ME613 - Análise de Regressão

Parte 8

Benilton S Carvalho e Rafael P Maia - 2S2020

Multicolinearidade

Introdução

Vamos considerar o modelo de regressão linear múltipla:

$$Y_i = \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i, \quad i = 1, \ldots, n, \quad \epsilon_i \ iid \ \sim N(0, \sigma^2)$$

em que X_1 e X_2 são variáveis tranformadas via transformação de correlação (ou seja estão centradas no 0 e limitadas entre -1 e 1).

A matrix de desenho do modelo é data por

$$\mathbf{X} = \begin{pmatrix} X_{11} & X_{12} \\ X_{21} & X_{22} \\ \vdots & \vdots \\ X_{n1} & X_{n2} \end{pmatrix} \ \ e \ \ \mathbf{X}^T\mathbf{X} = \begin{pmatrix} \sum X_{i1}^2 & \sum X_{i1}X_{i2} \\ \sum X_{i1}X_{i2} & \sum X_{i2}^2 \end{pmatrix}$$

Daí temos que

$$\det(\mathbf{X}^T\mathbf{X}) = \sum X_{i1}^2 \sum X_{i2}^2 - (\sum X_{i1}X_{i2})^2$$

Se X_1 e X_2 são não colineares então temos que $\sum X_{i1}X_{i2} = 0$ e a matrix $\mathbf{X}^T\mathbf{X} = diag(\sum X_{i1}^2, \sum X_{i2}^2)$.

Daí temos que

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} = \begin{pmatrix} \frac{\sum X_{i1}Y_i}{\sum X_{i1}^2} \\ \frac{\sum X_{i2}Y_i}{\sum X_{i2}^2} \end{pmatrix}$$

E neste caso obter os estimadoes para β_1 e β_2 na regressão múltipla é equivalente a estimá-los separadamente via modelos de regressão linear simples independentes; e

$SQReg(X_1 \mid X_2) = SQReg(X_1)$ e $SQReg(X_2 \mid X_1) = SQReg(X_2)$

Por outro lado se X_1 e X_2 são perfeitamente correlacionadas então

$$\sum X_{i1}^2 \sum X_{i2}^2 = (\sum X_{i1} X_{i2})^2 \ e$$

$$\det(\mathbf{X}^T \mathbf{X}) = 0$$

Neste caso não é possível obter a matriz $\mathbf{X}^T \mathbf{X}$ e o sistema de equações normais não possue solução única.

Quanto mais próximo de 1 for a correlação entre as variáveis mais próximo de 0 será o $\det(\mathbf{X}^T \mathbf{X}) = 0$

Exemplo

X_1: tamanho da equipe ; X_2: pagamento (dólares) ; e Y: produtividade .

```
dados = read.table('./dados/CH07TA06.txt')
colnames(dados) <- c("X1","X2","Y")
dados
```

```
##   X1 X2  Y
## 1  4  2 42
## 2  4  2 39
## 3  4  3 48
## 4  4  3 51
## 5  6  2 49
## 6  6  2 53
## 7  6  3 61
## 8  6  3 60
```

```
cor(dados$X1, dados$X2)
```

```
## [1] 0
```

Regressão de Y em X_1: $\hat{\beta}_1$.

```
##             Estimate Std. Error  t value   Pr(>|t|)
## (Intercept)   23.500  10.111359 2.324119 0.05911468
## X1             5.375   1.983001 2.710539 0.03508095
```

```
## Analysis of Variance Table
##
## Response: Y
##           Df Sum Sq Mean Sq F value  Pr(>F)
## X1         1 231.12 231.125   7.347 0.03508 *
## Residuals  6 188.75  31.458
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$$\hat{\beta}_1 = 5.375$$

$$SQReg(X_1) = 231.12$$

Regressão de Y em X_2: $\hat{\beta}_2$.

```
##             Estimate Std. Error  t value   Pr(>|t|)
## (Intercept)    27.25  11.607738 2.347572 0.05724814
## X2              9.25   4.552929 2.031659 0.08846031
```

```
## Analysis of Variance Table
##
## Response: Y
##           Df Sum Sq Mean Sq F value  Pr(>F)
## X2         1 171.12 171.125  4.1276 0.08846 .
## Residuals  6 248.75  41.458
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$$\hat{\beta}_2 = 9.25$$

$$SQReg(X_2) = 171.12$$

Regressão de Y em X_1 e X_2: $\hat{\beta}_1^*$ e $\hat{\beta}_2^*$.

```
##             Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)    0.375  4.7404509 0.0791064 0.9400164184
## X1             5.375  0.6637959 8.0973685 0.0004657066
## X2             9.250  1.3275918 6.9675031 0.0009365829
```

$$\hat{\beta}_1^* = 5.375$$

$$\hat{\beta}_2^* = 9.25$$

```
## Analysis of Variance Table
##
## Response: Y
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## X1         1 231.125 231.125  65.567 0.0004657 ***
## X2         1 171.125 171.125  48.546 0.0009366 ***
## Residuals  5  17.625   3.525
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$$SQReg(X_2|X_1) = SQE(X_2) - SQE(X_1, X_2) = 171.12 = SQReg(X_2)$$

```
## Analysis of Variance Table
##
## Response: Y
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## X2         1 171.125 171.125  48.546 0.0009366 ***
## X1         1 231.125 231.125  65.567 0.0004657 ***
## Residuals  5  17.625   3.525
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$$SQReg(X_1|X_2) = SQE(X_1) - SQE(X_1, X_2) = 231.12 = SQReg(X_1)$$

Exemplo 2

```
dados = read.table('./dados/CH07TA08.txt')
colnames(dados) <- c("X1","X2","Y")
dados
```

```
##   X1 X2   Y
## 1  2  6  23
## 2  8  9  83
## 3  6  8  63
## 4 10 10 103
```

```
cor(dados$X1, dados$X2)
```

```
## [1] 1
```

```r
modelo1 <- lm(Y ~ X1 + X2,data=dados)
summary(modelo1)$coef
```

```
## Warning in summary.lm(modelo1): essentially perfect fit: summary may be
## unreliable
```

```
##             Estimate   Std. Error      t value     Pr(>|t|)
## (Intercept)        3 6.064937e-15 4.946465e+14 4.087051e-30
## X1                10 8.492610e-16 1.177494e+16 7.212443e-33
```

que produz valores ajustados perfeitos (resíduo nulo):

```
##   1   2   3   4
##  23  83  63 103
```

$$\hat{Y} = -87 + X_1 + 18X_2$$

$$\hat{Y} = -7 + 9X_1 + 2X_2$$

também fornecem os mesmos valores para $\hat{Y}$.

Problema: X_1 e X_2 são perfeitamente correlacionadas ($X_2 = 5 + 0.5X_1$).

Podemos obter bons valores ajustados/preditos, mas não podemos interpretar os parâmetros do modelo (pois temos infinitas possibilidades).

Efeitos da multicolinearidade

Na prática, dificilmente encontraremos variáveis preditoras que sejam perfeitamente correlacionadas entre si.

No entanto, quando a correlação é alta, temos problemas similares aos vistos no exemplo anterior.

Efeito nos coeficientes de regressão

X_1: tríceps

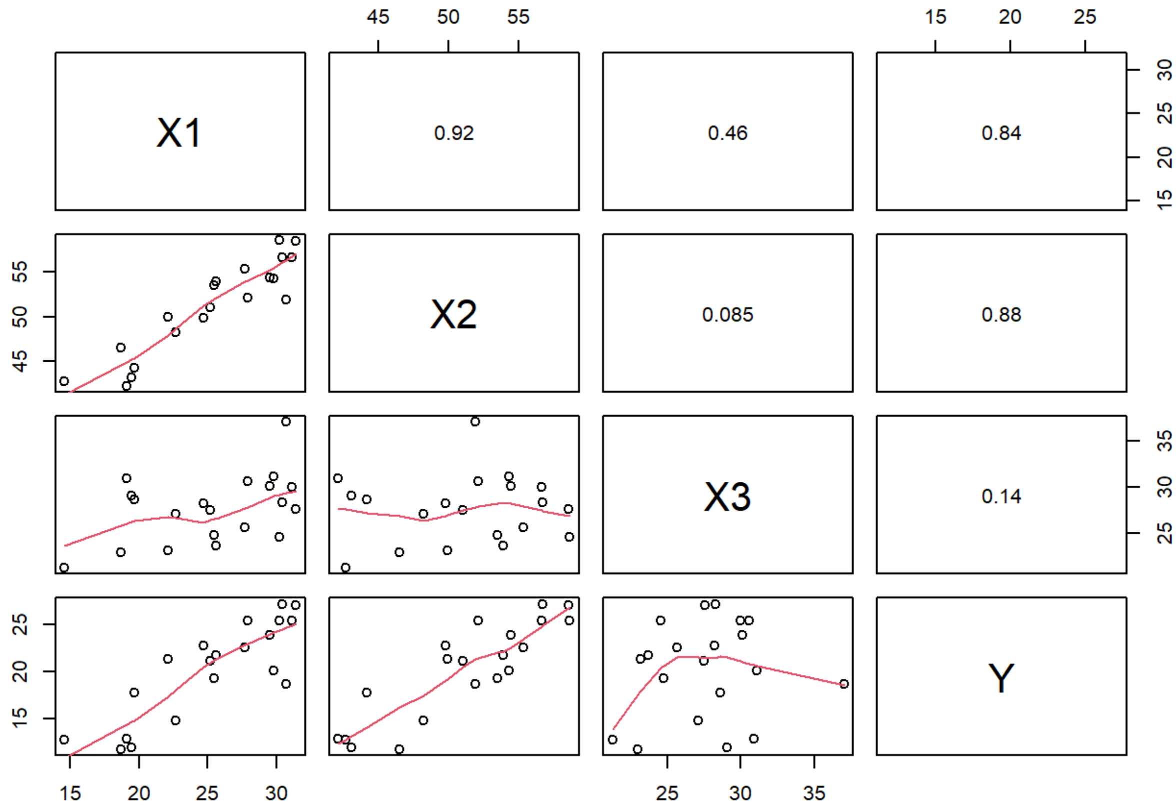
X_2: coxa

X_3: antebraço

Y: gordura corporal

```
##      X1   X2   X3    Y
## 1  19.5 43.1 29.1 11.9
## 2  24.7 49.8 28.2 22.8
## 3  30.7 51.9 37.0 18.7
## 4  29.8 54.3 31.1 20.1
## 5  19.1 42.2 30.9 12.9
## 6  25.6 53.9 23.7 21.7
## 7  31.4 58.5 27.6 27.1
## 8  27.9 52.1 30.6 25.4
## 9  22.1 49.9 23.2 21.3
## 10 25.5 53.5 24.8 19.3
## 11 31.1 56.6 30.0 25.4
## 12 30.4 56.7 28.3 27.2
## 13 18.7 46.5 23.0 11.7
## 14 19.7 44.2 28.6 17.8
```

Exemplo



X1
X2
X3
Y
0.92
0.46
0.84
0.085
0.88
0.14
45 50 55
15 20 25
15 20 25 30
45 50 55
25 30 35
15 20 25
15 20 25 30
25 30 35

Exemplo

Quando as preditoras têm correlação, os efeitos das variáveis são marginais ou parciais.

Variável no modelo	$\hat{\beta}_1$	$\hat{\beta}_2$
X_1	0.857	
X_2		0.857
X_1, X_2	0.222	0.659
X_1, X_2, X_3	4.334	-2.857

As estimativas do efeito de X_1 no modelo variam muito, dependendo das variáveis que são consideradas nos modelos. O mesmo pode ser dito sobre o efeito de X_2.

Efeito na soma extra de quadrados

Quando as variáveis preditoras apresentam correlação, a contribuição marginal de cada variável na redução da soma de quadrados do erro varia, dependendo de quais variáveis já estão no modelo.

Por exemplo: considerando apenas X_1 no modelo

```
## Analysis of Variance Table
##
## Response: Y
##           Df Sum Sq Mean Sq F value    Pr(>F)
## X1         1 352.27  352.27  44.305 3.024e-06 ***
## Residuals 18 143.12    7.95
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$$SQReg(X_1) = 352.27$$

Exemplo

Considerando X_1 e X_2 no modelo (primeiro X_2 e depois X_1):

```
## Analysis of Variance Table
##
## Response: Y
##           Df Sum Sq Mean Sq F value    Pr(>F)
## X2         1 381.97  381.97  59.057 6.281e-07 ***
## X1         1   3.47    3.47   0.537    0.4737
## Residuals 17 109.95    6.47
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$$SQReg(X_1 \mid X_2) = 3.473$$

Exemplo

O modelo de $SQReg(X_1 \mid X_1)$ ser tão pequeno quando comparado a $SQReg(X_1)$ é a alta correlação entre X_1 e X_2 (0.92) e de cada uma delas com a variável resposta (0.84 e 0.88, respectivamente).

Desta forma, quando X_2 já está no modelo, a contribuição marginal de X_1 é pequena na redução da soma de quadrados do erro, pois X_2 contém praticamente a mesma informação que X_1.

Efeito no desvio-padrão da estimativa

Variável no modelo	$\hat{\beta}_1$	$\hat{\beta}_2$
X_1	0.129	
X_2		0.11
X_1, X_2	0.303	0.291
X_1, X_2, X_3	3.016	2.582

Efeito nos valores ajustados e preditos

Variável no modelo	QME
X_1	7.95
X_1, X_2	6.47
X_1, X_2, X_3	6.15

QME diminui conforme variáveis são adicionadas ao modelo (caso usual).

Efeito nos valores ajustados e preditos

A precisão do valor ajustado não é tão afetada quando inserimos ou não uma variável preditora muito correlacionada com outra já no modelo.

Por exemplo, se considerarmos apenas o modelo com X_1, o valor estimado de gorduta corporal para $X_1 = 25$ é:

$$\hat{Y} = 19.934 \qquad \sqrt{\widehat{Var(\hat{Y})}} = 0.632$$

Quando incluímos X_2, altamente correlacionada à X_1, temos:

$$\hat{Y} = 19.356 \qquad \sqrt{\widehat{Var(\hat{Y})}} = 0.624$$

quando $X_1 = 25$ e $X_2 = 50$, por exemplo.

Efeito nos testes simultâneos de β_k

Considere os dados sobre gordura corporal e o modelo com X_1 e X_2 no modelo.

Queremos testar H_0: $\beta_1 = \beta_2 = 0$.

Calculamos:

$$t_1 = \frac{\hat{\beta}_1}{\sqrt{\widehat{Var}(\hat{\beta}_1)}} \qquad t_2 = \frac{\hat{\beta}_2}{\sqrt{\widehat{Var}(\hat{\beta}_2)}}$$

e não rejeitamos H_0 se ambos $|t_1|$ e $|t_2|$ forem menores do que $t_{n-3,\alpha/4} = 2.46$

para $\alpha = 0.05$.

Exemplo

```
##
## Call:
## lm(formula = Y ~ X1 + X2, data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9469 -1.8807  0.1678  1.3367  4.0147
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -19.1742     8.3606  -2.293   0.0348 *
## X1            0.2224     0.3034   0.733   0.4737
## X2            0.6594     0.2912   2.265   0.0369 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.543 on 17 degrees of freedom
## Multiple R-squared:  0.7781, Adjusted R-squared:  0.7519
## F-statistic:  29.8 on 2 and 17 DF,  p-value: 2.774e-06
```

Não rejeitamos H_0.

Exemplo

```
## Analysis of Variance Table
##
## Response: Y
##           Df Sum Sq Mean Sq F value    Pr(>F)
## X1         1 352.27  352.27 54.4661 1.075e-06 ***
## X2         1  33.17   33.17  5.1284    0.0369 *
## Residuals 17 109.95    6.47
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## Analysis of Variance Table
##
## Model 1: Y ~ 1
## Model 2: Y ~ X1 + X2
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     19 495.39
## 2     17 109.95  2    385.44 29.797 2.774e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Exemplo

Se utilizarmos o teste F para $H_0 : \beta_1 = \beta_2 = 0$, temos:

$$F_{obs} = \frac{QMReg}{QME} = \frac{385.44/2}{109.95/17} = 29.8$$

Sob H_0 a estatística do teste tem distribuição $F(2, 17)$, de maneira que o valor crítico para $\alpha = 0.05$ é 3.59.

Encontramos evidências para rejeitar H_0.

Resultado contrário ao obtido com os testes t com correção de Bonferroni.

Agradecimento

- Slides criados por Samara F Kiihl / IMECC / UNICAMP

Leitura

- Applied Linear Statistical Models: Seção 7.6.
- Faraway - Linear Models with R: Seção 7.3.