Taller Repaso 01

martes, 13 de mayo de 2025 9:10 p. m.

1. Bajo los supuestos del modelo, el estimador $\hat{\beta}$ por MCO coincide con el estimador de máxima verosimilitud.

- (A) Verdadera, porque ambos procedimientos minimizan la suma de cuadrados.
- (B) Verdadera, si y solo si se asume normalidad de los errores.
- (C) Falsa, ya que los métodos son conceptualmente distintos.
- (D) Falsa, porque MCO requiere menos supuestos que máxima verosimilitud.

$Y = \beta_0 + \beta_1 X_{i1} + \dots + \beta_2 X_{in} + \varepsilon_i$

$\underline{Y_n} = X\underline{\beta} + \underline{\varepsilon_n}$

$\hat{Y}_n = X\hat{\beta}$: (1). MCO; (2). MLE

(1). Método matemático:

. minimizar: $(Y_n - X\underline{\beta})^T (Y - X\underline{\beta}) = SSE$

$\hat{\beta} = (X^T X)^{-1} X^T Y_n$

. $\hat{\sigma}$: MSE: $\frac{SSE}{n-p}$ (insesgado); (2). Método estadístico: Distribución

. maximizar la observación de una muestra → $\hat{\sigma} = \frac{SSE}{n}$

2. Rechazar $H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$ implica que el modelo es significativo.

- (A) Verdadera, y puede comprobarse con el estadístico F.
- (B) Verdadera, solo si el número de predictores es mayor a 2.
- (C) Falsa, ya que el test F solo prueba significancia individual.
- (D) Falsa, debe emplearse el estadístico t para esta prueba.

3. La matriz H y $I - H$ son simétricas e idempotentes.

- (A) Verdadera, dado que H es una proyección ortogonal.
- (B) Falsa, solo H es idempotente. ✗
- (C) Verdadera, siempre y cuando X tenga rango completo.
- (D) Falsa, ninguna tiene esas propiedades. ✗

Matriz es idempotente: $H^n ; n \in \mathbb{N} = H$

$H = X(X^T X)^{-1} X^T$

4. El estadístico F para evaluar la regresión global es:

$$F = \frac{SSR/k}{SSE/(n-k)} \sim F_{k,n-k}$$

$F = \frac{SSR/k}{SSE/(n-p)} \sim f_{k,n-p}$

- (A) Verdadera, representa la relación entre variabilidad explicada y residual.
- (B) Falsa, el denominador debe ser $n - p$.
- (C) Falsa, porque el numerador es incorrecto.
- (D) Verdadera, pero solo con dos regresores.

$p = k+1$

5. Una observación es influencial si cumple: $D_i > 1, |DFBETAS_j(i)| > 2/\sqrt{n}, |DFFITS_i| > 2\sqrt{p/n}$.

- (A) Verdadera, esos umbrales indican influencia significativa.
- (B) Falsa, basta con que uno se cumpla.
- (C) Verdadera solo si n > 50.
- (D) Falsa, esos umbrales no son estándares.

*El material asociado a este taller puede encontrarse en el repositorio del curso, **(https://github.co tsssach/Estadistica-II)**

1. Verifique los supuestos del modelo de regresión lineal múltiple ajustado:

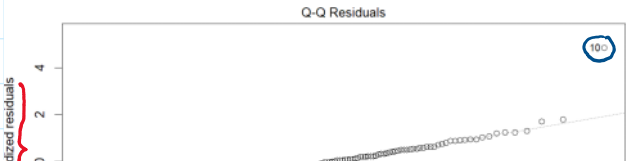
$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \varepsilon_i \; ; \quad \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$$

- Verifique: normalidad (prueba gráfica y prueba Shapiro-Wilk), homocedasticidad, independencia.

Y_i: Promedio (nota)
X_1: Horas estudio
X_2: Estrés
X_3: Sueño

(a). Normalidad: $\begin{cases} H_0: \varepsilon_i \sim Normal \\ H_1: \varepsilon_i \not\sim Normal \end{cases}$, $i = 1, \dots, n$.

. Realizar inferencia sobre errores: $\hat{\varepsilon}_i \rightarrow \varepsilon_i$

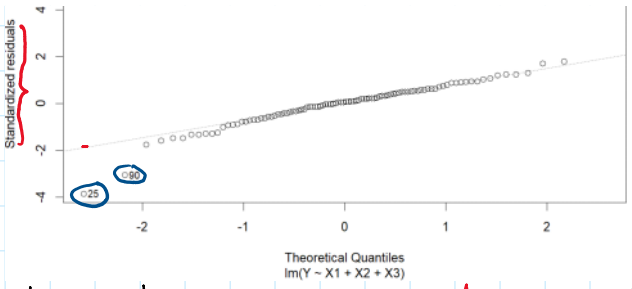



```
> shapiro.test(modelo$residuals)

        Shapiro-Wilk normality test

data:  modelo$residuals
W = 0.90719, p-value = 3.13e-06
```

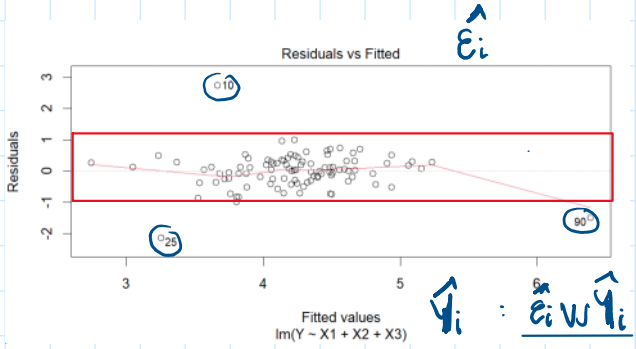
$< \alpha = 0,05$




```
data:  modelo$residuals
W = 0.90719, p-value = 3.13e-06
```
< $\alpha = 0,05$

. A un nivel de significancia del 5%, se puede establecer que existe evidencia suficiente para rechazar H_0 → Errores NO tienen una distribución normal.

(b). Media cero: Siempre se cumple; (c). Independencia: Asumimos que se cumple.
(d). Varianza constante:



$\hat{\varepsilon}_i$ vs $\hat{Y}_i$: $\hat{\varepsilon}_i$ vs $\hat{Y}_i$

- Verificar de forma gráfica: puntos atípicos e influenciales
- Verificar con los métodos analíticos
- Si hay varianza constante

2. Determine la significancia global del modelo ajustado usando el test F.

$$F = \frac{SSR/k}{MSE} \sim f_{k,n-p} \qquad P(f_{k,n-p} > F_{cal})$$

$H_0: \beta_1 = \dots = \beta_k = 0$
$H_1:$ Algún $\beta_j \neq 0;\ j = 1,\dots,k$

$H_0: \beta_1 = \beta_2 = \beta_3 = 0$
$H_1:$ Algún $\beta_j \neq 0;\ j = 1,\dots,3$

(3) métodos distintos:
(2). Sumas de cuadrados extra:

MF: $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \varepsilon_i$
MR: $Y_i = \beta_0 + \varepsilon_i$

(1). Regresión
(2). Errores

$$F = \frac{[SSR(MF) - SSR(MR)]/v}{MSE(MF)} = \frac{[SSE(MR) - SSE(MF)]/v}{MSE(MF)}$$

$$F = \frac{[SSR(\beta_0,\beta_1,\beta_2,\beta_3) - SSR(\beta_0|\beta_1,\beta_2,\beta_3)]/v}{MSE(\beta_0,\beta_1,\beta_2,\beta_3)} = \frac{[SSE(\beta_0|\beta_1,\beta_2,\beta_3) - SSE(\beta_0,\beta_1,\beta_2,\beta_3)]/v}{MSE(\beta_0,\beta_1,\beta_2,\beta_3)}$$

v = # de coeficientes en H_0 ; v = 3

```
> truco <- anova(modelo_reducido, modelo)
> truco
Analysis of Variance Table

Model 1: Y ~ 1
Model 2: Y ~ X1 + X2 + X3
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     99 55.931
2     96 31.335  3    24.596 25.117 4.425e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

SSE

Pval < α : A un nivel de sig. 5%, que la regresión es significativa (Al menos un β_j es distinto de cero)

- Prueba lineal general: $L\beta = 0$

$H_0: \beta_1 = 0,\ \beta_2 = 0,\ \beta_3 = 0$ vs $H_1: \beta_1 \neq 0 \vee \beta_2 \neq 0 \vee \beta_3 \neq 0$

$\beta_1 = 0 \rightarrow \begin{cases} 0\beta_0 + 1\beta_1 + 0\beta_2 + 0\beta_3 = 0 \\ 0\beta_0 + 0\beta_1 + 1\beta_2 + 0\beta_3 = 0 \\ 0\beta_0 + 0\beta_1 + 0\beta_2 + 1\beta_3 = 0 \end{cases} : \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$

L

↳ Sistema ecuaciones :

- r = # filas no nulas II
(r = 3)

$$\frac{[SSR(MF) - SSR(MR)]/r}{MSE(MF)}$$

$\frac{[SSR(MF) - SSR(MR)]/r}{MSE(MF)}$ $r = 3$

3. Estime los parámetros individuales, interprete sus p-valores e incluya intervalos de confianza al 95

$\hat{\beta}$: (MCO)

IC: $\hat{\beta}_j \pm t_{\alpha/2, n-p} \cdot \sqrt{Var(\hat{\beta}_j)}$: $P(|t_{n-p}| > |T_{cal}|)$; $T = \frac{\hat{\beta}_j - 0}{\sqrt{Var(\hat{\beta}_j)}} \sim t_{n-p}$

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.24701    0.48876   6.643 1.85e-09 ***
X1           0.05700    0.01125   5.067 1.96e-06 ***
X2          -0.18454    0.02941  -6.274 1.01e-08 ***
X3           0.05996    0.05011   1.197    0.234
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5713 on 96 degrees of freedom
Multiple R-squared:  0.4397,    Adjusted R-squared:  0.4222
F-statistic: 25.12 on 3 and 96 DF,  p-value: 4.425e-12
```

$\hat{Y}_i = 3{,}24 + 0{,}05 X_{i1} - 0{,}18 X_{i2} + 0{,}05 X_{i3}$

- A un nivel sig. 5%, el efecto de las horas de estudio es significativo.

$H_0: \beta_j = 0$ vs $H_1: \beta_j \neq 0$; $j = 0, 1, 2, 3$

VIMSE :

```
> confint(modelo, level = 0.95)
                  2.5 %      97.5 %
(Intercept)  2.27682537  4.21719730
X1           0.03467034  0.07932734
X2          -0.24291636 -0.12615438
X3          -0.03949946  0.15942115
```

- A un nivel de conf. del 95%: el efecto promedio de las horas de estudio varía entre 0,034 y 0,079

- A un nivel de conf. 95% : por un cambio unitario en las horas de estudio el cambio promedio del prom. académico está comprendido entre 0,03 y 0,07 manteniendo lo demás constante.

4. Identifique puntos atípicos, de balanceo e influenciales mediante: - $|r_i| > 2$ - $h_{ii} > 2p/n$ - D_i, DFBETAS_j, DFFITS_i

```
> balanceo <- which(hat_values > ((2 * p)/n))
> balanceo
64 74 90 97
64 74 90 97
```

Estos son balanceo

```
> atipicos_estandarizados <- which(abs(estandarizados) > 3)
> atipicos_estudentizados <- which(abs(estudentizados) > 3)
> atipicos_estandarizados
10 25 90
10 25 90
> atipicos_estudentizados
10 25 90
```

```
> DFBetas <- dfbetas(modelo) # Definir
> which(abs(DFBetas) > 2/sqrt(n)) # Verificar DFBETAS
 [1]  16  18  60  90 110 118 125 172 190 210 225 236 243 259
[15] 290 325 336 346 360 390
> DFFITS <- dffits(modelo) # Definir
> which(abs(DFFITS) > (2 * sqrt(p/n))) # Verificar DFFITS
10 25 72 90
10 25 72 90
```

5. Construya el intervalo de confianza para la respuesta media y el intervalo de predicción para el punto:

$$\mathbf{x}_0 = [1, 55, 32, 19]'$$

Verifique previamente que no es extrapolación.

```
> x01 <- c(1, 55, 32, 19)
> # ————————————————
> ifelse(t(x01)%*%solve(t(X)%*%X)%*%x01 < max(Hat_values), "Perten
ce a la region de diseno", "No pertenece")
     [,1]
[1,] "No pertenece"
```

No se puede realizar inferencias