Segundo Taller

miércoles, 11 de diciembre de 2024 1:59 p. m.

$\hat{\sigma}^2 = MSE$

1. Determine el valor de verdad de las siguientes afirmaciones.

- (a) Una suma de cuadrados extra mide la reducción marginal en el SSE cuando una o varias variables predictoras son agregadas al modelo de regresión, dado que las otras predictoras ya fueron agregadas o están en el modelo. Verdadero

n = # datos (información)
p = # parámetros.

- (b) El estadístico T correspondiente al procedimiento de prueba empleado para probar la significancia marginal del parámetro j es:

Falso

$$T_0 = \frac{\hat{\beta}_j - 0}{\sqrt{\sigma^2 c_{jj}}} \sim \text{Normal} \sim N(0,1)$$

Varianza conocida

$$T_0 = \frac{\hat{\beta}_j - 0}{\sqrt{\hat{\sigma}^2 c_{jj}}} \sim t_{n-p}$$

Con una región de rechazo asociada de $R_c = \{|T_0| > t_{\alpha/2,n-p}\}$ y p- valor $P(|t_{n-p}| > |T_0|)$. ✓

- (c) Valores grandes de R^2 implican que la superficie ajustada de respuesta es útil; sin embargo, es menos preferido que R^2_{adj} como medida de bondad de ajuste. ✓

Incorrecta: Falso

$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$; $R^2_{adj} = 1 - \frac{(n-1)MSE}{SST}$.

El R2 ajustado es más preferido que el R2 como medida de bondad de ajuste, dado que en el primero, hay una penalización que se realiza en función del número de covariables en el modelo de regresión.

- (d) El estadístico F correspondiente al procedimiento de prueba empleado para probar la significancia global del modelo de regresión lineal múltiple es:

$$F_0 = \frac{SSR/(k)}{SSE/n-k} \sim f_{k,n-k}$$

Falso

Con una región de rechazo asociada de $R_c = \{F_{calc} > f_{\alpha,k,n-k}\}$ y p- valor $P(f_{k,n-k} > F_{calc})$.

$Y_i = \beta_0 + \beta_1 X_{i1} + \dots + \beta_k X_{ik} + \varepsilon_i$; $\varepsilon_i \overset{iid}{\sim} N(0,\sigma^2)$

p: # parámetros; $p = k+1$
k: # covariables

- (e). Los grados de libertad del cuadrado medio debido a la hipótesis son iguales al rango de la matriz $\mathbf{L}$, asociada la prueba lineal general ($H_0 : \mathbf{L}\beta = 0$ vs $H_1 : \mathbf{L}\beta \neq 0$). Verdadera

- $F_0 = \frac{SSR/k}{SSE/n-p} \sim f_{k,n-p}$
- $F_0 = \frac{SSR/k}{SSE/n-(k+1)} \sim f_{k,n-k-1}$

- $R_c = \{F_0 > f_{\alpha,k,n-p}\}$; $P(f_{k,n-p} > F_0)$

$H_0 : L\beta = 0$ ~ L: m ecuaciones; $m \leq r$
r: Rango (L).

- Rango: Contar el # de filas no-nulas linealmente independientes: $L = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix}$ → 2

- $MSH = SSH/r$

- Parte práctica taller:

1. Determine cuál es el modelo empleado en esta situación, junto con sus supuestos, además, reporte la recta de regresión ajustada.

k = 3 covariables: Strength, Skills, Speed.
p = k+1; p = 4 parámetros ($\beta_0, \beta_1, \beta_2, \beta_3$)
Y = Performance (rendimiento)
n = 500

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 70.80225    2.45665   28.82   <2e-16 ***
Strength     0.88463    0.02848   31.06   <2e-16 ***
Skills       1.87368    0.02073   90.41   <2e-16 ***
Speed        3.04502    0.02789  109.19   <2e-16 ***
---
```

$\hat{Y}_i = 70,80 + 0,88 X_{i1} + 1,87 X_{i2} + 3,04 X_{i3}$
$\hat{Y}_i = \hat{\beta_0} + \hat{\beta_1} X_{i1} + \hat{\beta_2} X_{i2} + \hat{\beta_3} X_{i3}$

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

$Y = X\beta + \varepsilon$; $\varepsilon \sim Nn(0_{nxn}, \sigma^2 I_n)$

2. Determine la significancia de la regresión global. ¿Cree usted que puede realizarse esta prueba empleando otro método? De ser así, pruébelo.

Prueba F: Análisis Varianza (ANOVA).

prueba empleando otro método? De ser así, pruébelo.

Prueba F: Análisis Varianza (ANOVA).

Fuente	gl	Medias	Estadístico
SSR	k	MSR = SSR/k	F_0: MSR/MSE
SSE	n-p	MSE = SSE/n-p	
SST	n-1		

- $F_0 = \frac{SSR/k}{SSE/n-p} \sim f_{k,n-p}$
- $F_0 = \frac{SSR/k}{SSE/n-(k+1)} \sim f_{k,n-k-1}$
- $R_c = \{F_0 > f_{\alpha,k,n-p}\}$; $P(f_{k,n-p} > F_0)$

```
Analysis of Variance Table

Response: Performance
           Df  Sum Sq Mean Sq  F value    Pr(>F)
Strength    1   78308   78308   755.26 < 2.2e-16 ***
Skills      1  817773  817773  7887.27 < 2.2e-16 ***
Speed       1 1236243 1236243 11923.33 < 2.2e-16 ***
Residuals 496   51427     104
```

⇒

```
Anova Table
            Sum Sq  Df Mean Sq F value    Pr(>F)
Regression 2131351   3  710775  6855.3 < 2.2e-16 ***
Residuals    51427 496     104
Total      2183751 499
```

$\alpha = 0{,}05 > P_{val}$

$H_0: \beta_1 = \beta_2 = \beta_3 = 0$
$H_1:$ Algún $\beta_j \neq 0$; $j = 1,2,3$

Con un nivel de significancia del 5%, se puede establecer que, existe evidencia suficiente para rechazar la H0, esto es que, la regresión es significativa.

⇓

$H_0: \beta_1 = 0$
$H_1: \beta_1 \neq 0$

→ Relación regresión

- Hay otras dos maneras: SS extra ; Prueba lineal general

$H_0: \beta_1 = \beta_2 = \beta_3 = 0 \Rightarrow$ Hipótesis cierta
$H_1:$ Algún $\beta_j \neq 0$; $j = 1,2,3$

SS_{extra}:

MF: $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \varepsilon_i$; $\varepsilon_i \overset{iid}{\sim} N(0,\sigma^2)$ (1)

MR: $Y_i = \beta_0 + \varepsilon_i$; $\varepsilon_i \overset{iid}{\sim} N(0,\sigma^2)$ (2).

SS_{extra}:

$SSR(\beta_1,\beta_2,\beta_3 | \beta_0) = SSR(\beta_0,\beta_1,\beta_2,\beta_3) - SSR(\beta_0)$: $SSR(MF) - SSR(MR)$

$SSE = SSE(MR) - SSE(MF) : SSE(\beta_0) - SSE(\beta_0,\beta_1,\beta_2,\beta_3)$

- Grados libertad:

SSR: $gl(SSR(MF)) - gl(SSR(MR)) : 3 - 0 = 3$ ⇝

(# covariables) (# covariables)

- SSE: $gl(SSE(MR)) - gl(SSE(MF)) : (n-1) - (n-4) = n - 1 - n + 4 = 3$

$n-p$ → # parámetros MR ; $n-p$ → # parámetros MF

- $F_0: \frac{SSR(\beta_1,\beta_2,\beta_3|\beta_0)/3}{MSE(\beta_0,\beta_1,\beta_2,\beta_3)} = \frac{[SSR(MF) - SSR(MR)]/3}{MSE(MF)}$.

$= F_0: \frac{SSE(MR) - SSE(MF)/3}{MSE(MF)}$.

```
> #
> F0 <- (SSE_parcial/k)/MSE
> F0
[1] 6855.288
```

$F_0 = 6855{,}28$; $P_{val} \approx 0$

Prueba lineal general:

$H_0: \beta_1 = 0$ (1) $\quad H_1:$ Algún $\beta_j \neq 0$
$\beta_2 = 0$ (2)
$\beta_3 = 0$ (3)

$H_0: \beta_1 = \beta_2 = \beta_3 = 0 \Rightarrow$ Hipótesis cierta
$H_1:$ Algún $\beta_j \neq 0$; $j = 1,2,3$

$L = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$

$r = 3$

3. Determine la significancia de los parámetros individuales β_j, junto con intervalo de confianza. Brinde una interpretación apropiada.

$H_0: \beta_j = 0$ vs $H_1: \beta_j \neq 0$ $\quad j = 0,1,2,3$: $T_0 = \dfrac{\hat{\beta}_j - 0}{\sqrt{\hat{\sigma}^2 c_{jj}}} \sim t_{n-p}$ $\quad \alpha = 0.05$

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 70.80225    2.45665   28.82  <2e-16 ***
Strength     0.88463    0.02848   31.06  <2e-16 ***
Skills       1.87368    0.02073   90.41  <2e-16 ***
Speed        3.04502    0.02789  109.19  <2e-16 ***
---
```

2α

A un nivel de significancia del 5%, se puede determinar que el parámetro B1 es diferente de cero dado que las demás covariables están presentes en el modelo. -> Efecto de la covariable fuerza es significativamente distinto de cero, dado que las demás covariables están presentes en el modelo. (CONCLUIR DE FORMA MARGINAL - LOS EFECTOS SON PARCIALES).

```
> confint(modelo)
                2.5 %     97.5 %
(Intercept) 65.9755369 75.6289713
Strength     0.8286681  0.9405966
Skills       1.8329626  1.9144035
Speed        2.9902333  3.0998132
```

$\hat{\beta}_i \pm t_{\alpha/2, n-p} \sqrt{\hat{\sigma}^2 c_{ii}}$

$\hat{\beta}_i \pm t_{\alpha/2, n-p} \; s.e(\hat{\beta}_i)$

Con un nivel de confianza del 95%, se puede determinar que por un cambio unitario de la covariable fuerza, el cambio promedio de la variables respuesta (PERFORMANCE) -el cambio promedio del rendimiento- está comprendido entre 0.82 y 0.94 unidades, dado que las demás covariables están en el modelo -manteniendo el efecto de las demás covariables fijo-.

4. Determine si el efecto de la primera covariable es el mismo que el efecto de la tercera covariable; al mismo tiempo, verifique si el correspondiente efecto de la primera covariable es el mismo que el de la segunda covariable. Plantee una prueba de hipótesis para ello y realice el procedimiento adecuado. **Reporte el modelo completo y el modelo reducido.**

$H_0: \beta_1 = \beta_3 \Rightarrow H_0: \beta_1 - \beta_3 = 0$ (1) vs $H_1: \beta_1 \neq \beta_3$

$\beta_1 = \beta_2 \quad\quad \beta_1 - \beta_2 = 0$ (2) $\quad\quad \beta_1 \neq \beta_2$

$H_0: L\beta = 0$ vs $H_1: L\beta \neq 0$: $F = \dfrac{SSH/r}{MSE} \sim f_{r, n-p}$

$r = Rango(L)$: # filas L.I distintas de cero

$R_c: \{ F_{calc} > F_{\alpha, r, n-p} \}$ $\quad P(f_{r,n-p} > F_{calc}) < \alpha$

MF

$$L = \begin{bmatrix} 0 & 1 & 0 & -1 \\ 0 & 1 & -1 & 0 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

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

Modelo reducido: MR

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

$Y_i = \beta_0 + \beta_1 (X_{i1} + X_{i2} + X_{i3}) + \varepsilon_i$

$Y_i = \beta_0 + \beta_1 X_{i123} + \varepsilon_i$

SSH: $SSR(MF) - SSR(MR)$

$SSE(MR) - SSE(MF)$

$SSR(\beta_0, \beta_1, \beta_2, \beta_3) - SSR(\beta_0, \beta_1)$

$SSE(\beta_0, \beta_1) - SSE(\beta_0, \beta_1, \beta_2, \beta_3)$ $\Big\}$ = SSH

$F = \dfrac{SSH/r}{MSE} : \dfrac{[SSE(\beta_0, \beta_1) - SSE(\beta_0, \beta_1, \beta_2, \beta_3)]/2}{MSE(\beta_0, \beta_1, \beta_2, \beta_3)} = \dfrac{[366319 - 51427]/2}{51427/(500-4)} = 1518{,}5$ $\quad P_v < \alpha$

```
Analysis of Variance Table

Model 1: Y ~ X1 + X2 + X3
Model 2: Y ~ X123
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    496  51427
2    498 366319 -2   -314893 1518.5 < 2.2e-16 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

SSE

Se puede determinar, a un nivel de significancia del 5%, que el efecto de X1 es significativamente distinto del efecto de X3. Al mismo tiempo, el efecto de X1 es significativamente distinto del efecto de X2.