Cuarto taller

jueves, 8 de mayo de 2025 1:58 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**.

Verdadera

$Y_i^* = \beta_1 X_{i1}^* + \beta_2 X_{i2}^* + \ldots + \beta_k X_{ik}^* + \varepsilon_i$
$\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$

- Comparar efectos parciales (diferentes unidades de medida) : ⇒ Estandarización unitaria
- No se estandariza el intercepto (β_0)

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

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

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

Falsa

- (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.

Verdadera

- (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

- La matrix X^TX : Singular / invertible : $(X^TX)^{-1}$
- X^TX → Estimación $\hat{\beta}$: $(X^TX)^{-1}X^TY$ (OLS)
  - Dependencia lineal
- Las correlación : No es criterio válido; diagnóstica

$VIF_j = \dfrac{1}{1 - R_j^2}$; R_j^2 : Variabilidad que es explicada por el grado de asociación lineal entre la covariable X_j y las demás covariables : $X_j \sim \ldots$ $X_j = \beta_0 + \beta_1 X_{j1} + \ldots + \beta_{j-1} X_{j-1} \ldots + \beta_k + \varepsilon$

- Cuantificar el grado de asociación lineal + Minimizar el VIF_j
- $VIF_j \leq 5$ no hay multi.
  $5 < VIF_j < 10$ moderada
  $VIF_j \geq 10$ severa

: Asociados con los eig. valores propios: $tr((X^TX^{-1}) = \sum_{j=1}^{k+1} \dfrac{1}{\lambda_j}$

Multicol. grave

Al realizar prueba de significancia global → significativa: Regresión signif.
prueba de sign. ind → No sig. para algún parámetro

- (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.

Verdadera

Minimizar C_p :

$C_p = \dfrac{SSE_p}{MSE} - (n - 2p)$

- (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 sea esta métrica.**

Falsa

menor

$PRESS_p = \sum_{i=1}^{n} e_{(i)}^2 = Y_i - \hat{Y}_{i(i)}$

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

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

```
> # Realizar un análisis descriptivo
> cor(datos[, -1])
           Strength      Skills       Speed
Strength 1.000000000 -0.006905304 -0.02712214
Skills   -0.006905304 1.000000000 -0.01449966
Speed    -0.027122140 -0.014499658 1.00000000
```

- Parece que no hay multicolinealidad

Todos los efectos parciales son significativos

```
Skills    -0.006905304  1.000000000 -0.01449966
Speed     -0.027122140 -0.014499658  1.00000000
```

Todos los efectos parciales son significativos

```
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
```

- La significancia global se cumple

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

Índice de condición: Descomposición en valores propios

- $K_j = \frac{\lambda_{max}}{\lambda_j}$: $10 < \sqrt{K_j}$ No hay mult.
  $10 \leq \sqrt{K_j} < 31,6$ moderada
  $\sqrt{K_j} \geq 31,6$ severa

- Número condición: Máximo índice condición

$K = \frac{\lambda_{max}}{\lambda_{min}}$: $10 < \sqrt{K}$ No hay mult.
$10 \leq \sqrt{K} < 31,6$ moderada
$\sqrt{K} \geq 31,6$ severa

- Proporción de descomposición de varianza: $\pi_{ij} > 0,5$

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

< 5: No hay mult.

```
Eigenvalue and Condition Index
   Eigenvalue Condition Index   Intercept          X1
1 3.77122408        1.000000 0.0023700794 0.006560913
2 0.10575817        5.971511 0.0002384626 0.323451648
3 0.09759469        6.216244 0.0004951339 0.337753030
4 0.02542306       12.179438 0.9968463217 0.332234409
            X2          X3
1 0.007072986 0.005957841
2 0.644487364 0.026440677
3 0.068484017 0.565467283
4 0.279955633 0.402134199
```

Multicolinealidad moderada

Se requiere que al menos un par > 0,5

- Conclusión: el modelo no tiene mult. (ningún criterio lo identificó)

VIF, I. condición, # cond: Detectar si hay o no hay multicolinealidad y en qué grado.
Proporción V. (π_{ij}): Identificar el conjunto de variables que causan multicoli.

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.**

Observación: Es idóneo recordar que los criterios R_p^2, 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.

Para seleccionar modelos:

R_p^2: Más alto | C_p: Más bajo
R_{adj}^2: Más alto
MSE_p: Más bajo

```
> 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
```

Método de todas las reg. posibles

(1). Escoger el mejor modelo para $p = 2, \ldots, 4$
(2). Seleccionar el mejor modelo global

- La selección de modelos puede llegar a ser subjetiva:

Supongamos que los supuestos se cumplen

p	R_p^2	R_{adj}^2	SSE	C_p	Ecuación

le cumplen

p	R_p^2	R^2_{adj}	SSE	C_p	Ecuación
2	X_3	X_3	X_3	X_3	$Y_i = \beta_0 + \beta_3 X_3 + \varepsilon_i$
3	$X_2\ X_3$	$X_2\ X_3$	$X_2\ X_3$	$X_2\ X_3$	$Y_i = \beta_0 + \beta_2 X_{i2} + \beta_3 X_{i3} + \varepsilon_i$
4	$X_1\ X_2\ X_3$	$X_1\ X_2\ X_3$	$X_1 X_2 X_3$	$X_1\ X_2\ X_3$	

- El principio de parsimonia: Lo más ideal es escoger el modelo con el menor número de covariables (el modelo menos complejo) (Validación de los supuestos).
- La decisión es subjetiva