & \cdot & x_n \end{pmatrix} \begin{pmatrix} y_1 \\ \cdot \\ \cdot \\ y_n \end{pmatrix} = \begin{pmatrix} \sum y_i \\ \sum x_i y_i \end{pmatrix} = \begin{pmatrix} n\bar{y} \\ n_1 \bar{y}_1 \end{pmatrix}$$

Estimando los parámetros

$$\hat{\beta} = (X^t X)^{-1} X^t Y = \frac{1}{n_0 n_1} \begin{pmatrix} n_1 n \bar{y} - n_1^2 \bar{y}_1 \\ -n_1 n \bar{y} + n n_1 \bar{y}_1 \end{pmatrix}$$

$$\hat{\beta}_0 = \frac{n \bar{y} - n_1 \bar{y}_1}{n_0}, \text{ pero : } n \bar{y} = n_1 \bar{y}_1 + n_0 \bar{y}_0 \Rightarrow \hat{\beta}_0 = \bar{y}_0$$

$$\hat{\beta}_1 = \frac{-n\bar{y} + n\bar{y}_1}{n_0} = \frac{n\bar{y}_1 - n_1\bar{y}_1 - n_0\bar{y}_0}{n_0} = \frac{n_0}{n_0}(\bar{y}_1 - \bar{y}_0) = \bar{y}_1 - \bar{y}_0$$

Si $x = 1$

$$E(y) = \hat{\beta}_0 + \hat{\beta}_1 = \bar{y}_0 + \bar{y}_1 - \bar{y}_0 = \bar{y}_1$$

Si $x = 0$

$$E(y) = \begin{cases} \bar{y}_1 & \text{si } x = 1 \\ \bar{y}_0 & \text{si } x = 0 \end{cases}$$

a) Dé la expresión del R^2 en función de la varianza empírica de y , \bar{y}_1 e \bar{y}_0 . Muestre que el test $H_0 : \beta_1 = 0$ equivale a un test de comparación de medias.

$$R^2 = \frac{(\sum x_i y_i - n\bar{x}\bar{y})^2}{(\sum x_i^2 - n\bar{x}^2)(\sum y_i^2 - n\bar{y}^2)} \quad \sum x_i y_i = n_1 \bar{y}_1$$

$$n\bar{x}\bar{y} = n \frac{n_1}{n} \bar{y}$$

$$R^2 = \frac{n_1^2 (\bar{y}_1 - \bar{y})^2}{n(n_1 - \frac{n_1^2}{n}) \text{var}(y)} = \frac{n_1^2 (\bar{y}_1 - \bar{y})^2}{n_1 n_0 \text{var}(y)} = \frac{n_1}{n_0} \frac{(\bar{y}_1 - \bar{y})^2}{\text{var}(y)}$$

$H_0 : \beta_1 = 0 \Rightarrow \frac{\hat{\beta}_1}{\hat{\sigma}_{\beta_1}} \sim t_{n-2} \Rightarrow \frac{\bar{y}_1 - \bar{y}_0}{\hat{\sigma}_{\beta_1}} \sim t_{n-2}$ lo que equivalente al test de comparación de medias con

$$H_0 : \mu_1 = \mu_0$$

$$H_1 : \mu_1 \neq \mu_0$$

Problema 3

Una compañía de teléfonos celulares estudia la permanencia de sus clientes con el objeto de emprender algunas acciones. Se recogieron los datos de permanencia mensuales durante un periodo de 3 años, suponiendo M el total de clientes al mes 0 se definen los datos recogidos como $\{(x_i, y_i) / i = 0, 1, \dots, 35\}$ donde y_i es el número de clientes en el mes i (que quedan del total $M = y_0$ inicial, e x_i es el mes ($x_i = i$)). Para explicar la permanencia de los clientes, se propone un modelo exponencial:

$$y = \alpha e^{\beta x}$$

- a) Interprete el modelo. En particular considere los cuocientes $\frac{y_{i+1}}{y_i}$. Interprete el coeficiente β precise su signo e interprete α
- b) Transforme el modelo de manera de encontrar un modelo lineal
- c) Los resultados del modelo lineal se presentan en la siguiente tabla. Complete la tabla. Opine sobre la validez local de los coeficientes del modelo. Dé los grados de libertad de las t-student.

| Variable | Estimación | Desv típica estimación | t-student | Pvalor |
|-----------|------------|------------------------|-----------|--------|
| Constante | 8.5092 | | 196.3138 | 0.000 |
| x | -0.0284 | 0.0020 | | 0.000 |

Coeficiente de correlación entre los valores observados y estimados: 0.92

- d) Deduzca el modelo de permanencia exponencial estimado. Comente. Estime el número de clientes del mes 0 inicial.
- e) Después de cuantos meses se esperan que permanezcan solamente 1500 clientes.

Solución

- a) $\frac{y_{i+1}}{y_i}$ es la tasa de deserción de los clientes. Al realizar el cuociente vemos que
- $$\frac{y_{i+1}}{y_i} = \frac{\alpha e^{\beta X_{i+1}}}{\alpha e^{\beta X_i}} = \frac{e^{\beta(i+1)}}{e^{\beta i}} = e^{\beta}$$
- que no depende de i. Es decir, la tasa de deserción es constante en el tiempo.

Luego $\beta = \ln\left(\frac{y_{i+1}}{y_i}\right)$ que tiene que ser negativo ya que los y_i son decrecientes.

Por otro lado $y_0 = \alpha e^{\beta \cdot 0} = \alpha = M$ que es el número de clientes inicial > 0

- b) Si aplicamos ln a la función y tenemos que $\ln(y) = \ln(\alpha) + \beta X = \ln(M) + \beta X$

- c) Utilizando el estadístico del test local $t_{n-r} = \frac{\hat{\beta}_j}{\hat{\sigma}_{\beta_j}}$ lo que permite completar la tabla

es así como $\hat{\sigma}_{\beta_0} = 0.0433$ y $t(\beta_1) = -13.9119$

tenemos 36 pares de datos (meses) y 2 variables que estimar por tanto los grados de libertad de las t-student son 34. Los coeficientes son claramente significativos (pvalor=0) bastante menor que 0.05, por lo que podemos concluir que la variable X aporta al modelo. Más aún, el coeficiente de correlación entre los datos observados y estimados es casi igual a 1, por lo que existe una relación bastante lineal entre ambos.

d) Tenemos que $\ln(y) = \ln(\alpha) + \beta X = \beta_0 + \beta_1 X$, on β_0 y β_1 obtenidos de la tabla,

$$\text{entonces } \beta = \beta_1 = -0.0284 \quad y \quad \ln(\alpha) = \beta_0 = 8.5092 \Rightarrow \alpha = e^{8.5092}$$

$$\Rightarrow y = e^{8.5092} * e^{-0.0284 X} = 4960.3e^{-0.0284 X}$$

Por tanto, el número de clientes el primer mes se estima como $\alpha = e^{8.5092} = 4960$

f) Para un valor de $X=1500$ clientes

$$X = \frac{1}{\beta}(\ln(y) - \ln(\alpha)) = \frac{\ln(1500) - 8.5092}{0.0284} \approx 42 \text{ meses.}$$

Problema 3

Se estudia la relación del peso con otras medidas antropométricas de 30 niñas de 10 años, la variable y representa el peso, x_1 la talla, x_2 la circunferencia del tórax, x_3 la circunferencia del brazo y x_4 la circunferencia de la pierna. Se presentan la matriz de las correlaciones de todas las variables en la tabla 2 y los resultados de 3 modelos de regresión lineal distintos en las tablas 3, 4 y 5.

3.1) Analice los resultados de la regresión dados en la tabla 3. Considerando la tabla 2, ¿qué opina del efecto del tórax en el modelo?

3.2) Dé un intervalo de confianza para el coeficiente β_1 de la talla. Realice el test

$$H_0: \beta_1 = 4 \quad \text{contra} \quad H_1: \beta_1 < 4$$

3.3) En el modelo de la tabla 4, se eliminó la variable “Torax” y en la tabla 5 se eliminó la variable “Talla” del modelo de la tabla 3. Examine y compare las regresiones.

3.4) Dé los grados de libertad de la F de Fisher y de las T-student dadas en la tabla 4.

3.5) ¿Cuáles son los supuestos usuales sobre los errores de un modelo lineal? ¿Cómo se relacionan estos supuestos con las propiedades de los β ?

Tabla 2 : matriz de correlaciones

| | Peso | Talla | Tórax | Brazo | Pierna |
|--------|------|-------|-------|-------|--------|
| Peso | 1.00 | 0.67 | 0.82 | 0.75 | 0.83 |
| Talla | 0.67 | 1.00 | 0.45 | 0.14 | 0.47 |
| Tórax | 0.82 | 0.45 | 1.00 | 0.76 | 0.65 |
| Brazo | 0.75 | 0.14 | 0.76 | 1.00 | 0.61 |
| Pierna | 0.83 | 0.47 | 0.65 | 0.61 | 1.00 |

Tabla 3: Modelo $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$

| Variable | Coefficiente | Desv. típica | t-student | P(X>T) |
|-----------|--------------|--------------|-----------|--------|
| Constante | -6130.72 | 584.06 | -10.50 | 0.000 |
| Talla | 3.54 | 0.54 | 6.51 | 0.000 |
| Tórax | 1.33 | 1.05 | 1.28 | 0.214 |
| Brazo | 7.58 | 1.62 | 4.67 | 0.000 |
| Pierna | 7.43 | 1.70 | 4.36 | 0.000 |

Coefficiente de correlación múltiple : 0.94

F- Fisher para las tres variables : 95,24 con $P(X>F)=0.0001$

Tabla 4: Modelo $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$

| Variable | Coefficiente | Desv. típica | t-student | P(X>T) |
|-----------|--------------|--------------|-----------|--------|
| Constante | -6134.00 | 586.03 | -10.30 | 0.000 |
| Talla | 3.84 | 0.50 | 7.74 | 0.000 |
| Brazo | 8.98 | 1.20 | 7.47 | 0.000 |
| Pierna | 7.74 | 1.70 | 4.54 | 0.000 |

Coefficiente de correlación múltiple : 0.93

F- Fisher para las tres variables : 123.47 con $P(X>F)=0.0000$

Tabla 5: Modelo $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$

| Variable | Coefficiente | Desv. típica | t-student | P(X>T) |
|-----------|--------------|--------------|-----------|--------|
| Constante | -3515.57 | 582.74 | -5.15 | 0.009 |
| Tórax | 4.28 | 1.52 | 2.81 | 0.200 |
| Brazo | 3.08 | 2.36 | 1.30 | 0.000 |
| Pierna | 11.48 | 2.55 | 4.50 | 0.000 |

Coefficiente de correlación múltiple : 0.83

F- Fisher para las tres variables : 43.54 con $P(X>F)=0.0000$

Solución:

3.1) De la matriz de correlaciones vemos que todas las variables están altamente correlacionadas con la variable y, por lo tanto, se espera que todas ellas vayan en el modelo y sean estadísticamente significativas. Sin embargo, la variable Tórax está altamente relacionada con las variables Brazo y Pierna, por lo que es probable que alguna de estas tres variables no vaya en el modelo por no cumplir con el supuesto de independencia de las variables x.

De hecho si incluimos todas las variables en el modelo, vemos en la tabla 3 que la variable Tórax no es estadísticamente significativa. Por otro lado, el pvalor de la Fisher es nulo y el coeficiente de correlación múltiple bastante alto.

3.2) Para calcular el intervalo de confianza para β_1 , tenemos que

$P(t_{25} < -2.06) = P(t_{25} > 2.06) = 0.025$, el intervalo queda entonces como

$$(3.54 - 2.06 \cdot 0.54, 3.54 + 2.06 \cdot 0.54) = (2.43, 4.65)$$

El test de hipótesis $H_0 : \beta_1 = 4$ contra $H_1 : \beta_1 < 4$ tiene como región crítica para un

error de tipo I del 5% : $P\left(\frac{\hat{\beta}_1 - 4}{\hat{\sigma}_{\beta_1}} < -1.71\right) = P(\hat{\beta}_1 < 4 - 1.71 \cdot 0.54) = P(\hat{\beta}_1 < 3.08) = 0.05$

Como $\hat{\beta}_1 = 3.54$ no pertenece a la región crítica y no se rechaza H_0 .

Si calculamos el pvalor $P\left(\frac{\hat{\beta}_1 - 4}{\hat{\sigma}_{\beta_1}} < \frac{3.54 - 4}{0.54}\right) = P(t_{25} < -0.852) = 0.20$ que es demasiado

alto para rechazar la hipótesis nula.

3.3) El modelo de la tabla 4 parece el más adecuado ya que conserva un coeficiente de correlación múltiple alto y tiene todos sus coeficientes significativos.

3.4) En la tabla 4, la F de Fisher tiene 3 y 26 grados de libertad y la t- student tiene 26 grados de libertad.

3.5) Los supuestos usuales son los que los errores son normales de media nula, de misma varianza e independientes entre sí. El hecho que $E(\varepsilon) = 0$ implica que los $\hat{\beta}_j$ son insesgados. La independencia de los errores implica que los $\hat{\beta}_j$ son de mínima varianza.