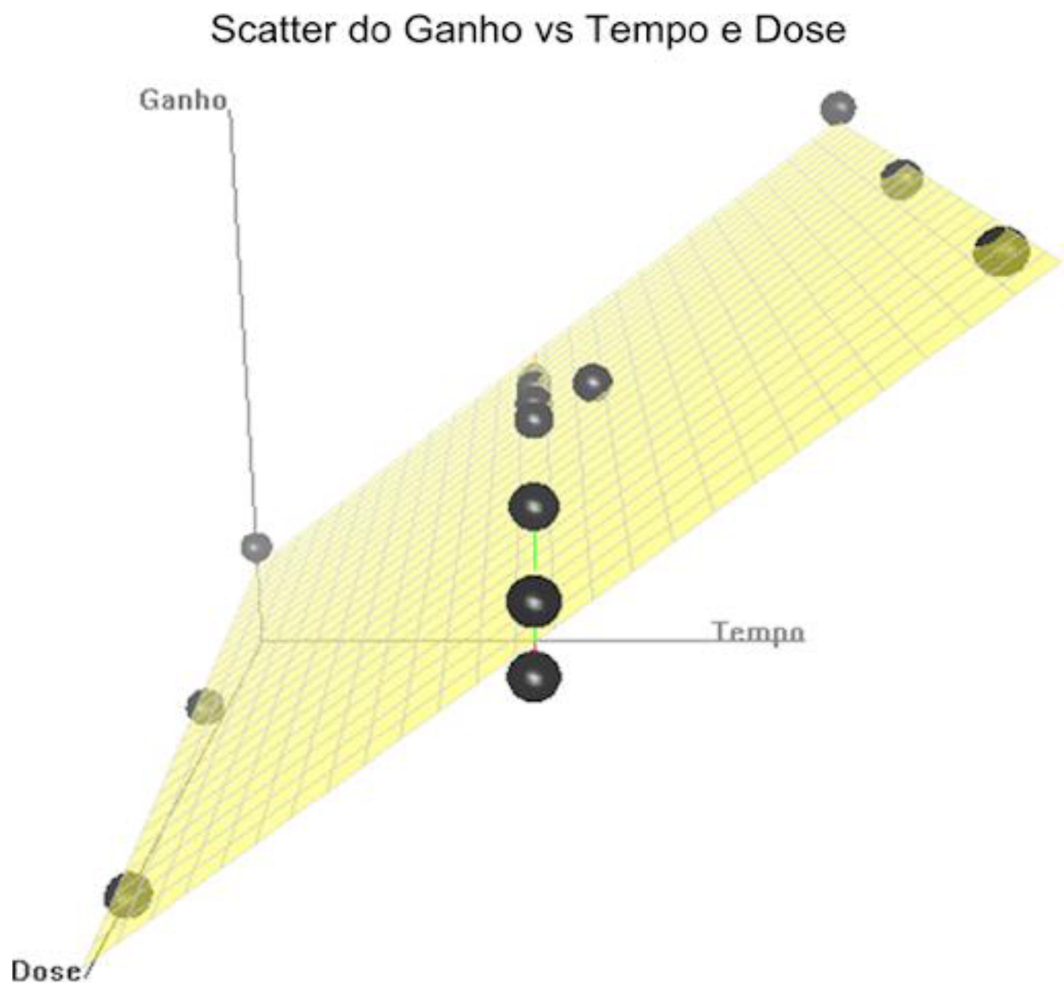
MODELO DE DE REGRESSÃO LINEAR MÚLTIPLA

Apostila Suely Ruiz Giolo



02/12 - Análise de Regressão Múltipla - ARM

09/12 - Análise de Regressão Múltipla - ARM

16/12 - Aula Prática (Exercício)

1. ARM (Interpretação, ANOVA, Coeficientes (Determinação e correlação), Matricial, MMQO)
2. IC
3. TH
4. Diagnósticos
5. Multicolineariedade
6. Diagnóstico de Influência
7. Métodos para tratar multicolineariedade
8. Seleção de variáveis e construção do modelo
9. Extrapolações
10. Validações MRLM
11. Regressão com parte categórica (Variáveis Dammy)
12. Regressão Polinomial
13. Exemplos

13/01 - ARM – Seleção de variáveis

20/01 - Variáveis Dummy

27/01 – 03/02 – 10/02 – 17/02

3- Teste de Hipóteses

3.1 – Teste para significância da regressão (ANOVA)

$$F = QMreg / QMres$$

e que, sob Ho, tem distribuição $F_{p-1;\, n-p}$. Se Ho for rejeitada, haverá evidências de que pelo menos um β_j difere de zero.

3.2 – Teste para os coeficientes individuais da Regressão

$$H_0\colon \beta_j = 0$$
$$H_a\colon \beta_j \neq 0.$$

A estatística de teste usada, em geral, para testar as hipóteses apresentadas é dada por:

$$t^* = \frac{\hat{\beta}_j}{\sqrt{\hat{\sigma}^2 C_{ii}}} = \frac{\hat{\beta}_j}{e.p.(\hat{\beta}_j)} \overset{\text{sob } H_0}{\sim} t_{n-p} \quad (j = 0, 1, \ldots, k \;\text{ e }\; i = j+1),$$

em que C_{ii} é o i-ésimo elemento da diagonal da matriz $\mathbf{(X'X)}^{-1}$ e $\hat{\sigma}^2$ = QMres. Se H_o não for rejeitada, haverá evidências de que a contribuição da regressora X_j para a explicação de Y não é significativa e, desse modo, X_j pode ser excluída do modelo. Caso contário, a regressora deve ser mantida no modelo.

3- Teste de Hipóteses

3.2 – Método da SQextra para testar coeficientes

SQextra

mede o **acréscimo** marginal na SQreg, quando uma ou diversas regressoras são adicionadas ao modelo de regressão ou,

equivalentemente, a **redução** marginal na SQres, quando uma ou mais regressoras são adicionadas ao modelo;

Para isso temos a estatística t*, alternativamente temos a estatística F*;

Coeficiente de determinação parcial;

Coeficiente de correlação parcial.

3- Teste de Hipóteses

3.2.1 – Método da SQextra para testar coeficientes

Para determinar a contribuição da regressora X_j para a SQreg, na presença das demais regressoras X_i ($i \neq j$) no modelo;

Determinar a contribuição de um subconjunto de variáveis regressoras para o modelo.

n = 20, Y = variável resposta e as regressoras contínuas X_1, X_2 e X_3 , os modelos

- Regressão de Y em X_1: $\hat{Y} = -1{,}496 + 0{,}8572x_1$.

F.V.	SQ	g.l.	QM
Reg	352,27	1	352,27
Res	143,12	18	7,95
Total	495,39	19	---

d.p. $(\hat{\beta}_1) = 0{,}1288$

- Regressão de Y em X_2: $\hat{Y} = -23{,}634 + 0{,}8565x_2$.

F.V.	SQ	g.l.	QM
Reg	381,97	1	381,97
Res	113,42	18	6,30
Total	495,39	19	---

d.p. $(\hat{\beta}_2) = 0{,}11$

- Regressão de Y em X_1 e X_2: $\hat{Y} = -19{,}174 + 0{,}2224x_1 + 0{,}6594x_2$.

F.V.	SQ	g.l.	QM
Reg	385,44	2	192,72
Res	109,95	17	6,47
Total	495,39	19	---

d.p. $(\hat{\beta}_1) = 0{,}3034$

d.p. $(\hat{\beta}_2) = 0{,}2912$

- quando X_1 e X_2 estão no modelo tem-se SQres $(X_1,X_2) = 109{,}95$ e
- quando somente X_1 está no modelo tem-se SQres $(X_1) = 143{,}12$

e, ainda,

- quando X_1 e X_2 estão no modelo tem-se SQreg $(X_1,X_2) = 385{,}44$ e
- quando somente X_1 está no modelo tem-se SQreg $(X_1) = 352{,}27$.

- **efeito marginal** de adicionar X_2 em X_1

$$
\begin{aligned}
SQ_E\,(X_2 | X_1) &= SQres\,(X_1) - SQres(X_1, X_2) \\
&= SQreg\,(X_1, X_2) - SQreg\,(X_1) \\
&= 33{,}17.
\end{aligned}
$$

- **efeito marginal** de adicionar X_3 ao modelo quando X_1 e X_2 estão presentes.

$$\boxed{\begin{array}{c} SQ_E\,(X_3 | X_1, X_2) = SQres\,(X_1, X_2) - SQres\,(X_1, X_2, X_3) = 109{,}95 - 98{,}41 = 11{,}54 \\ \textbf{ou} \\ SQ_E\,(X_3 | X_1, X_2) = SQreg\,(X_1, X_2, X_3) - SQreg\,(X_1, X_2) = 396{,}98 - 385{,}44 = 11{,}54. \end{array}}$$

- Regressão de Y em X_1 e X_2: $\hat{Y} = -19{,}174 + 0{,}2224x_1 + 0{,}6594x_2$.

F.V.	SQ	g.l.	QM
Reg	385,44	2	192,72
Res	109,95	17	6,47
Total	495,39	19	---

$d.p.(\hat{\beta}_1) = 0{,}3034$

$d.p.(\hat{\beta}_2) = 0{,}2912$

- Regressão de Y em X_1, X_2 e X_3: $\hat{Y} = 117{,}08 + 4{,}344x_1 - 2{,}857x_2 - 2{,}186x_3$.

F.V.	SQ	g.l.	QM
Reg	396,98	3	132,33
Res	98,41	16	6,15
Total	495,39	19	---

$d.p.(\hat{\beta}_1) = 3{,}016$

$d.p.(\hat{\beta}_2) = 2{,}582$

$d.p.(\hat{\beta}_3) = 1{,}596$

- **efeito marginal** de adicionar X_2 e X_3 ao modelo quando X_1 está presente.

$$SQ_E\,(X_2, X_3 | X_1) = SQres\,(X_1) - SQres\,(X_1, X_2, X_3) = 143{,}12 - 98{,}41 = 44{,}71$$

ou

$$SQ_E\,(X_2, X_3 | X_1) = SQreg\,(X_1, X_2, X_3) - SQreg\,(X_1) = 396{,}98 - 352{,}27 = 44{,}71.$$

- Regressão de Y em X_1: $\hat{Y} = -1{,}496 + 0{,}8572x_1$.

F.V.	SQ	g.l.	QM
Reg	352,27	1	352,27
Res	143,12	18	7,95
Total	495,39	19	---

$d.p.(\hat{\beta}_1) = 0{,}1288$

- Regressão de Y em X_1, X_2 e X_3: $\hat{Y} = 117{,}08 + 4{,}344x_1 - 2{,}857x_2 - 2{,}186x_3$.

F.V.	SQ	g.l.	QM
Reg	396,98	3	132,33
Res	98,41	16	6,15
Total	495,39	19	---

$d.p.(\hat{\beta}_1) = 3{,}016$

$d.p.(\hat{\beta}_2) = 2{,}582$

$d.p.(\hat{\beta}_3) = 1{,}596$

Interesse não está somente em obter tais reduções ou acréscimos, mas saber se a variável (ou as variáveis) Xj deve, ou não, ser incluída no modelo.

Finalidade, já foi visto que a estatística de teste parcial t^* é apropriada.

Alternativamente, pode-se usar a estatística de teste parcial F*,que usa as SQ extra.

Exemplo: Testar se a variável X_3 deve ser adicionada ao modelo contendo X_1 e X_2, o que equivale a testaras hipóteses:

Ho: $\beta 3 = 0$
Ha: $\beta 3 \neq 0$.

Se H_0 não for rejeitada tem-se o **modelo reduzido** $E(Y|\mathbf{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ e, se H_0 for rejeitada tem-se o **modelo completo** $E(Y|\mathbf{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$.

A estatística de teste parcial F* para testar tais hipóteses é expressa por:

$$F^* = \frac{SQ_E(X_3|X_1, X_2) / [(n-3)-(n-4)]}{SQres(X_1, X_2, X_3) / (n-4)} = \frac{SQ_E(X_3|X_1, X_2)/1}{QMres(X_1, X_2, X_3)} \sim F_{1;\, n-4}.$$

Para os dados do exemplo tem-se:

$F^* = 11{,}54 / 6{,}15 = 1{,}88$ (p-valor = 0,189)

$t^* = -2{,}186 / 1{,}596 = -1{,}37$ (p-valor = 0,189).

Obs: Relembre que $F^* = (t^*)^2$.

o teste F*, pode também ser utilizado para testar se um subconjunto de regressoras pode ser retirado do modelo completo.

- Testar se X_2 e X_3 podem ser retiradas do modelo completo, isto é, do modelo contendo X_1, X_2 e X_3. Nesse caso tem-se as hipóteses:

$$H_o: \beta_2 = \beta_3 = 0 \quad \textit{versus} \quad H_a: \ \beta_2 \neq 0 \text{ ou } \beta_3 \neq 0.$$

Se H_o não for rejeitada $\Rightarrow$ tem-se o **modelo reduzido**: $Y = \beta_o + \beta_1 x_1 + \varepsilon$ e,
se H_0 for rejeitada $\Rightarrow$ tem-se o **modelo completo**: $Y = \beta_o + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$.

Em sendo, $F^* = [(33{,}17 + 11{,}54)/2] / [98{,}41/16] = [44{,}71/2] / 6{,}15 = 3{,}63$, para o qual o p-valor associado à distribuição $F_{2;16}$ é igual a 0,05, é possível concluir pela rejeição da hipótese nula e, desse modo, opta-se pelo modelo completo.

3- Teste de Hipóteses

3.3 – Coeficiente de determinação parcial R^2

- **coeficiente de determinação parcial entre Y e X_2 dado X_1 no modelo**

$$\boxed{r^2_{Y2\bullet 1} = \frac{SQ_E(X_2|X_1)}{SQres(X_1)}.}$$

- **coeficiente de determinação parcial entre Y e X_1 dado X_2 e X_3 no modelo**

$$\boxed{r^2_{Y1\bullet 23} = \frac{SQ_E(X_1|X_2,X_3)}{SQres(X_2,X_3)}}$$

(a) $r^2_{Y2\bullet 1} = 33{,}17/143{,}12 = 0{,}2317$ (23,17%),

(b) $r^2_{Y3\bullet 12} = 11{,}54/109{,}95 = 0{,}105$ (10,5%) e,

(c) $r^2_{Y1\bullet 2} = 3{,}47/113{,}42 = 0{,}031$ (3,1%).

quando X2 é adicionada ao modelo contendo X1 a SQres(X1) é reduzida em 23,17%

SQres(X1, X2) é reduzida em 10,5% quando X3 é adicionada ao modelo

E, se o modelo contém X2, adicionar X1 reduz a SQres em 3,1%.

3- Teste de Hipóteses

3.3 – Coeficiente de correlação parcial

$$\boxed{r = +\sqrt{R^2}}$$

(a) $r_{Y2\bullet 1} = (0{,}2317)^{1/2} = -0{,}48$ (sinal negativo pois $\hat{\beta}_2 = -2{,}857$),

(b) $r_{Y3\bullet 12} = (0{,}105)^{1/2} = -0{,}324$ (sinal negativo pois $\hat{\beta}_3 = -2{,}186$) e,

(c) $r_{Y1\bullet 2} = (0{,}031)^{1/2} = 0{,}176$ (sinal positivo pois $\hat{\beta}_1 = 4{,}344$).

4- Diagnóstico nos MRLM

4.1 – Análise dos resíduos

(a) **Resíduos em papel de probabilidade Normal** $(e_i \; x \; F_i)$

- examinar se os erros apresentam distribuição aproximadamente Normal;
- auxiliar na detecção de pontos atípicos.

(b) **Resíduos *versus* valores ajustados** $(e_i \; \mathrm{x} \; \hat{Y}_i)$

- verificar homogeneidade das variâncias dos erros;
- fornecer informações sobre pontos atípicos.

(c) **Resíduos *versus* seqüência de coleta (se conhecida)** $(e_{(i)} \; \mathrm{x} \; \mathrm{i})$

- informações sobre possível correlação entre os erros.

(d) **Resíduos *versus* cada X_j incluída no modelo** $(e_i \; \mathrm{x} \; X_{ij})$

- informações adicionais sobre a adequacidade da função de regressão com respeito a j-ésima variável independente, ou seja, auxilia na detecção de não-linearidade na regressora X_j;
- informações sobre possível variação na magnitude da variância dos erros em relação a variável independente X_j;
- informações sobre dados atípicos.

4- Diagnóstico nos MRLM

4.1 – Análise dos resíduos

(e) **Resíduos parciais *versus* X_{ij} para cada X_j no modelo** $(e_{ij}^* \text{ x } X_{ij})$

- revelar mais precisamente a relação entre os resíduos e cada regressora X_j. O i-ésimo resíduo parcial para a regressora X_j é definido por:

$$e_{ij}^* = e_i + \hat{\beta}_j x_{ij} \qquad (i = 1, \ldots, n)$$

$$e_{ij}^* = (Y_i - \hat{Y}_i) + \hat{\beta}_j x_{ij} \qquad (i = 1, \ldots, n).$$

Permite avaliar falhas de linearidade, presença de *outliers* e heterogeneidade de variâncias.

Se, por exemplo, a relação entre Y e X_j não for linear, o gráfico dos resíduos parciais indicará mais precisamente do que o gráfico *ei versus* X_j como transformar os dados para obter a linearidade.

4- Diagnóstico nos MRLM

4.1 – Análise dos resíduos

(f) **Resíduos *versus* X_k omitidas do modelo**

- ajuda a revelar a dependência da resposta Y com uma ou mais das regressoras não presentes no modelo. Qualquer estrutura (padrão sistemático), que não o aleatório, indicarão que a inclusão daquela variável pode melhorar o modelo.

(g) **Resíduos *versus* interações não incluídas no modelo**

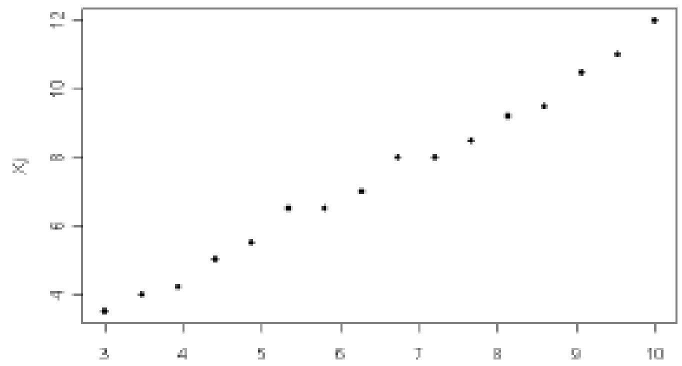
- úteis para examinar se alguma, algumas ou todas as interações são requeridas no modelo. Um padrão sistemático nestes gráficos, que não o aleatório, sugere que o efeito da interação pode estar presente.

(h) **Gráfico da regressora X_i *versus* regressora X_j ($i \neq j$)**

- útil para estudar a relação entre as variáveis regressoras e a disposição dos dados no espaço $\mathbf{X}$;
- encontrar pontos atípicos.

4- Diagnóstico nos MRLM

4.1 – Análise dos resíduos



As regressoras X1 e X2 são altamente correlacionadas consequentemente, pode não ser necessário incluir ambas no modelo.

Quando duas ou mais variáveis regressoras forem altamente corelacionadas,
multicolinearidade está presente nos dados.

A presença de multicolinearidade pode afetar seriamente o ajuste por MQO e, em algumas situações, produzir modelos quase inúteis.

MULTICOLENERIEDADE

$$\mathbf{r}_{XX} = \begin{bmatrix} 1 & r_{12} & .. & r_{ik} \\ r_{21} & 1 & & r_{2k} \\ .. & .. & .. & .. \\ r_{k1} & r_{k2} & .. & 1 \end{bmatrix}$$

A matriz **r**XX é simétrica

Se rij for próximo de zero, então Xi e Xj não são altamente correlacionadas

Se rji for próximo de |1|, então Xi e Xj são altamente correlacionadas.

4- Diagnóstico nos MRLM

4.2 – Propriedades dos resíduos

Foi visto anteriormente que $\mathbf{e} = (\mathbf{I} - \mathbf{H})\mathbf{Y}$, então segue que:

$$\begin{aligned}\mathbf{e} &= (\mathbf{I} - \mathbf{H})(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) = \mathbf{X}\boldsymbol{\beta} - \mathbf{H}\mathbf{X}\boldsymbol{\beta} + (\mathbf{I} - \mathbf{H})\boldsymbol{\varepsilon} \\ &= \mathbf{X}\boldsymbol{\beta} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} + (\mathbf{I} - \mathbf{H})\boldsymbol{\varepsilon} = (\mathbf{I} - \mathbf{H})\boldsymbol{\varepsilon}.\end{aligned}$$

Logo, $\quad \mathbf{E}(\mathbf{e}) = \mathbf{0}$

e, $\quad \mathbf{V}(\mathbf{e}) = \mathbf{V}[(\mathbf{I} - \mathbf{H})\boldsymbol{\varepsilon}] = (\mathbf{I} - \mathbf{H})\mathbf{V}(\boldsymbol{\varepsilon})(\mathbf{I} - \mathbf{H})'$

$$= (\mathbf{I} - \mathbf{H})\sigma^2\mathbf{I}(\mathbf{I} - \mathbf{H})' = \sigma^2(\mathbf{I} - \mathbf{H})$$

pois, $(\mathbf{I} - \mathbf{H})$ é simétrica $((\mathbf{I} - \mathbf{H}) = (\mathbf{I} - \mathbf{H})')$ e idempotente $((\mathbf{I} - \mathbf{H})(\mathbf{I} - \mathbf{H})) = (\mathbf{I} - \mathbf{H})$.

$$\boxed{\begin{array}{ll} e_i \sim \mathrm{N}(0, \sigma^2(1 - \mathrm{h}_{ii})) & i = 1, \ldots, n \\ \mathrm{Cov}(e_i, e_j) = -\sigma^2(\mathrm{h}_{ij}) & i, j = 1, \ldots, n \quad (i \neq j). \end{array}}$$

4- Diagnóstico nos MRLM

4.2 – Propriedades dos resíduos

1) Resíduos *standardized*	$d_i = \dfrac{e_i}{\sqrt{QMres}}$
2) Resíduos *studentized*	$r_i = \dfrac{e_i}{\sqrt{QMres(1-h_{ii})}}$
3) Resíduos PRESS	$e_{(i)} = \dfrac{e_i}{1-h_{ii}}$
4) Resíduos *studendized* externamente (R-Student)	$t_i = \dfrac{e_i}{\sqrt{S^2_{(i)}\ (1-h_{ii})}}$ sendo $S^2_{(i)} = \dfrac{(n-p-1)QMres - e_i^2(1-h_{ii})}{n-p}$

h_{ii} corresponde ao *i*-ésimo componente da diagonal da matriz $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$.

- Pontos com grande resíduo e grande h_{ii} são observações, possivelmente, altamente influentes no ajuste por MQO.
- Resíduos associados a pontos os quais h_{ii} é grande terão resíduos PRESS grandes. Esses pontos geralmente serão altamente influentes.

5- Multicolineariedade

5.1 Fatores de inflação da variância (VIF)

O VIF para o j-ésimo coeficiente de regressão pode ser escrito por:

$$\boxed{\mathrm{VIF}_j = \frac{1}{1 - R^2_j}},$$

em que R^2j é o coeficiente de determinação múltiplo obtido da regressão de Xj com as demais variáveis regressoras.

Se Xj for quase linearmente dependente com alguma das outras regressoras, então R^2j será próximo de 1 e VIFj será grande.

VIF maiores que 10 implicam multicolineridade severa.

5- Multicolineariedade

5.2 Análise dos autovalores da matriz $\mathbf{r}_{XX}$

As raízes características, ou autovalores de $\mathbf{r}_{XX}$, denotadas por $\lambda_1, \lambda_2, \ldots, \lambda_k$, podem ser usados para medir a extensão da multicolinearidade nos dados. Se existirem uma, ou mais, dependência linear nos dados, então uma, ou mais, das raízes características serão pequenas. Auto valores de $\mathbf{r}_{XX}$ são as raízes características da equação $|\,\mathbf{r}_{XX} - \lambda\mathbf{I}\,| = 0$.

Alguns analistas preferem, no entanto, examinar o número de condição da matriz $\mathbf{r}_{XX}$ dado por:

$$\boxed{k = \frac{\lambda_{max}}{\lambda_{min}}.}$$

Geralmente, se $k < 100 \Rightarrow$ não existem sérios problemas de multicolinearidade,

se $100 < k < 1000 \Rightarrow$ moderada a forte multicolinearidade e,

se $k > 1000 \Rightarrow$ severa multicolinearidade.

Os índices de condição da matriz $\mathbf{r}_{XX}$ são dados por: $k_j = \frac{\lambda_{max}}{\lambda_j}.$

5- Multicolineariedade

Exemplo: Suponha Y = variável resposta e X_1, ..., X_9 as regressoras, de modo que os autovalores obtidos sejam:

$\lambda_1 = 4{,}2048$	$\lambda_4 = 1{,}0413$	$\lambda_7 = 0{,}0136$
$\lambda_2 = 2{,}1626$	$\lambda_5 = 0{,}3845$	$\lambda_8 = 0{,}0051$
$\lambda_3 = 1{,}1384$	$\lambda_6 = 0{,}0495$	$\lambda_9 = 0{,}0001$.

Assim, $k = 42048$, o que implica em severa multicolinearidade. Ainda,

$k_1 = 1{,}0$	$k_4 = 4{,}04$	$k_7 = 309{,}18$
$k_2 = 1{,}94$	$k_5 = 10{,}94$	$k_8 = 824{,}47$
$k_3 = 3{,}69$	$k_6 = 84{,}96$	$k_9 = 42048$,

e como k_7 e $k_8 > 100$ e $k_9 > 1000$, há indícios de multicolinearidade envolvendo as variáveis X_7, X_8 e X_9.

5.3 Determinante da matriz $\mathbf{r}_{xx}$

O determinante da matriz $\mathbf{r}_{XX}$ pode ser usado como um indicador de existência de multicolineridade. Os valores possíveis deste determinante são $0 \leq \det(\mathbf{r}_{XX}) \leq 1$. Se $\det(\mathbf{r}_{XX}) = 1$, as regressoras são ortogonais, enquanto $\det(\mathbf{r}_{XX}) = 0$ implica em dependência linear exata entre as regressoras. O grau de multicolinearidade torna-se mais severo quando o determinante aproxima-se de zero.

6- Diagnósticos de Influência

6.1 Pontos de Alavancagem

A disposição dos pontos no espaço **X** é importante para a determinação das propriedades do modelo. Em particular, observações potencialmente remotas têm desproporcional alavancagem nos parâmetros estimados, bem como nos valores preditos e nas usuais estatísticas sumárias. Para localizar esses pontos remotos no espaço **X**, Hoaglin e Welsh (1978) sugeriram o uso da matriz *chapéu*, a qual é obtida por $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. De acordo com esses autores, a inspeção dos elementos da matriz **H** pode revelar pontos que são potencialmente influentes em virtude de sua localização no espaço **X**. Atenção é usualmente centrada nos elementos da diagonal da matriz **H**, ou seja, nos h_{ii}. Como,

$$\sum_{i=1}^{n} h_{ii} = \text{rank}(\mathbf{H}) = \text{rank}(\mathbf{X}) = p,$$

em que p é o número de parâmetros do modelo, tem-se que o tamanho médio de um elemento da diagonal da matriz **H** é $\bar{h} = p/n$ e, assim, como uma regra um tanto grosseira, tem-se que:

$$\boxed{\text{se } h_{ii} > 2(p/n) \Rightarrow \text{ a observação } i \text{ é um possível ponto de alta alavancagem.}}$$

6- Diagnósticos de Influência

6.2 Influência nos coeficientes da regressão

Se for desejado, contudo, considerar a localização do ponto, bem como a variável resposta, Cook (1979) sugeriu o uso de uma medida que considera o quadrado da distância entre as estimativas $\hat{\boldsymbol{\beta}}$ obtidas com todas as n observações (pontos) e as estimativas obtidas retirando-se a i-ésima observação (ponto), denotada por $\hat{\boldsymbol{\beta}}_{(i)}$. Essa medida é expressa por:

$$D_i = \frac{(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)})'\mathbf{X}'\mathbf{X}(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)})}{p \text{ QMres}} \qquad i = 1, \ldots, n.$$

Pontos com grandes valores de D_i têm considerável influência nas estimativas $\hat{\boldsymbol{\beta}}$. Os valores de D_i são comparados com a distribuição $F_{\alpha,\, p,\, n-p}$. Se $D_i \approx F_{\alpha,\, p,\, n-p}$, então retirar o ponto i deve deslocar $\hat{\boldsymbol{\beta}}$ para o limite de uma região de confiança de 50% de $\boldsymbol{\beta}$ baseado nos dados completos. Isto é uma grande discordância e indica que as estimativas obtidas por MQO são sensíveis ao i-ésimo ponto. Como $F_{0.5;\, n;\, n-p} \approx 1$, usualmente consideram-se os pontos em que $D_i > 1$ como sendo possivelmente influentes. Idealmente, seria desejado que cada estimativa $\hat{\boldsymbol{\beta}}_{(i)}$ permanecesse dentro dos limites de uma região de confiança de 10 ou 20%.

6- Diagnósticos de Influência

Belsley, Kuh e Welsch (1980) sugeriram, ainda, uma estatística que indica o quanto cada coeficiente de regressão $\hat{\beta}_j$ muda, em unidades de desvio-padrão, se a i-ésima observação for removida. Esta estatística é dada, para j = 0, 1, ..., k, por:

$$\text{DFBeta}_{j,i} = \frac{\hat{\beta}_j - \hat{\beta}_{j(i)}}{\sqrt{(S_{(i)})^2 C_{j+1,j+1}}} \qquad i = 1, \ldots, n,$$

sendo $C_{j+1,j+1}$ o (j +1)-ésimo elemento da diagonal da matriz $\mathbf{C} = (\mathbf{X'X})^{-1}$.

Um valor grande de $\text{DFBeta}_{j,i}$ indica que a observação i tem considerável influência no j-ésimo coeficiente de regressão. O ponto de corte $2/\sqrt{n}$ é, em geral, usado para comparar os $\text{DFBetas}_{j,i}$. Para amostras grandes, observações as quais $|\text{DFBeta}_{j,i}| > 2/\sqrt{n}$ merecem atenção. Para amostras pequenas ou moderadas, as observações que merecem atenção são aquelas em que $|\text{DFBeta}_{j,i}| > 1$.

6- Diagnósticos de Influência

6.3 Influência nos valores ajustados

É possível, também, investigar a influência da i-ésima observação nos valores ajustados (preditos). Uma medida razoável é:

$$\mathrm{DFFit}_i = \frac{\hat{y}_i - \hat{y}_{(i)}}{\sqrt{(S_{(i)})^2 h_{ii}}} \qquad i = 1, \ldots, n,$$

sendo $\hat{y}_{(i)}$ o valor predito de $\hat{y}_i$ sem o uso da i-ésima observação. O denominador é somente uma padronização. Assim, DFFit_i é o quanto o valor ajustado muda, em unidades de desvio-padrão, se a i-ésima observação for removida.

Geralmente, observações em que $|\mathrm{DFFit}_i| > 1$, para amostras pequenas ou moderadas, e $|\mathrm{DFFit}_i| > 2\sqrt{p/n}$, para amostras grandes, merecem atenção.

6.4 Influência na precisão da estimação

As medidas D_i, $\text{DFBeta}_{j,i}$ e DFFit_i fornecem uma visão do efeito de cada observação nos coeficientes estimados e nos valores ajustados. Elas não fornecem, contudo, qualquer informação sobre a precisão geral da estimação. Para expressar o papel da i-ésima observação na precisão da estimação pode ser definido a medida a seguir.

$$\text{Covratio}_i = \frac{\left|(X'_{(i)}X_{(i)})^{-1}(S_{(i)})^2\right|}{\left|(X'X)^{-1}\text{QMres}\right|} \qquad i = 1, \ldots, n.$$

Pontos de corte para Covratio_i não são fáceis de serem obtidos. Belsley et al.(1980) sugeriram que se $\text{Covratio}_i > 1 + (3p/n)$ ou se $\text{Covratio}_i < 1 - (3p/n)$, então, o i-ésimo ponto deve ser um possível ponto influente. O limite inferior é somente válido quando $n > 3p$. Em geral, esses pontos de corte são mais apropriados para amostras grandes.

7- Métodos para tratar com a multicolineariedade

- **Coleta adicional de dados**

Coletar dados adicionais para combinações de X_i e X_j

- **Reespecificação do modelo**

$X = (X_1+X_3)/X_2$ ou $X = X_1*X_2*X_3$

Eliminação de regressoras

- **Regressão Ridge**

Encontrar um estimador $\hat{\boldsymbol{\beta}}^*$ viciado, tal que seu vício seja pequeno, mas que sua variância seja menor do que a de $\boldsymbol{\beta}$. O termo *regressão ridge* é usado para denominar um modelo de regressão em que esse tipo de estimador é considerado. Para mais detalhes sobre esse assunto, pode ser consultado, por exemplo, o livro de Montgomery e Peck (1992)

8- Seleção de variáveis e construção de modelos

1) o modelo deveria incluir tantas quantas regressoras fossem necessárias para auxiliar na predição de Y e,

2) o modelo deveria ser parcimonioso (conter poucas regressoras), visto que a variância da predição cresce conforme o número de regressoras cresce. Além disso, quanto mais regressoras existirem no modelo, maior o custo para coleta e manutenção do modelo.

O processo de encontrar um modelo que concilie esses objetivos é denominado
seleção da melhor equação de regressão

8.1- Critérios para avaliação dos modelos

a) Coeficiente de determinação múltiplo R^2

O valor de R^2 cresce quando *k* cresce e é máximo quando todas as *k* regressoras são usadas. Assim, o analista pode usar o critério de adicionar regressoras até o ponto em que a adição de uma variável não for mais útil, pois fornece um acréscimo muito pequeno em R^2

b) Coeficiente de determinação múltiplo ajustado R^2a ou QMres

O critério é escolher o subconjunto de regressoras que forneça o valor máximo de R^2a, o que equivale a encontrar o subconjunto que minimize o QMres.

c) Estatística Cp de Mallows

$$\boxed{C_p = \frac{SQres(p)}{\sigma^2} - n + 2p}$$

em que σ^2 = QMres e p é o número de parâmetros em cada modelo. Valores pequenos de Cp são desejáveis. Modelos de regressão com Cp próximos da linha Cp = p e abaixo dela são candidatos ao ***melhor modelo.***