Modelo de regressão linear geral

Em geral, a variável resposta Y pode estar relacionada com p variáveis explicativas. O modelo de regressão é definido por:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip} + \varepsilon_i, \qquad (2)$$

é conhecido como modelo de primeira ordem, pois é linear nos parâmetros e linear nas variáveis explicativas. Esse modelo é conhecido como um modelo de regressão linear múltipla ou também como um modelo de regressão linear geral.

- Y_i é o valor da variável resposta para a i-ésima observação,
- β_0, $\beta_1, \ldots, \beta_p$ são parâmetros desconhecidos,
- $X_{i1}, X_{i2}, \ldots, X_{ip}$ são constantes conhecidas,
- ε_i são independentes e $N(0, \sigma^2)$, $i = 1, \ldots, n$.

Modelo de Regressão Linear Geral em termos matriciais

- Dessa forma, o modelo (2) pode ser escrito em termos matriciais:

$$\mathbf{Y}_{nx1} = \mathbf{X}_{nx(p+1)}\boldsymbol{\beta}_{(p+1)x1} + \boldsymbol{\varepsilon}_{nx1}$$

- As condições do erro do modelo (2) são:

$$E(\boldsymbol{\varepsilon})_{nx1} = \mathbf{0}_{nx1} \quad \text{e} \quad Var(\boldsymbol{\varepsilon})_{nxn} = \sigma^2\mathbf{I}_{nxn}$$

- Consequentemente,

$$E(\mathbf{Y}) = E(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) = \mathbf{X}\boldsymbol{\beta} + E(\boldsymbol{\varepsilon}) = \mathbf{X}\boldsymbol{\beta}$$
$$Var(\mathbf{Y}) = Var(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) = \mathbf{0} + Var(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}$$

- O estimador de mínimos quadrados de $\boldsymbol{\beta}$ é dada por:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \qquad (3)$$

Estimador de σ^2

- Como definido anteriormente, o estimador de σ^2 é dado por:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-(p+1)} = \frac{SQRes}{n-(p+1)} = MSRes$$

- Em termos matriciais:

$$\hat{\sigma}^2 = \frac{\mathbf{Y}'\mathbf{Y} - \hat{\beta}'\mathbf{X}'\mathbf{Y}}{n-(p+1)} = \frac{SQRes}{n-(p+1)} = MSRes$$

Multicolinearidade

- Em regressão múltipla, o interesse principal é verificar quais variáveis explicativas estão relacionadas com a variável resposta de forma significativa. Algumas perguntadas de interesse:
  - Qual a importância relativa dos efeitos das diferentes variáveis explicativas?
  - Qual a magnitude do efeito de uma dada variável explicativa sobre a variável resposta ?
  - Pode alguma variável explicativa ser retirada do modelo porque seu efeito é pequeno ou não tem efeito sobre a variável resposta?
  - Qualquer variável explicativa ainda não incluída no modelo deve ser considerada para possível inclusão ?

Multicolinearidade

- Para responderas as questões anteriores e fazer a análise de regressão múltipla correta é necessário verificar se:
  - As variáveis explicativas são **Não** correlacinadas entre si, ou,
  - Se as variáveis explicativas **São** correlacionadas entre si.

Multicolinearidade

Exemplo: Considere o estudo sobre o efeito do tamanho do grupo de trabalho e do número de bônus pagos sobre o escore de produtividade do grupo.

Tabela : Variáveis explicativas não correlacionadas

Tamanho do grupo	Bônus pago	Escore de produtividade
4	2	42
4	2	39
4	3	48
4	3	51
6	2	49
6	2	53
6	3	61
6	3	60

Multicolinearidade

Tabela : Matriz de correlação das variáveis

cor	y	x1	x2
y	1.000000000000000	0.741930890891587	0.638405650302063
x1	0.741930890891587	1.000000000000000	0.000000000000000
x2	0.638405650302063	0.000000000000000	1.000000000000000

Multicolinearidade

Tabela : $\hat{Y} = 0.375 + 5.375X_1 + 9.250X_2$

Fonte de Variação	g.l.	SQ	MS
Regressão	2	402.250	201.125
Residuo	5	17.625	3.525
SQT	7	419.875	

Multicolinearidade

Tabela : $\hat{Y} = 23.50 + 5.375X_1$

Fonte de Variação	g.l.	SQ	MS
Regressão	1	231.125	231.125
Residuo	6	188.750	31.458
SQT	7	419.875	

Multicolinearidade

Tabela : $\hat{Y} = 27.250 + 9.250X_2$

Fonte de Variação	g.l.	SQ	MS
Regressão	1	171.125	171.125
Residuo	6	248.750	41.458
SQT	7	419.875	

Multicolinearidade

- Ao analisar as Tabelas anteriores, verifica-se que as estimativas dos coeficientes de regressão são as mesmas. Tanto no modelo com uma única variável explicativa, como no modelo com as duas variáveis explicativas.

- Esse resultado indica que as variáveis explicativas são **Não correlacionadas**.

Multicolinearidade

1) Encontre a soma de quadrados extra: $SQReg(X_1|X_2)$

2) Encontre a soma de quadrados extra: $SQReg(X_2|X_1)$

Multicolinearidade

1)

$$SQReg(X_1|X_2) = 248.750 - 17.625 = 231.125 = SQReg(X_1)$$

2)

$$SQReg(X_2|X_1) = 188.750 - 17.625 = 171.125 = SQReg(X_2)$$

Multicolinearidade

Conclusões:

- Se as variáveis explicativas **são não correlacionadas**, os efeitos atribuídos a elas pelo modelo de regressão de 1ª ordem são os mesmos não importando que outras variáveis estão incluídas no modelo.

- A contribuição marginal de uma variável explicativa na redução da *SQRes* quando outras variáveis explicativas estão no modelo é exatamente a mesma de quando a variável explicativa está sozinha no modelo.

Multicolinearidade

Efeitos da Multicolinearidade:

- Adicionar e retirar uma variável explicativa do modelo muda os coeficientes de regressão;
- A soma de quadrados associada com uma variável explicativa varia dependendo de quais variáveis independentes já estão no modelo;
- Os coeficientes de regressão podem não ser, individualmente, estatisticamente significantes mesmo que exista relação estatística entre a variável resposta e o conjunto de variáveis explicativas.

Multicolinearidade

Voltando ao estudo sobre a quantidade de gordura corporal

Multicolinearidade

- Um estudo sobre a relação entre a quantidade de gordura corporal e algumas possíveis variáveis explicativas foi realizado com base em uma amostra de 20 mulheres saudáveis com idade entre 25 e 34 anos.
- As possíveis variáveis explicativas são:
  - X_1: tríceps cutâneo
  - X_2: circunferência da coxa
  - X_3: circunferência do braço médio
- A quantidade de gordura no corpo de cada uma das 20 pessoas da amostra foi obtida por um procedimento incômodo e caro que requer a imersão da pessoa em água. Seria, portanto, muito útil se um modelo regressivo com algumas ou todas as variáveis preditoras pudesse fornecer estimativas confiáveis da gordura do corpo, uma vez que as medidas necessárias para estas variáveis são fáceis de serem obtidas.

Multicolinearidade

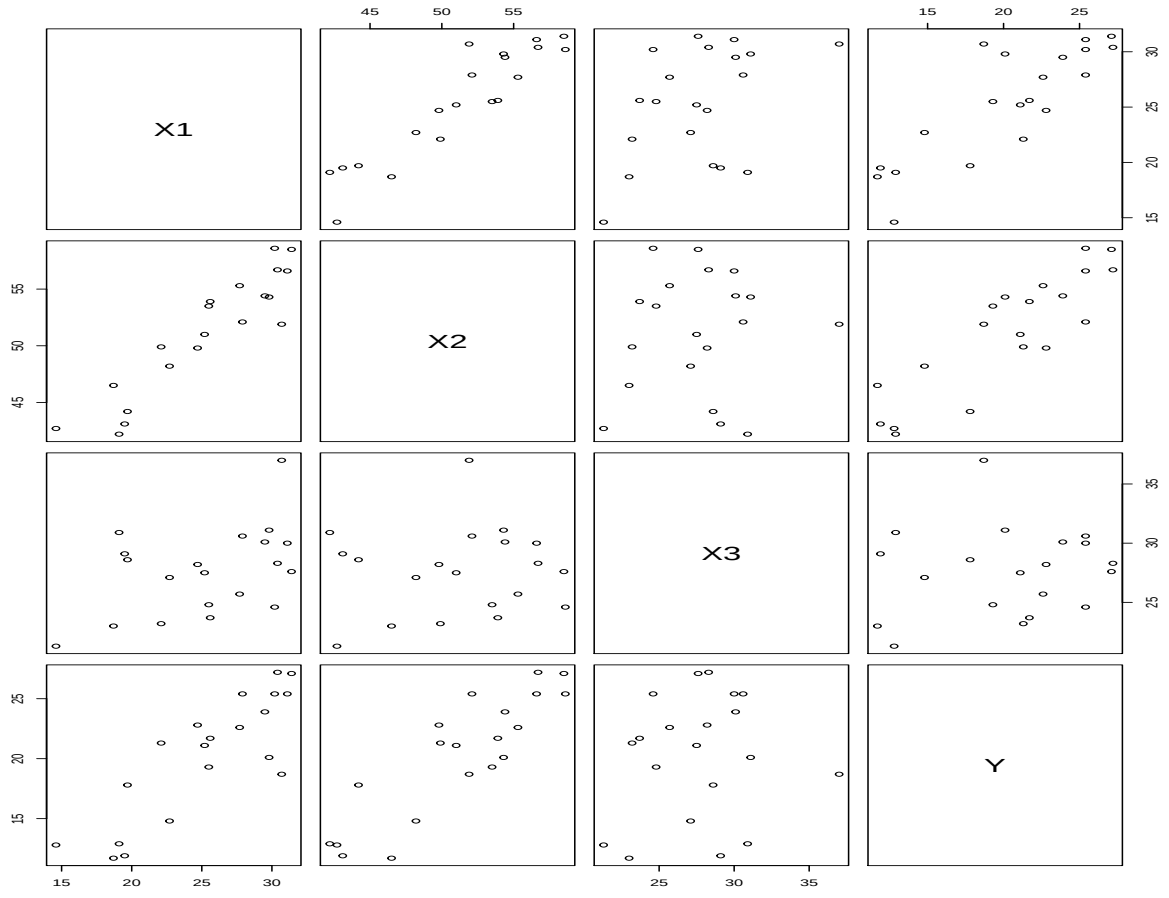

Figura : Diagrama de dispersão dos dados

Multicolinearidade

cor	X1	X2	X3	Y
X1	1.0000	0.9238	0.4578	0.8433
X2	0.9238	1.0000	0.0847	0.8781
X3	0.4578	0.0847	1.0000	0.1424
Y	0.8438	0.8781	0.1424	1.0000

Multicolinearidade

Tabela : **Efeito nos coeficientes de regressão**

Variáveis no Modelo	$\hat{\beta}_1$	$\hat{\beta}_2$
X_1	0.8572	-
X_2	-	0.8565
X_1, X_2	0.2224	0.6594
X_1, X_2, X_3	4.334	-2.857

$\rightarrow$ Pode-se dizer, que o efeito da variável explicativa na variável resposta é apenas marginal ou parcial, dado que outras variáveis explicativas são incluídas no modelo.

Multicolinearidade

Efeito na soma de quadrados extra

$SQReg(X_1) = 352.27$ $\qquad$ $SQReg(X_2) = 381.97$

$SQReg(X_1|X_2) = 3.47$ $\qquad$ $SQReg(X_2|X_1) = 33.17$

- A razão pela soma de quadrados extra ser pequena quando comparada com a soma de quadrados de uma única variável explicativa no modelo, é que X_1 e X_2 são altamente correlacionadas entre elas e cada uma com a variável resposta.
- Então, quando uma variável já está no modelo, a contribuição marginal da outra variável em reduzir a soma de quadrados do erro é pequena.

Multicolinearidade

3) $R^2_{Y1} = 0.71$

4) $R^2_{Y1|2} = 0.03$

- Fica fácil de observar que o **efeito da multicolinearidade** também afeta o resultado do **Coeficiente de determinação parcial**.

Multicolinearidade

Tabela : **Efeito no erro padrão dos coeficientes de regressão**

Variáveis no Modelo	EP($\hat{\beta}_1$)	EP($\hat{\beta}_2$)
X_1	0.1288	-
X_2	-	0.1100
X_1, X_2	0.3034	0.2912
X_1, X_2, X_3	3.016	2.582

$\rightarrow$ Pode-se perceber que o efeito da multicolinearidade também é responsável pela inflação na variabilidade das estimativas dos coeficentes de regressão.

TABELA : **Efeito nos valores ajustados e em predição**

Variáveis no Modelo	*MSRes*
X_1	7.95
X_1, X_2	6.47
X_1, X_2, X_3	6.15

$\rightarrow$ Pode-se verificar uma alteração nos valores do quadrado médio residual quando variáveis correlacionadas são inseridas no modelo. Esse fato implica alteração nas estimativas dos valores ajustados e preditos. Pois, $Var(\hat{Y}_0) = \sigma^2 \mathbf{X}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_0$ e

$$Y_N - \hat{Y}_N \sim \mathrm{N}\left(0, \sigma^2\left[1 + \mathbf{X}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_0\right]\right)$$

Multicolinearidade

- Como observar efeito da multicolinearidade:
  - Mudanças grandes nas estimativas dos coeficientes de regressão quando: uma variável explicativa é adicionada ou retirada do modelo ou quando, uma observação é alterada ou apagada.
  - Nos testes individuais sobre os coeficientes de regressão para variáveis explicativas importantes do modelo aceita-se $H_0 : \beta_k = 0$.
  - Coeficientes de regressão estimados com sinal algébrico que é o oposto do que se espera através de considerações teóricas ou experiência anterior.
  - Coeficientes de correlação alto entre pares de variáveis explicativas.
  - Intervalos de confiança para os coeficientes de regressão de variáveis explicativas importantes apresentam grande amplitude.

Multicolinearidade

Limitações:

- Estes métodos informais não fornecem medidas quantitativas do impacto da multicolinearidade e não podem identificar a natureza da multicolinearidade.

- Algumas vezes o comportamento observado pode ocorrer sem que exista de fato multicolinearidade.

Diagnóstico

- A análise de diagnóstico é um importante procedimento para avaliar a adequabilidade do modelo de regressão múltipla.

- Para verificar a adequabilidade do modelo serão usados:
  - Métodos Gráficos
  - Testes de Hipóteses.

Diagnóstico

- É interessante iniciar a análise de dados com os seguintes gráficos: Box plot, ramo e folhas, diagrama de dispersão univariado de cada uma das variáveis explicativas e também da variável resposta.
- O próximo passo é realizar um estudo bidimensional entre cada variável explicativa e a variável resposta e também entre duas variáveis explicativas. Para esse estudo pode-se usar o diagrama de dispersão, bem como, o coeficiente de correlação linear de Pearson quando as duas variáveis em estudo são quantitativas.

Diagnóstico

- Quando ambas as variáveis são qualitativas, é indicado o cálculo do coeficiente de contingência modificado. Quando uma variável é quantitativa e a outra qualitativa, pode-se fazer análise via Box plot ou cálculo da medida R^2 com base na variância da variável quantitativa e variância separada pelas categorias da variável qualitativa.
- Esse estudo dará informações preliminares sobre os dados.

Diagnóstico

- Gráficos do resíduo versus os valores ajustados é utilizado para:
  - Avaliar a adequação da função de regressão múltipla
  - Os erros do modelo tem Variância constante
  - Presença de valor atípico
- Gráfico do resíduo absoluto versus os valores ajustados também é usado para verificar se os erros do modelo tem Variância constante.
- Gráfico dos resíduos versus tempo ou uma outra sequência é usado para verificar se os Erros são Independentes.
- Box plot e o gráfico normal de probabilidade são usados para verificar se os Erros são Normais.

Diagnóstico

- Gráficos dos resíduos versus cada variável explicativa dará informação sobre a adequação da função de regressão com respeito a variável explicativa considerada.
- Gráficos dos resíduos versus cada variável explicativa também informará possível variação da variância que pode estar relacionada com a variável explicativa considerada.
- Gráficos dos resíduos versus variável explicativa não introduzida no modelo ou ainda, versus interação, $X_1X_2, X_1X_3, X_2X_3, ...$, mostrará se a variável ou a interação devem ser introduzidas no modelo.

Diagnóstico

- Forma gráfica ideal e não ideal para os pressupostos do modelo serem válidos:

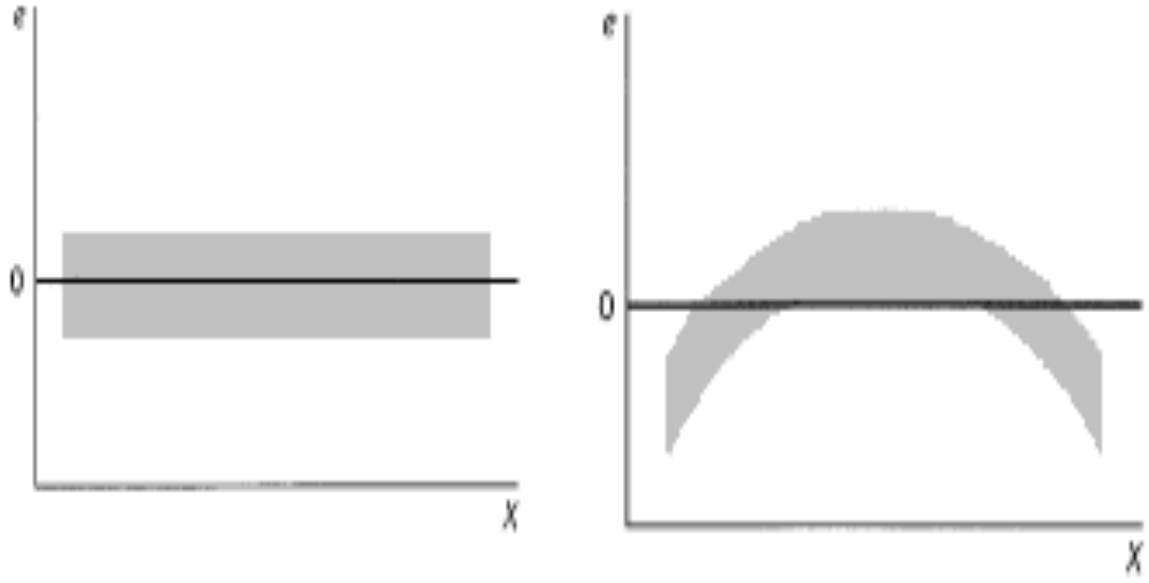

Figura : Modelo de regressão linear é apropriado. (b) Modelo de regressão linear não é apropriado

Diagnóstico

- Teste Shapiro Wilks - para normalidade dos erros
- Teste Brown-Forsythe e Breusch-Pagan para variância constante dos erros. São utlizados quando há evidências de que a variância do erro ou cresce ou decresce com uma determinada variável explicativa.
- Teste F para falta de ajustamento do modelo de regressão

Identificando observações discrepantes em relação à Y

Resíduos

- Resíduos - como já definido, o resíduo é dado por:

$$e_i = Y_i - \hat{Y}_i$$

- Resíduo semistudentizado ou padronizado

$$e_i^* = \frac{e_i}{\sqrt{MSRes}}$$

Matriz H - "chapéu"

- A matriz **H** é definida por:

$$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

- Os valores de $\hat{Y}_i$ podem ser expressos como uma combinação linear dos Y_i através da matriz **H**:

$$\hat{\mathbf{Y}} = \mathbf{H}\mathbf{Y}$$

- E os resíduos e_i também podem ser expressos como uma combinação linear dos Y_i através da matriz **H**:

$$\mathbf{e} = (\mathbf{I} - \mathbf{H})\mathbf{Y}$$

Matriz H - "chapéu"

- A matriz de variância e covariância do resíduo é definida como:

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

- Dessa forma, a variância do resíduo, e_i é dada por:

$$Var(e_i) = \sigma^2(1 - h_{ii})$$

- E a covariância entre e_i e e_j, para $i \neq j$, é:

$$Cov(e_i, e_j) = \sigma^2(0 - h_{ij}) = -h_{ij}\sigma^2$$

em que h_{ii} são os elementos da diagonal principal da matriz $\mathbf{H}$ e h_{ij} são os elementos da i-ésima linha e j-ésima coluna da matriz $\mathbf{H}$.

Matriz H - "chapéu"

- Ao usar MSRes, o estimador da variância do erro - σ^2, as estimativas das variâncias e covariâncias são definidas por:

$$\hat{Var}(e_i) = MSRes(1 - h_{ii})$$

e

$$\hat{Cov}(e_i, e_j) = -h_{ij} MSRes$$

Resíduo Studentizado

- Os resíduos e_i podem ter variâncias substancialmente diferentes.
- Por isso, é importante considerar a magnitude de cada resíduo relativa ao desvio padrão estimado de cada resíduo. Dando origem ao Resíduo Studentizado:

$$r_i = \frac{e_i}{\sqrt{MSRes(1 - h_{ii})}}$$

- O resíduo studentizado tem variância constante, $Var(r_i) = 1$.

Resíduo excluído

- Se a i-ésima observação, Y_i, é realmente incomum, discrepante, o modelo de regressão ajustado usando todas as observações pode ser influenciado por essa observação.
- Uma outra medida para verificar se a i-ésima observação, Y_i, é discrepante, é o resíduo excluído definido por:

$$e_{(i)} = Y_i - \hat{Y}_{i(i)},$$

também conhecido como erro de predição PRESS para o i-ésimo caso.

- Uma expressão equivalente para $e_{(i)}$ é:

$$e_{(i)} = \frac{e_i}{(1 - h_{ii})}$$

- Dessa forma, o critério PRESS pode ser obtido por:

$$PRESS_{p+1} = \sum_{i=1}^{n} \left(\frac{e_i}{(1 - h_{ii})} \right)^2$$

Resíduo excluído

- A variância do resíduo excluído é dada por:

$$\begin{aligned} Var(e_{(i)}) &= \frac{1}{(1-h_{ii})^2} Var(e_i) = \frac{1}{(1-h_{ii})^2}[\sigma^2_{(i)}(1-h_{ii})] \\ &= \frac{\sigma^2_{(i)}}{(1-h_{ii})} \end{aligned}$$

- Consequentemente, a variância estimada é definida por:

$$\hat{Var}(e_{(i)}) = \frac{MSRes_{(i)}}{(1-h_{ii})},$$

em que $MSRes_{(i)}$ é o MSRes do modelo quando a i-ésima observação é excluída.

Resíduo excluído studentizado

- Seguindo a definição do resíduo studentizado, o resíduo excluído studentizado é definido por:

$$t_i = \frac{e_{(i)}}{\sqrt{Var\hat{(}e_{(i)})}} \sim \text{t-Student}_{n-1-(p+1)}$$

ou ainda,

$$t_i = e_{(i)} \sqrt{\frac{n-1-(p+1)}{SQRes(1-h_{ii}) - e^2_{(i)}}}$$

Identificando observações discrepantes em relação à X

- Pontos potencialmente distantes tem impacto nas estimativas dos parâmetros, erro padrão, valores preditos e estatísticas do modelo. A matriz $\mathbf{H}$

$$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

é importante para detectar observações influentes.

- Os elementos h_{ii} da matriz $\mathbf{H}$ são definidos por:

$$h_{ii} = \mathbf{X}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_i,$$

em que $\mathbf{X}_i$ é a i-ésima linha da matriz $\mathbf{X}$. A diagonal da matriz $\mathbf{H}$ é uma medida padronizada da distância da i-ésima observação do centro do espaço de X.

- Valor grande de h_{ii} indica que a i-ésima observação está distante do centro das observações e que essa observação pode ser considerada um ponto de alavanca (leverage point).

- Outra forma de verificar ponto de alavanca é quando: $h_{ii} > 2\bar{h}$, em que $\bar{h} = \sum_{i=1}^{n} h_{ii}/n$.

- Após identificar valores discrepantes com respeito aos valores de Y ou X, o próximo passo é verificar se essas observações são ou não observações influentes.

- Medidas para identificar observações influentes são:
  - DFFITS
  - Distância de Cook
  - DFBETAS

DFFITS

- Medida da influência que a i-ésima observação tem sobre o valor ajustado $\hat{Y}_i$ é dada por:

$$(\text{DFFITS})_i = \frac{\hat{Y}_i - \hat{Y}_{i(i)}}{\sqrt{MSRes_{(i)} h_{ii}}}$$

em que:

- $\hat{Y}_i$ valor ajustado para o i-ésimo caso quando todas as n observações são usadas no ajuste do modelo
- $\hat{Y}_{i(i)}$ valor ajustado para o i-ésimo caso quando o i-ésimo caso é omitido no ajuste do modelo
- $MSRes_{(i)}$ quando o i-ésimo caso é omitido no ajuste do modelo.

DFFITS

- O valor de DFFITS pode ser obtido usando apenas o resultado do ajuste do modelo com todos os dados por meio de:

$$(\text{DFFITS})_i = t_i \left(\frac{h_{ii}}{1 - h_{ii}} \right)^{1/2}$$

- Para identificar se uma observação é influente:
- Se $|(\text{DFFITS})_i| > 1$, para conjunto de dados pequenos ou médios,
- Se $|(\text{DFFITS})_i| > 2\sqrt{(p+1)/n}$, para conjunto de dados grandes.

Distância de Cook

- Cook(1977,1979) sugeriu uma medida usando a distância ao quadrado entre todas as estimativas $\hat{\boldsymbol{\beta}}$, e a estimativa obtida ao exluir a i-ésima observação, $\hat{\boldsymbol{\beta}}_{(i)}$. Essa medida da distância pode ser expressa por:

$$D_i = \frac{(\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}})' \mathbf{M} (\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}})}{c},$$

em que $\mathbf{M}$ e c são usualmente $\mathbf{M} = (\mathbf{X}'\mathbf{X})$ e $c = (p+1)MSRes$. Logo:

$$D_i = \frac{(\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}})' (\mathbf{X}'\mathbf{X}) (\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}})}{(p+1)MSRes}.$$

Distância de Cook

- A distância de Cook, medida da influência que a i-ésima observação tem sobre todos os n valores ajustados, também pode ser definida por:

$$D_i = \sum_{i=1}^{n} \frac{(\hat{Y}_i - \hat{Y}_{i(i)})^2}{(p+1)MSRes}$$

- A distância de Cook pode ser obtida usando apenas o resultado do ajuste do modelo com todos os dados por meio de:

$$D_i = \frac{e_i^2}{pMSRes}\left[\frac{h_{ii}}{(1-h_{ii})^2}\right]$$

- Valores de $D_i > 1$ são consideradas observações influentes.

DFBETAS

- Medida da influência que a i-ésima observação tem sobre cada um dos coeficientes de regressão estimados é dada por:

$$(\text{DFBETAS})_{k(i)} = \frac{\hat{\beta}_k - \hat{\beta}_{k(i)}}{\sqrt{MSRes_i C_{kk}}}$$

em que:

- $\hat{\beta}_k$ coeficiente de regressão estimado considerando todos os n casos no ajuste do modelo
- $\hat{\beta}_{k(i)}$ coeficiente de regressão estimado considerando que o i-ésimo caso é omitido no ajuste do modelo
- $MSRes_i$ é o MSRes quando o i-ésimo caso é omitido no ajuste do modelo
- C_{kk} o k-ésimo elemento da diagonal da matriz $(\mathbf{X}'\mathbf{X})^{-1}$.

DFBETAS

Interpretação da medida DFBETAS:

- Sinal - indica se a inclusão do i-ésimo caso leva ao aumento ou a diminuição da estimativa do coeficiente de regressão
- Valor grande de $(\text{DFBETAS})_{k(i)}$ indica grande influência do i-ésimo caso na estimativa do k-ésimo coeficiente de regressão.
- Se $|(\text{DFBETAS})_{k(i)}| > 1$, para conjunto de dados pequenos ou médios, $\Rightarrow$ observação é influente
- Se $|(\text{DFBETAS})_{k(i)}| > 2/\sqrt{n}$, para conjunto de dados grandes $\Rightarrow$ observação é influente.

Diagnóstico de Multicolinearidade

Diagnóstico de Multicolinearidade

Diagnóstico informal

- Coeficientes de correlação alto entre pares de variáveis explicativas.
- Mudanças grandes nas estimativas dos coeficientes de regressão quando:
  - uma variável explicativa é adicionada ou retirada do modelo;
  - uma observação é alterada ou apagada.
- Nos testes individuais sobre os coeficientes de regressão para variáveis explicativas importantes do modelo aceita-se $H_0 : \beta_k = 0$
- Coeficientes de regressão estimados com sinal algébrico oposto do que se espera através de considerações teóricas ou experiência anterior
- Intervalos de confiança para os coeficientes de regressão de variáveis explicativas importantes apresentam grande amplitude.

Diagnóstico de Multicolinearidade

Limitações:

- Estes métodos informais não fornecem medidas quantitativas do impacto da multicolinearidade e não podem identificar a natureza da multicolinearidade.

- Algumas vezes o comportamento observado pode ocorrer sem que exista de fato multicolinearidade.

Diagnóstico de Multicolinearidade

Fator de inflação da Variância (VIF)

- Esse fator mede quanto a variância dos estimadores de mínimos quadrados são influenciadas quando comparada com variáveis explicativas que não são correlacionadas.
- Para entender esse fator, começa-se com a matriz de variâncias e covariâncias dos estimadores de mínimos quadrados:

$$Var(\beta) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$$

- Para medir o impacto da multicolinearidade, é útil utilizar o modelo de regressão padronizado definido em aulas anteriores para reduzir os erros de arredondamento no cálculo da matriz $(\mathbf{X}'\mathbf{X})^{-1}$.

Diagnóstico de Multicolinearidade

- No modelo transformado os coeficientes de regressão estimados $\hat{\beta}_k^*$ são padronizados e sua matriz de covariâncias é dada por:

$$Var(\boldsymbol{\beta}^*) = (\sigma^*)^2(\mathbf{r_{xx}})^{-1},$$

em que $\mathbf{r_{xx}}$ é a matriz de correlação simples entre os pares de variáveis X e $(\sigma^*)^2$ é a variância do erro do modelo transformado.

Diagnóstico de Multicolinearidade

- Porém, $(\mathbf{r_{xx}})^{-1}$ é o fator de inflação da variância (VIF) para β_k^*. Logo:

$$Var(\beta_k^*) = (\sigma^*)^2(\text{VIF})_k,$$

- Os elementos da diagonal de $(\text{VIF})_k$ é o fator de inflação da variância (VIF) para β_k^* e é igual a:

$$(\text{VIF})_k = (1 - R_k^2)^{-1},$$

em que R_k^2 é o coeficiente de determinação múltipla quando X_k é retirado do modelo que contém $(p = (p + 1) - 1)$ variáveis .

Diagnóstico de Multicolinearidade

- Se $R_k^2 = 0 \Rightarrow (\text{VIF})_k = 1$ e X_k não está correlacionada com as demais variáveis.
- Se $R_k^2 \neq 0 \Rightarrow (\text{VIF})_k > 1$ e X_k está correlacionada com as demais variáveis.
- Quando X_k está perfeitamente relacionada com as outras variáveis do modelo $\Rightarrow R_k^2 = 1$ e $(\text{VIF})_k$ será ilimitado (valor grande).

Diagnóstico de Multicolinearidade

Indicador de Multicolinearidade

- Se o máximo dos $(\text{VIF})_k > 10 \Rightarrow$ Multicolinearidade está influenciando as estimativas dos parâmetros.
- Se $R_k^2 = 0 \Rightarrow (\text{VIF})_k = 1 \Rightarrow$ Não existe Multicolinearidade
- Outra forma de verificar é obter:

$$(\bar{\text{VIF}})_k = \frac{\sum_{k=1}^{p} (\text{VIF})_k}{p}$$

- Se $(\bar{\text{VIF}})_k$ for um valor consideravelmente maior que $1 \Rightarrow$ Multicolinearidade está influenciando as estimativas dos parâmetros.