Cuarto Taller

miércoles, 29 de enero de 2025 1:59 p. m.

(a) Bajo el modelo de regresión lineal múltiple $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_k X_{ik} + \varepsilon_i; \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$, la comparación de los efectos parciales de las variables debe realizarse a través del **escalamiento normal unitario de las covariables**.

Verdadero: $X_{ij}^* = \frac{X_{ij} - \bar{X}}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_{ij} - \bar{X})^2}}$

$Y_i^* = \frac{Y_i - \bar{Y}}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$: $Y_i^* = \beta_1^* X_{i1}^* + \ldots + \beta_k^* X_{ik}^* + e_i; \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$

Las interpretaciones que se realizan con este modelo, son en base a las variables originales.

(b) La multicolinealidad refiere a la dependencia lineal casi perfecta entre covariables, afectando la matriz $X'X$. Esta puede destacartarse si la correlación entre un par de variables $X_i \neq X_j$ es pequeña.

Falsa → No invertible

- NO se puede descartar la multicolinealidad a partir de la correlación entre variables. Acudir a las pruebas formales.

- (c) La multicolinealidad causa 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.

Verdadera

$tr(X'X)^{-1} = \sum_{j=1}^{k+1} \frac{1}{\lambda_j}$ ↑

$VIF_j = \frac{1}{1 - R_j^2}$ ↑ : $\begin{cases} VIF_j < 5 & \text{No mult.} \\ 5 \leq VIF_j < 10 & \text{Mod.} \\ VIF_j > 10 & \text{severa.} \end{cases}$

R^2 : $X_j \sim \ldots$

(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.

Verdadero

(e) Dado que el estadístico C_p es una medida del sesgo del modelo, se prefiere el estadístico más bajo, tal que $|C_p - p|$ es mínima, puesto que a mayor sesgo mayor C_p.

Verdadera

$C_p = \frac{SSE_p}{MSE} - (n - 2p)$ $|C_p - p|$ mínima

- $E(C_p) = p$. ✓

(f) La suma de cuadrados de los errores de predicción $e_{(i)} = Y_i - \hat{Y}_{i(i)}$ mide qué tan bien los valores ajustados por un submodelo predicen las respuestas observadas. **Mejor se considerará el modelo entre mayor** ~~son esta métrica~~.

Falsa → Menor

- $PRESS_p = \sum_{i=1}^{n} e_{(i)}^2$ ↝ Minimice

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.

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

$e_i \overset{iid}{\sim} N(0, \sigma^2)$

```
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
```

(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.

Verdadero

$\begin{cases} H_0: \beta_j = 0 \; ; j = 0, \ldots, 3 \\ H_1: \beta_j \neq 0 \end{cases}$

$P(|t_{n-p}| > |T_{calc}|) < \alpha$

- Los parámetros son significativos

Probar significancia global: $\begin{cases} H_0: \beta_1 = \beta_2 = \beta_3 = 0 \\ H_1: \beta_j \neq 0 \; ; j = 1, 2, 3 \end{cases}$

- El modelo es significativo

$P(f_{n-p} > F_{cal}) < \alpha = 0.05$

H_1: "$\beta_j \neq 0$ $\forall j = 1,2,3$

- El modelo es significativo

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 ✓
- $\sqrt{k}$
- $\sqrt{\kappa_j}$
- π_{ij}

$k = \frac{\lambda_{max}}{\lambda_{min}}$ (X'X)

```
> vif(modelo)
       X1       X2       X3
1.000790 1.000264 1.000952
```

VIF_1 VIF_2 VIF_3

< 5 : Según este criterio, no hay mult.

$\sqrt{\kappa_j} < 10$ no hay mult.
$10 \leq \sqrt{\kappa_j} < 31$ moderada
$\sqrt{\kappa_j} \geq 31$ severa

Un par X_i, X_j tengan Un $\pi_{ij} > 0,5$

```
Eigenvalue and Condition Index
  Eigenvalue Condition Index    intercept          X1
1 3.77122408        1.000000 0.002700794 0.006560913
2 0.10575817        5.971511 0.00028 4626 0.323451648
3 0.09759469        6.216244 0.00045 369 0.337753030
4 0.02542306       12.179438 0.996846311 0.332234409
          X2          X3
1 0.007072986 0.005957841
2 0.644487364 0.026440677
3 0.068484017 0.565467283
4 0.279955633 0.402134199
```

Valores propios X'X

π_{ij}

$\sqrt{\kappa_j}$

0,5

Indice condición: Moderada

La columna del índice de condición, me va a decir a mí, qué tan grave es la multicolinealidad: No hay, si es moderada o es severa.

- El factor pi, es el factor que me indica qué relación de variables induce la multicolinealidad.

En conclusión, estos criterios no detectan multicolinealidad.

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

Observación: Es idóneo recordar que los criterios R^2_p, MSE_p permiten escoger modelos que ajusten bien a los datos, mientras que, por su parte, C_p de Mallows permite escoger el mejor modelo para predecir.

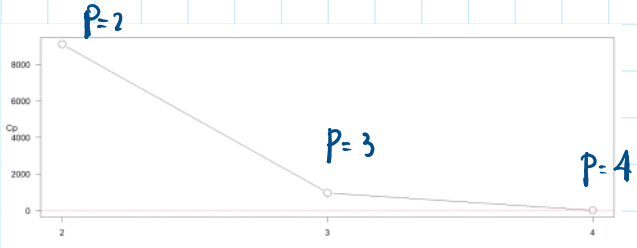
- Modelos seleccionados:

p	R^2	R^2_{adj}	SSE	C_p
2	X_3	X_3	X_3	X_3 ✓
3	X_1X_3	X_1X_3	X_2X_3	X_2X_3 ✓
4	$X_1X_2X_3$	$X_1X_2X_3$	$X_1X_2X_3$	$X_1X_2X_3$ ✓

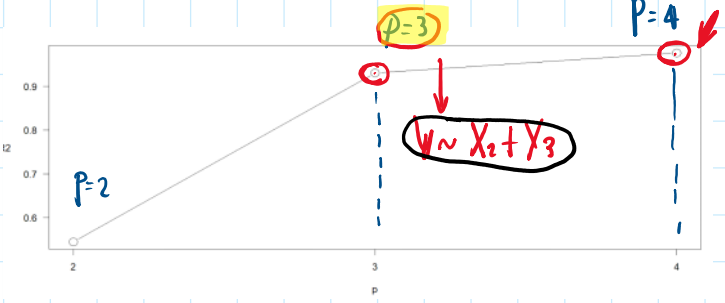
```
> myAllRegTable(modelo) # Función curso
  k  R_sq adj_R_sq      SSE      Cp Variables_in_model
1 1 0.545    0.544 994638.7 9097.102               X3
2 1 0.373    0.372 1369507.8 12712.643              X2
3 1 0.036    0.034 2105443.3 19810.602              X1
4 2 0.931    0.930  151433.3  966.546           X2 X3
5 2 0.588    0.587  898833.1 8175.076            X1 X3
6 2 0.410    0.408 1287670.0 11925.333            X1 X2
7 3 0.976    0.976   51426.6    4.000         X1 X2 X3
```

Mayor — Mayor — Menor — Menor

No tiene penalización.




```
Models are Indexed in rows
 k p        Cp Variables.in.model
1 2 9097.1024               X3
2 3  966.5458            X2 X3
3 4    4.0000         X1 X2 X3
```



- Sin tener en cuenta otros chequeos como el de normalidad, etc. (Otros supuestos)

Según R^2_{adj}

```
> myAdj_R2_criterion(modelo) # Gráfico R2 adj
Models are Indexed in rows
 k p     adjR2 Variables.in.model
1 2 0.5436128               X3
2 3 0.9303754            X2 X3
3 4 0.9763079         X1 X2 X3
```

```
> myR2_criterion(modelo) # Gráfica R2
Models are Indexed in rows
 k p        R2 Variables.in.model
1 2 0.5445274               X3
2 3 0.9306545            X2 X3
```

- Tener en cuenta el principio de parsimonia.

```
Models are Indexed in rows
 k p        R2 Variables.in.model
1 2 0.5445274                 X3
2 3 0.9306545              X2 X3
3 4 0.9764503           X1 X2 X3
```

el principio de parsimonia.

```
3 4 0.9763079           X1 X2 X3
```

$Y \sim X_1 + X_3$

Aclaración:

```
> olsrr::ols_step_all_possible(modelo)
  Index N Predictors    R-Square Adj. R-Square Mallow's Cp
3     1 1         X3 0.54452741    0.54361280  0.54084807
2     2 1         X2 0.37286451    0.37160520  0.36773621
1     3 1         X1 0.03585927    0.03392325  0.02771424
6     4 2      X2 X3 0.93065449    0.93037544  0.92979240
5     5 2      X1 X3 0.58839944    0.58674310  0.58334347
4     6 2      X1 X2 0.41034029    0.40796741  0.40308268
7     7 3   X1 X2 X3 0.97645033    0.97630789  0.97607344
```

$|C_p - p|$