

Cuarto taller

sábado, 25 de octubre de 2025 12:01 p. m.

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

Verdadera

- (a) Bajo el modelo de regresión lineal múltiple $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_k X_{ik} + \varepsilon_i$; $\varepsilon_i \stackrel{iid}{\sim} N(0, \sigma^2)$, la comparación de los efectos parciales de las variables debe realizarse a través del escalamiento normal unitario de las covariables.
- (b) La multicolinealidad refiere a la dependencia lineal casi perfecta entre covariables, afectando la matriz $X'X$. Esta puede destacarse si la correlación entre un par de variables $X_i \neq X_j$ es pequeña.
- (c) La multicolinealidad puede causar la inflación de las varianzas de los estimadores, además de estimadores $\hat{\beta}_j$ muy grandes en términos absolutos y valores de los coeficientes estimados con signo contrario a lo esperado.
- (d) Una forma en la que se manifiesta la multicolinealidad grave es cuando el modelo de regresión ajustado es significativo (globalmente), pero los parámetros individuales no lo son.

(b). Falsa: $X'X$ Singulair/invertible

$(X'X)^{-1}$: Estimación $\hat{\beta} = (X'X)^{-1}X'Y$ (OLS).

(c): La correlación: Fuerte y dirección: Criterio no válido

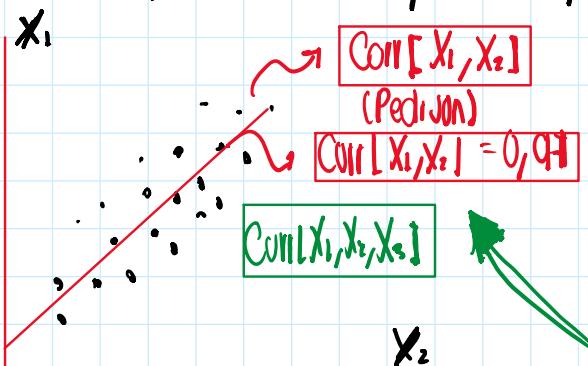
(c). Verdadera: $VIF_j = \frac{1}{1 - R_j^2}$

Cuantificando el grado de asociación lineal

- $VIF_j \leq 5$ no hay mult.
- $5 < VIF_j < 10$ moderada
- $VIF_j \geq 10$ severa

$$Y_i = \beta_0 + \beta_1 X_{i1} + \dots + \beta_j X_{ij} + \dots + \beta_k X_{ik} + \varepsilon_i$$

$$X_1 \sim \dots$$



R_j^2 : Grado de variabilidad que es explicada por el grado de asociación lineal entre la variable X_j y las demás covariables en general $X_i \sim \dots$

Multivariado

(d). Verdadera: (1). Pasa la prueba F; (2). No pasa la prueba T.

• Verificación en vía PH

- (e) Dado que el estadístico C_p es una medida del sesgo del modelo, se prefiere el estadístico más bajo, puesto que a mayor sesgo mayor C_p .
- (f) La suma de cuadrados de los errores de predicción $e_{(i)} = Y_i - \hat{Y}_{(i)}$ mide qué tan bien los valores ajustados por un submodelo predicen las respuestas observadas. Mejor se considerará el modelo entre menor sea esta métrica.

(f). Falsa: $PRESS_{(i)} = \sum_{i=1}^n e_{(i)}^2$ menor

Parte práctica:

1. Escriba el modelo de regresión lineal múltiple, junto con sus supuestos. Deduzca a partir del modelo ajustado si podría haber problemas de multicolinealidad.

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_k X_{ik} + \varepsilon_i$$

$\varepsilon_i \stackrel{iid}{\sim} N(0, \sigma^2)$

• Comparar efectos parciales: estandarización unitaria

$$Y_i^* = Y_i - \bar{Y}$$

$$\sqrt{\frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y})^2}$$

$$X_i^* = X_i - \bar{X}$$

$$\sqrt{\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2}$$

$$\frac{Y_i - \bar{Y}}{\sigma}$$

Dijo que un poco de las notas de clase

(c). Verdadera: $VIF_j = \frac{1}{1 - R_j^2}$; R_j^2 : Variabilidad que es explicada por el grado de asociación lineal entre la variable X_j y las demás covariables en general $X_i \sim \dots$

$$X_i = \beta_0 + \beta_1 X_{i1} + \dots + \beta_j X_{ij} + \dots + \beta_k X_{ik} + \varepsilon_i$$

$$\text{Corr}[X_1, X_2]$$

(Pearson)

$$\text{Corr}[X_1, X_2] = 0.95$$

$$\text{Corr}[X_1, X_2, X_3]$$

R_j^2 : Grado de variabilidad que es explicada por la relación lineal entre $Y \sim \dots$

(e). Verdadero.

$$Q = \frac{SSE}{MSE} - (n - 2p)$$

$$|Q - p| \rightarrow 0$$

$$E(C_p) = p$$

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

$\varepsilon_i \stackrel{iid}{\sim} N(0, \sigma^2)$

1. Escriba el modelo de regresión lineal múltiple, junto con sus supuestos. Deduzca a partir del modelo ajustado si podría haber problemas de multicolinealidad.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 70.80225  2.45665 28.82 <2e-16 ***
X1          0.88463  0.02848 31.06 <2e-16 ***
X2          1.87368  0.02073 90.41 <2e-16 ***
X3          3.04502  0.02789 109.19 <2e-16 ***

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.18 on 496 degrees of freedom
Multiple R-squared:  0.9765, Adjusted R-squared:  0.9763 
F-statistic: 6855 on 3 and 496 DF,  p-value: < 2.2e-16
```

```
> cor(datos[, -1]) # Sin la inclusión de la respuesta
      Strength Skills Speed
Strength 1.00000000 -0.006905304 -0.02712214
Skills   -0.006905304 1.00000000 -0.01449966
Speed    -0.027122140 -0.014499658 1.00000000
```

| | X_1 | X_2 | X_3 |
|-------|-------------------------|-------------------------|-------------------------|
| X_1 | $\text{Corr}(X_1, X_1)$ | $\text{Corr}(X_1, X_2)$ | $\text{Corr}(X_1, X_3)$ |
| X_2 | $\text{Corr}(X_2, X_1)$ | $\text{Corr}(X_2, X_2)$ | $\text{Corr}(X_2, X_3)$ |
| X_3 | $\text{Corr}(X_3, X_1)$ | $\text{Corr}(X_3, X_2)$ | $\text{Corr}(X_3, X_3)$ |

$$\cdot X_1, X_2, X_3;$$

$$\text{Corr}[X_1, X_2] = \text{Corr}[X_2, X_1]$$

$$\text{Corr}[X_1, X_3] = \text{Corr}[X_3, X_1]$$

$$\text{Corr}[X_2, X_3] = \text{Corr}[X_3, X_2]$$

$$-1 \leq \text{Corr} \leq 1$$

Da una idea

$\text{Corr} > 0, \text{ f. directa}$

$\text{Corr} < 0, \text{ f. inversa}$

2. Realice un análisis de multicolinealidad a través del criterio del factor de inflación de varianza, número de condición, índice de condición y proporción de descomposición de varianza.

- $VIF_j \leq 5$ no hay mult.
- $5 < VIF_j < 10$ moderada
- $VIF_j \geq 10$ severa

- Número condición: $\text{Núm}(x_{\text{máx}})$ número condición

| |
|---|
| $10 > \sqrt{\lambda_j}$ no hay mult. |
| $10 \leq \sqrt{\lambda_j} \leq 31,6$ moderada |
| $\sqrt{\lambda_j} > 31,6$ severa |

índice de condición: Decomposición en valores propios

$$R_i = \frac{\lambda_{\max}}{\lambda_i}$$

$10 > \sqrt{\lambda_j}$ No hay multicoll.

$10 \leq \sqrt{\lambda_j} \leq 31,6$ moderada

$\sqrt{\lambda_j} > 31,6$ severa

$j = 1, \dots, n$

proporción de descomposición varianza $\pi_{ij} > 0,5$

- VIF_j, I condición, # cond: Detectar si hay o no hay multicolinealidad y en qué grado

- π_{ij} : Identificar el conjunto de variables que causan la mult.

```
> car::vif(modelo) # Identificar multicolinealidad
      X1        X2        X3
1.000790 1.000264 1.000952
```

$VIF_j < 5$: No identifica mult

| olsrr::ols_coll_diag(modelo) # Análisis general | | | |
|---|-----------|----------|--|
| Tolerance and Variance Inflation Factor | | | |
| Variables | Tolerance | VIF | |
| 1 X1 | 0.9992111 | 1.000790 | |
| 2 X2 | 0.9997365 | 1.000264 | |
| 3 X3 | 0.9990487 | 1.000952 | |

| Eigenvalue and Condition Index | | | |
|--------------------------------|-------------|-----------------|-----------|
| | Eigenvalue | Condition Index | |
| 1 3.77122408 | | 1.000000 | intercept |
| 2 0.10575817 | | 5.971511 | X1 |
| 3 0.09759469 | | 6.216244 | X2 |
| 4 0.02542306 | | 12.179438 | X3 |
| | | | |
| 1 0.007072986 | 0.005957841 | | |
| 2 0.644487364 | 0.026440677 | | |
| 3 0.068484017 | 0.565467283 | | |
| 4 0.279955633 | 0.402134199 | | |

3. Use el método de todas las regresiones posibles para seleccionar los mejores submodelos en función de los criterios R_p^2 , $R_{adj(p)}^2$ (o bien MSE_p) y C_p . Posteriormente seleccione el mejor modelo.

| TERCER PUNTO | | | | | | |
|--------------|-------|-------|----------|-----------|-----------|--------------------|
| | k | R_sq | adj_R_sq | SSE | Cp | Variables_in_model |
| 1 1 | 0.545 | 0.544 | 0.544 | 994638.7 | 9097.182 | X3 |
| 2 1 | 0.373 | 0.372 | 0.372 | 1369507.8 | 12712.643 | X2 |
| 3 1 | 0.036 | 0.034 | 0.034 | 2105443.3 | 19810.602 | X1 |
| 4 2 | 0.931 | 0.930 | 0.930 | 151433.3 | 966.546 | X2 X3 |
| 5 2 | 0.588 | 0.587 | 0.587 | 898833.1 | 8175.076 | X1 X3 |
| 6 2 | 0.418 | 0.408 | 0.408 | 1287670.0 | 11925.333 | X1 X2 |
| 7 3 | 0.976 | 0.976 | 0.976 | 51426.6 | 4.000 | X1 X2 X3 |

(1). Escoger el mejor modelo para $p=2,3,4$

(2). Selecciona el modelo global. Suponiendo que los supuestos se cumplen

| P | R_p^2 | R_{adj}^2 | SSE | Cp | Ecuación |
|---|-----------------|-----------------|-----------------|-----------------|--|
| 2 | X_3 | X_3 | X_3 | X_3 | $Y_i = \beta_0 + \beta_3 X_3 + \epsilon_i$ |
| 3 | X_2, X_3 | X_2, X_3 | X_2, X_3 | X_2, X_3 | $Y_i = \beta_0 + \beta_2 X_2 + \beta_3 X_3 + \epsilon_i$ |
| 4 | X_1, X_2, X_3 | X_1, X_2, X_3 | X_1, X_2, X_3 | X_1, X_2, X_3 | $Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon_i$ |

$$R^2 = 0,6 \rightarrow R^2 \approx 0,8$$

$R^2 \approx 1$: No penaliza # CW

$$t_{ij} > 0,5$$

$$\sqrt{V_{ii}} = 12,17 > 10$$

moderada

• No se debe a una relación entre variables

- Mejores métricas de rendimiento ajustado
- Complemento supuestos

→ AJUSTE

Para seleccionar modelos:

R_p^2 más alto *
 R_{adj}^2 más alto
 SSE_p más bajo
 C_p más bajo

Sobre punto : No penaliza

• Principio de parsimonia