MODELO LINEAR CLÁSSICO

Frederico Machado Almeida
frederico.almeida@unb.br

Departamento de Estatística
Instituto de Exatas
Universidade de Brasília

Inverso Generalizado, g-inverso

- O conceito clássico de inversa de uma matriz apresentada anteriormente (Slide 21, Aula 01), supõem que a matriz A seja quadrada e não singular, i.e., $|A| \neq 0$.
- Acontece com alguma frequência em Estatística situações em que a matriz A não é quadrada, ou mesmo, caso seja, pode ocorrer de ser singular.
- Para contornar situações descritas no item anterior, uma teoria unificada é desejável. Tal teoria consiste em usar uma técnica que permite obter a inversa de uma matriz mesmo que ela seja singular ou não quadrada.

Inverso Generalizado, g-inverso

- Antes de apresentar a definição do g-inverso vamos considerar os seguintes resultados:
- Sejam A e B matrizes $m \times n$. Considere igualmente as seguintes matrizes $C_{p\times m}$ e $D_{n\times p}$
  - Se $CA = CB$ então $A = B$
  - Se $AD = BD$ então $A = B$
  - Se $CAD = CBD$ então $A = B$
  - $AB = AC$ se e somente se $A^\top AB = A^\top AC$
  - $EA^\top = FA^\top$ se e somente se $EA^\top A = FA^\top A$.

Inverso Generalizado, g-inverso

Seja A uma matriz $m \times n$; $\mathbf{x}$ um vetor n-dimensional e $\mathbf{y}$ um vetor m-dimensional. De tal forma que, $\mathbf{y}$ pode ser escrito da seguinte forma, $\mathbf{y} = A\mathbf{x}$.

- Definição 1: Uma matriz G é dita ser inversa generalizada (g-inversa), se e somente se $\mathbf{x} = G\mathbf{y}$ é solução de $\mathbf{y} = A\mathbf{x}$, então, segue que $G \equiv A^{-1}$.

- Definição 2: De acordo com a definição anterior, um inverso generalizado (g-inverso) de uma matriz $m \times n$, A é uma matriz $G_{n \times m}$ se ela satisfaz a seguinte relação:

$$AGA = A \tag{1}$$

- Denotamos o g-inverso por A^{-1}.

Inverso Generalizado, g-inverso

Nota 2: Para qualquer matriz $A \in \mathcal{R}^{m\times n}$, sua inversa generalizada sempre existirá, no entanto, ela não será única.

Para verificar tal afirmação, assuma que $A = [1,2] \in \mathcal{R}^{1\times 2}$. A sua inversa generalizada será a matriz $G = [x,y]^\top \in \mathcal{R}^{2\times 1}$ satisfazendo:

$$AGA = A \text{ então, } [1,2]\begin{bmatrix} x \\ y \end{bmatrix}[1,2] = (x+2y)[1,2] = [1,2].$$

Este resultado mostra que, qualquer vetor $G = \begin{bmatrix} x \\ y \end{bmatrix} \in \mathcal{R}^{2\times 1}$ satisfazendo a equação $x + 2y = 1$ será uma inversa generalizada de A. Por exemplo,

$$G = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \text{ ou } G = \begin{bmatrix} 3 \\ -1 \end{bmatrix}.$$

Inverso Generalizado, g-inverso

A matriz inversa generalizada é particularmente importante para resolver sistemas lineares no qual seus coeficientes estão em A.

Teorema

A matriz G, g-inverso de uma matriz real A sempre existe, e $G = A^{-1}$ se A é não singular.

Exemplo 1: Prove que,

$$A_1^{-1} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix} \text{ e } A_2^{-1} = \begin{bmatrix} -42 & 18 \\ -2 & 2 \\ 1 & 0 \end{bmatrix}$$

são inversas generalizadas de

$$A = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 4 & 6 \end{bmatrix}.$$

Inverso Generalizado, g-inverso

Seja $A_{n\times n}$ uma matriz de posto r, se G denota a inversa generalizada de A, então,

- $A^{-1}G$ e GA^{-1} são idempotente
- $(I - A^{-1}G)$ e $(I - GA^{-1})$ são idempotentes

Seja G, g-inverso de $A^\top A$, então

- $G^\top$ é g-inverso de $A^\top A$
- $GA^\top$ é g-inverso de A, tal que $AGA^\top A = A$
- $AGA^\top$ é invariante a escolha de G, ou seja,

$$AG_1A^\top = AG_2A^\top$$

- $AGA^\top$ é simétrica.

Inverso Generalizado, g-inverso

A matriz $\mathbf{X}^\top\mathbf{X}$ tem um papel importante na Estatística, com particular aplicação nos equações dos mínimos quadrados. Propriedades do g-inversa para $\mathbf{X}^\top\mathbf{X}$, são de extrema importância.

Teorema

Seja $\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}$ uma inversa generalizada de $\mathbf{X}^\top\mathbf{X}$. Então, as seguintes propriedades são válidas:

1. $\left[\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\right]^\top$ é também uma g-inversa de $\mathbf{X}^\top\mathbf{X}$.
2. $\mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{X} = \mathbf{X}$, isto é, $\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top$ é uma g-inversa de $\mathbf{X}$.
3. $\mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{X}$ é invariante para $\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}$.
4. $\mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{X}$ é simétrica, mesmo que $\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}$ seja simétrica ou não.

Algumas Propriedades Matricial

Considere as seguintes notações:

- $A_{n\times n}$ é uma matriz de constantes
- $\mathbf{x}_{n\times 1}$ é um vetor de variáveis (ou parâmetros)
- $\mathbf{a}_{n\times 1}$ é um vetor de constantes.

Derivadas de Matrizes:

- $\frac{\partial \mathbf{a}^\top \mathbf{x}}{\partial \mathbf{x}} = \mathbf{a}$
- $\frac{\partial \mathbf{x}^\top \mathbf{x}}{\partial \mathbf{x}} = 2\mathbf{x}$
- $\frac{\partial \mathbf{a}^\top A\mathbf{x}}{\partial \mathbf{x}} = A^\top \mathbf{a}$
- $\frac{\partial \mathbf{x}^\top A\mathbf{x}}{\partial \mathbf{x}} = A\mathbf{x} + A^\top \mathbf{x}.$

Algumas Propriedades Matricial

Frequentemente estamos interessados em calcular a média e a variância de vetores. Por exemplo, assuma que o vetor $\mathbf{W}$ é obtido como uma combinação linear do vetor $\mathbf{y}$ e a matriz $A_{m\times n} = (a_{ij})$, com $i \in \{1, \cdots, m\}$ e $j \in \{1, \cdots, n\}$. Ou seja, $\mathbf{W} = A\mathbf{y}$. Tal que, $\mathbb{E}(\mathbf{Y}) = \boldsymbol{\mu}$ e $\text{Var}(\mathbf{Y}) = \sigma^2 I_n$.

Assim, segue que:

1. $\mathbb{E}(A) = A$
2. $\mathbb{E}(\mathbf{W}) = \mathbb{E}(A\mathbf{Y}) = A\mathbb{E}(\mathbf{Y}) = A\boldsymbol{\mu}$
3. $\text{Var}(\mathbf{W}) = \text{Var}(A\mathbf{Y}) = A\text{Var}(\mathbf{Y})A^\top(*)$
4. $\mathbb{E}(A\mathbf{W}B) = A\mathbb{E}(\mathbf{W})B$
5. $\mathbb{E}\left(\mathbf{Y}^\top A\mathbf{Y}\right) = tr\,(A\Sigma) + \boldsymbol{\mu}^\top A\boldsymbol{\mu}$ (verifique!),

onde $\Sigma = \sigma^2\mathbf{I}_n$. $^{(*)}\text{Var}(\mathbf{Y}) = \mathbb{E}\left[(\mathbf{Y} - \mathbb{E}(\mathbf{Y}))\,(\mathbf{Y} - \mathbb{E}(\mathbf{Y}))^\top\right] = \mathbb{E}\left(\mathbf{Y}\mathbf{Y}^\top\right) - \boldsymbol{\mu}^\top\boldsymbol{\mu}$.

Regressão Linear Múltipla

- Podemos afirmar que estamos perante uma situação envolvendo o modelo de regressão múltipla quando se admite que a resposta Y pode ser modelada por meio de duas ou mais variáveis independentes.
- Assim, como na regressão linear simples, uma regressão linear múltipla supõem que função de regressão $E(Y|X = x)$ é linearmente dependente dos preditores $x_1, \cdots, x_p$.
- Regressões lineares são simples e comumente fornecem um descrição adequada e interpretável de como as variáveis exploratórias afetam as resposta.

Regressão Linear Múltipla

- O modelo estatístico de uma regressão linear múltipla com p-variáveis explicativas $(x_1, \cdots, x_p)^\top$ é dado por:

$$Y = \beta_0 + \sum_{j=1}^{p} \beta_j x_j + \epsilon, \tag{2}$$

  onde β_j's são os coeficientes de regressão, e ϵ o termo de erro.
- As variáveis x_j's podem vir de diferentes fontes; a saber:
  - Inputs quantitativos (valores reais, medidas)
  - Transformação de inputs quantitativos (ex.: log, $\sqrt{}$,etc)
  - Expansão de base (ex.: $x_2 = x_1^2$, $x_3 = x_1^3$, etc)
  - inputs qualitativos ("dummy" ex.: gênero, classes sociais, escolaridade)
  - Interação (ex.: $x_3 = x_1.x_2$)

- Como no caso da regressão linear simples, normalmente tem-se um banco de observações do tipo $(x_1, y_1), \cdots, (x_n, y_n)$, cujo objetivo é estimar o vetor de parâmetros $\boldsymbol{\beta} = (\beta_0, \beta_1, \cdots, \beta_p)^\top$.
- Ao estender a noção de regressão simples para regressão múltipla o método de mínimos quadrados continua sendo utilizado, no qual as estimativas dos β_i's são escolhidas derivando a SQE em relação a $(\beta_0, \beta_1, \cdots, \beta_p)^\top$. Ou seja,

$$
\begin{aligned}
SQE(\boldsymbol{\beta}) &= \sum_{i=1}^{n} \left(Y_i - f(x_i, \boldsymbol{\beta})\right)^2 \\
&= \sum_{i=1}^{n} \left(Y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij}\right)^2
\end{aligned}
$$

Regressão Linear Múltipla

Source: Elements of Statistical Learning (2nd Ed.) Hastie, Tibshirani & Friedman 2009 Chap 3

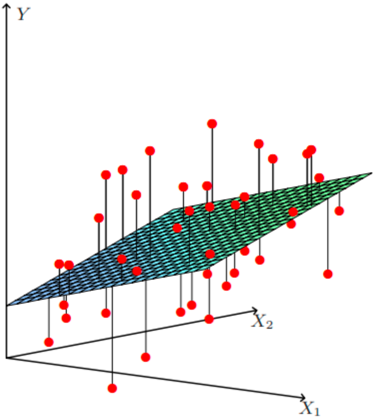

FIGURE 3.1. *Linear least squares fitting with* $X \in \mathbb{R}^2$. *We seek the linear function of* X *that minimizes the sum of squared residuals from* Y.

Regressão Linear Múltipla

- Podemos representar tanto a regressão linear simples, assim como a regressão linear múltipla na forma matricial. Para tal, defina $p = k + 1$ e

$$\underset{n\times 1}{\mathbf{y}} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \ \underset{n\times p}{\mathbf{X}} = \begin{pmatrix} 1 & X_{11} & \cdots & x_{1k} \\ 1 & X_{21} & \cdots & X_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & \cdots & X_{nk} \end{pmatrix}, \ \underset{p\times 1}{\boldsymbol{\beta}} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}, \ \underset{n\times 1}{\boldsymbol{\epsilon}} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}.$$

- Aqui, $\mathbf{X}$ é chamada de matriz desenho (experimental ou de covariáveis). Assim, o modelo de regressão linear múltipla, pode ser escrito como:

$$\underset{n\times 1}{\mathbf{y}} = \underset{n\times p}{\mathbf{X}} \underset{p\times 1}{\boldsymbol{\beta}} + \underset{n\times 1}{\boldsymbol{\epsilon}}, \tag{3}$$

- Portanto, $\mathbb{E}\left(\mathbf{Y}|\mathbf{X} = \mathbf{x}_i\right) = \mathbf{x}_i\boldsymbol{\beta}$. Sendo $\mathbf{x}_i$ um vetor $1 \times p$.

Regressão Linear Múltipla

- Nesse caso, a primeira coluna da matriz $\mathbf{X}_{n\times p}$ foi incluída para acomodar o intercepto.
- É importante salientar que, a menos que haja o problema de multicolinearidade, a matriz $\mathbf{X}$ será de posto completo, i.e., $r\left(\mathbf{X}\right) = p$.
- Tal como no modelo de regressão linear simples, temos que:
  1. Para $\forall_i$, y_i é uma função linear de de $\mathbf{x}_i$
  2. Os $\mathbf{x}_i$'s são fixos (controlados) e medidos sem erros
  3. Os erros tem média 0, i.e., $\mathbb{E}\left(\boldsymbol{\epsilon}\right) = \mathbf{0}$
  4. Os erros são homocedásticos
  5. Os erros são independentes
  6. De (3)-(5) segue que, $\boldsymbol{\epsilon} \sim \mathcal{NM}_n\left(\mathbf{0}_{n\times 1}, \sigma^2\mathbf{I}_n\right)$.

Regressão Linear Múltipla

- Tal como no caso da regressão linear simples, da equação (3) segue que, $\epsilon = \mathbf{y} - \mathbf{X}\beta$.
- Sendo um vetor observado $n \times 1$ de respostas, segue que, a $SQE(\beta)$ será:

$$
\begin{aligned}
SQE(\beta) &= \epsilon^\top \epsilon = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{1i} \cdots - \beta_p x_{pi})^2 \\
&= (\mathbf{y} - \mathbf{X}\beta)^\top (\mathbf{y} - \mathbf{X}\beta) \\
&= \mathbf{y}^\top \mathbf{y} - 2\beta^\top \mathbf{X}^\top \mathbf{y} + \beta^\top \mathbf{X}^\top \mathbf{X}\beta.
\end{aligned}
$$

- Para minimizar a $SQE(\beta)$ devemos derivar em relação ao vetor β, e igualar a 0. Isto é,

$$
\frac{\partial}{\partial \beta} SQE(\beta) = -2\mathbf{X}^\top (\mathbf{y} - \mathbf{X}\beta) \text{ (verifique!)} \tag{4}
$$

Regressão Linear Múltipla

- Igualando (4) ao vetor nulo $\mathbf{0}$, e resolvendo o sistema resultante, obtemos:

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{y},$$

sendo $\hat{\boldsymbol{\beta}} = \left(\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_p\right)^\top$ os estimadores dos mínimos quadrados.

- Um caso particular, seria obter os estimadores dos mínimos quadrados no caso que $\mathbf{X}$ é uma matriz $n \times 2$ (`regressão linear simples`). Assim,

$$\underset{n\times 1}{\mathbf{y}} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad \underset{n\times 2}{\mathbf{X}} = \begin{pmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{pmatrix}, \quad \underset{n\times 1}{\boldsymbol{\epsilon}} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}.$$

Regressão Linear Múltipla

- Passo 1:

$$\mathbf{X}^\top\mathbf{X} = \begin{pmatrix} 1 & 1 & \cdots 1 \\ X_1 & X_2 & \cdots X_n \end{pmatrix} \begin{pmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{pmatrix} = \begin{pmatrix} n & \sum_{i=1}^{n} X_i \\ \sum_{i=1}^{n} X_i & \sum_{i=1}^{n} X_i^2 \end{pmatrix}$$

- Passo 2: Observe que,

$$|\mathbf{X}^\top\mathbf{X}| = n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2 = n\left[\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\right] = nS_{XX}.$$

- Passo 3: Com base no resultado dos passos 1 e 2, segue que:

$$\left(\mathbf{X}^\top\mathbf{X}\right)^{-1} = \frac{1}{nS_{xx}}\begin{pmatrix} \sum_{i=1}^{n} x_i^2 & -\sum_{i=1}^{n} x_i \\ -\sum_{i=1}^{n} x_i & n \end{pmatrix} = \begin{pmatrix} \frac{\sum_{i=1}^{n} x_i^2}{nS_{xx}} & -\frac{\bar{x}}{S_{xx}} \\ -\frac{\bar{X}}{S_{xx}} & \frac{1}{S_{xx}} \end{pmatrix},$$

- Passo 4: Obter a quantidade,

$$\underset{2\times n}{\mathbf{X}^\top}\underset{n\times 1}{\mathbf{y}} = \begin{pmatrix} 1 & 1 & \cdots 1 \\ x_1 & x_2 & \cdots x_n \end{pmatrix}\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i y_i \end{pmatrix}$$

- Passo 5: Por fim, obtemos (Verifique!)

$$\hat{\beta} = \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{y} = \begin{pmatrix} \bar{y} - \bar{x}\frac{S_{xy}}{S_{xx}} \\ \frac{S_{xy}}{S_{xx}} \end{pmatrix} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix}.$$

Regressão Linear Múltipla

- Voltando para a situação envolvendo o modelo de regressão linear múltipla, a reta de regressão estimada no caso da regressão múltipla será:

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\beta} = \mathbf{X}\left(\mathbf{X}^{\top}\mathbf{X}\right)^{-1}\mathbf{X}^{\top}\mathbf{y} = \underset{n\times n}{\mathbf{H}}\,\underset{n\times 1}{\mathbf{y}},$$

  sendo $\mathbf{H} = \mathbf{X}\left(\mathbf{X}^{\top}\mathbf{X}\right)^{-1}\mathbf{X}^{\top}$ a **matriz de projeção** (ou **hat matrix**).

- A **matriz de projeção** apresenta propriedades de extrema importância na estatística.

Regressão Linear Múltipla

Propriedades da matriz de projeção:

1. $\mathbf{H}$ é simétrica, ou seja, $\mathbf{H}^\top = \mathbf{H}$
2. $\mathbf{H}$ é idempotente, ou sejam $\mathbf{H}^2 = \mathbf{H}$
3. $tr\left(\mathbf{H}\right) = \sum_{i=1}^{n} h_{ii} = p$, sendo p o posto de $\mathbf{X}$, caso seja de posto completo
4. $\mathbf{H}$ é a matriz de projeção no plano gerado pelas colunas de $\mathbf{X}$, ou seja, $\mathbf{HX} = \mathbf{X}$, veja a Figura a seguir.

Regressão Linear Múltipla

Source: Elements of Statistical Learning (2nd Ed.) Hastie, Tibshirani & Friedman 2009 Chap 3

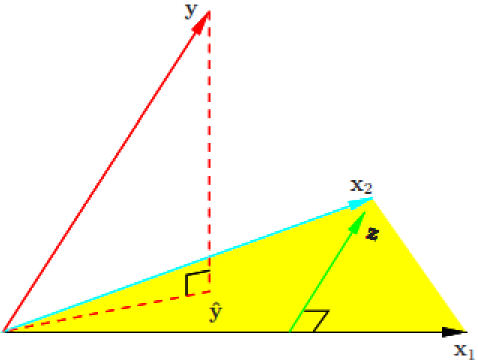

FIGURE 3.4. *Least squares regression by orthogonalization of the inputs. The vector* $\mathbf{x}_2$ *is regressed on the vector* $\mathbf{x}_1$*, leaving the residual vector* $\mathbf{z}$*. The regression of* $\mathbf{y}$ *on* $\mathbf{z}$ *gives the multiple regression coefficient of* $\mathbf{x}_2$*. Adding together the projections of* $\mathbf{y}$ *on each of* $\mathbf{x}_1$ *and* $\mathbf{z}$ *gives the least squares fit* $\hat{\mathbf{y}}$*.*

Regressão Linear Múltipla

- Quando a matriz desenho $\mathbf{X}$ não é de posto completo, isto é, quando $r(\mathbf{X}) < p$, então $\hat{\beta}$ não pode ser estimado de forma única.
- Porém, $\hat{\mathbf{y}} = \mathbf{X}\hat{\beta}$ é único, pois, $\mathbf{H} = \mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top$ unicamente define a projeção de $\hat{\mathbf{y}}$ no espaço gerado por $\mathbf{X}$.
- Portanto apesar das infinitas soluções para $\hat{\beta}$ temos que, $\hat{\mathbf{y}}$ continua sendo estimado de forma única.

Regressão Linear Múltipla

Algumas propriedades importantes:

1. Conforme apresentado no caso da regressão linear simples, é possível mostrar que $\hat{\beta}$ é um estimador não viesado, i.e., $\mathbb{E}\left(\hat{\beta}\right) = \beta$ com variância $\text{Var}\left(\hat{\beta}\right) = \sigma^2 \left(\mathbf{X}^\top \mathbf{X}\right)^{-1}$ (Verifique!).
2. Os elementos de $\hat{\beta}$ são combinações lineares dos y_i's, isto é, $\hat{\beta} = \sum_{i=1}^{n} d_{ij} y_i$. Sendo d_{ij} elementos da matriz $\left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top$.
3. Com base nos resultados anteriores, segue que, $\hat{\beta} \sim \mathcal{NM}_P\left(\beta, \sigma^2 \left(\mathbf{X}^\top \mathbf{X}\right)^{-1}\right)$, e $\mathbf{y} \sim \mathcal{NM}_n\left(\mathbf{X}\beta, \sigma^2 \mathbf{I}_n\right)$.

4. Dada uma combinação linear dos parâmetros $\mathbf{c}^\top\beta = \sum_{j=1}^{p} c_j\beta_j$, com $\mathbf{c}^\top = (c_1, \cdots, c_p)^\top$, o estimador dos MQ's não viesado e de variância mínima é dado por $\mathbf{c}^\top\hat{\beta}$. Tal que,

$$\mathbb{E}\left(\mathbf{c}^\top\hat{\beta}\right) = \mathbf{c}^\top\beta \text{ e } \mathrm{Var}\left(\mathbf{c}^\top\hat{\beta}\right) = \mathbf{c}^\top\mathrm{Var}\left(\hat{\beta}\right)\mathbf{c} = \sigma^2\mathbf{c}^\top\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{c}.$$

Assim, seque imediatamente que,
$\mathbf{c}^\top\hat{\beta} \sim \mathcal{N}\left(\mathbf{c}^\top\beta, \sigma^2\mathbf{c}^\top\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{c}\right)$.

Inferência sobre os $\hat{\beta}$'s

- Seguindo a mesma ideia apresentada no MRL simples, a questão lógica que pode ser feita agora é: `entre os coeficientes de regressão`, quais deles são significativos para descrever as variações de $\mathbf{y}$?
- As hipóteses para testar individualmente a significância de qualquer coeficiente de regressão, digamos β_j são:

$$H_0 : \beta_j = 0 \text{ vs } H_0 : \beta_j \neq 0,$$

caso a H_0 não seja rejeitada, isso significa que o fator x_j não é significativo, e portanto, pode ser retirado das análises.

Inferência sobre os $\hat{\beta}$'s

- A estatística de teste é:

$$t_{obs} = \frac{\hat{\beta}_j}{\sqrt{\text{Var}\left(\hat{\beta}_j\right)}} = \frac{\hat{\beta}_j}{\hat{\sigma}\sqrt{c_{jj}}} \sim t_{n-p}.$$

onde c_{jj} é o j-ésimo elemento da diagonal da matriz $\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}$.

- É importante lembrar que, $\text{Var}\left(\hat{\beta}\right) = \sigma^2\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}$ e, portanto, a variância de $\hat{\beta}_j$ é dada por:

$$\text{Var}\left(\hat{\beta}_j\right) = \sigma^2 c_{jj}.$$

- A H_0 será rejeitada sempre que $|t_{obs}| > t_{(\alpha/2, n-p)}$.

Inferência sobre os $\hat{\beta}$'s

- Assumindo um nível de confiança $100\,(1-\alpha)$, o intervalo de confiança para o j-ésimo coeficiente de regressão ($j = 0, \cdots, p$) é dado por:

$$\hat{\beta}_j - t_{\left(\frac{\alpha}{2}, n-p\right)} \sqrt{\hat{\sigma}^2 c_{jj}} \leq \beta_j \leq \hat{\beta}_j + t_{\left(\frac{\alpha}{2}, n-p\right)} \sqrt{\hat{\sigma}^2 c_{jj}},$$

  sendo $\hat{\beta}_j$ a estimativa pontual de β_j, e $t_{\left(\frac{\alpha}{2}, n-p\right)}$ quantis da distribuição t-Student com $n-p$ gruas de liberdade.

Obs.: podemos igualmente construir o intervalo de confiança para o parâmetro $\mu_{y|x_0} = \mathbb{E}\,(Y|X = x_0)$.

Análise de Variância (ANOVA)

- Note que até o momento não foi necessária nenhuma hipótese quanto ao ajuste do modelo.
- Para questões de inferência, vamos supor que $\boldsymbol{\epsilon} \sim \mathcal{N}\left(\mathbf{0}, \sigma^2 \mathbf{I}_n\right)$. Sendo $\sigma^2 \mathbf{I}_n$ matriz de variância e covariância.
- Conforme apresentado no caso da regressão linear simples, os resíduos serão obtidos fazendo a seguinte diferença:

$$\underset{n\times 1}{\mathbf{e}} = \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{H}\mathbf{y} = \left(\mathbf{I} - \mathbf{H}\right)\mathbf{y}.$$

- Pode-se mostrar facilmente que $\mathbf{I}_n - \mathbf{H}$ é uma matriz `simétrica` e `idempotente`.
- Usando a mesma ideia do modelo de regressão linear simples, pode-se mostrar facilmente que:

$$\begin{aligned} \mathrm{Var}(\mathbf{e}) &= \mathrm{Cov}\left[(\mathbf{I}_n - \mathbf{H})\mathbf{Y}, (\mathbf{I}_n - \mathbf{H})\mathbf{Y}\right] \\ &= (\mathbf{I}_n - \mathbf{H})\,\mathrm{Var}(\mathbf{Y})\,(\mathbf{I}_n - \mathbf{H})^\top \\ &= \sigma^2(\mathbf{I}_n - \mathbf{H}). \end{aligned}$$

- Como σ^2 é desconhecido, segue que, $\mathrm{Var}(\mathbf{e}) = QME\,(\mathbf{I}_n - \mathbf{H})$. Onde,

$$QME = \frac{SQE(\beta)}{n-p} = \frac{\mathbf{y}^\top(\mathbf{I}_n - \mathbf{H})\mathbf{y}}{n-p},$$

onde $n - p = tr(\mathbf{I}_n - \mathbf{H}) = tr(\mathbf{I}_n) - tr(\mathbf{H})$.

ANÁLISE DE VARIÂNCIA

- Para construir uma tabela de ANOVA, a soma dos quadrados dos totais é dada por:

$$
\begin{aligned}
SQT &= \sum_{i=1}^{n} (y_i - \bar{y})^2 \text{ , em termo matricial fica} \\
&= \left(\mathbf{y} - \frac{1}{n}\mathbf{1}\mathbf{1}^\top\mathbf{y}\right)^\top \left(\mathbf{y} - \frac{1}{n}\mathbf{1}\mathbf{1}^\top\mathbf{y}\right) \\
&= \mathbf{y}^\top \left(\mathbf{I}_n - \frac{1}{n}\mathbf{J}\right)\mathbf{y}
\end{aligned}
$$

$$
SQReg = SQT - SQE = \mathbf{y}^\top \left(\mathbf{H} - \frac{1}{n}\mathbf{J}\right)\mathbf{y}.
$$

Regressão Linear Múltipla: ANOVA

- Cada soma dos quadrados apresentada pode ser vista da forma $\mathbf{y}^\top A\mathbf{y}$, onde A tem a seguinte formulação: $(i)\, \mathbf{I} - \frac{1}{n}\mathbf{J}$; $(ii)\, \mathbf{I} - \mathbf{H}$; $(iii)\, \mathbf{H} - \frac{1}{n}\mathbf{J}$.
- As matrizes dos casos (i)-(iii) são simétricas e idempotentes. A prova é trivial, e segue o fato de que $\mathbf{I}$, $\frac{1}{n}\mathbf{J}$ e $\mathbf{H}$ serem simétricas e idempotentes.
- De igual forma, a matriz $\mathbf{H} - \frac{1}{n}\mathbf{J}$ é simétrica e idempotente. A prova deste fato resume-se em mostrar que $\mathbf{HJ} = \mathbf{J}$ que é simétrica, mas não é idempotente.

Regressão Linear Múltipla: ANOVA

- Usando o fato de que $\mathbb{E}\left(\mathbf{Y}^\top A\mathbf{Y}\right) = \sigma^2 tr\,(A) + \boldsymbol{\mu}^\top A\boldsymbol{\mu}$, podemos mostrar que (verifique!):

  1. $\mathbb{E}\,(SQT) = (n-1)\sigma^2 + (\mathbf{X}\beta)^\top \left(\mathbf{I} - \frac{1}{n}\mathbf{J}\right)(\mathbf{X}\beta)$,
  2. $\mathbb{E}\,(SQE) = (n-p)\sigma^2 + \beta^\top\mathbf{X}^\top\mathbf{X}\beta - (\mathbf{X}\beta)^\top\mathbf{H}\mathbf{X}\beta = (n-p)\sigma^2$,
  3. $\mathbb{E}\,(SReg) = (p-1)\sigma^2 + \beta^\top\mathbf{X}_1^\top\mathbf{X}_1\beta$, sendo $\mathbf{X}_1$ a matriz experimental excluindo a coluna de 1's.

- Assim, o estimador não tendencioso para a variância dos erros é,

$$QME = \frac{\mathbb{E}\,(SQE)}{n-p} = \sigma^2.$$

Regressão Linear Múltipla: ANOVA

- O teste da análise de variância para a regressão linear múltipla tem por objetivo, testar as seguintes hipóteses:

$$
\begin{aligned}
H_0 &: \beta_1 = \beta_2 = \cdots = \beta_p = 0 \\
H_1 &: \exists \beta_j \neq 0 \ \text{ para } \ j = 1, \cdots, p
\end{aligned}
$$

Tabela: Análise de variância para a regressão linear múltipla.

F.V	SQ	G.L	QM	F
Regressão	$SQReg$	$p-1$	$\frac{SQReg}{p-1}$	$F_{cal} = \frac{QMReg}{QME} \overset{sob\ H_0}{\sim} F_{(\alpha,p-1,n-p)}$
Erro	SQE	$n-p$	$\frac{SQE}{n-p}$	
Total	SQT	$n-1$		

- Assim, a H_0 será rejeitada à favor da H_1 se $F_{cal} > F_{(\alpha,p-1,n-p)}$.

Exemplo: Aplicação do MRLM no R

Seleção de Modelos

- Existem duas razões do porquê as vezes não nos satisfazemos com o estimador de mínimos quadrados.

  1. `Precisão da Predição`: Os estimadores de mínimos quadrados normalmente são não viciados mas possuem grande variabilidade.
  2. `Interpretação`: Com um grande número de preditores, gostaríamos de escolher um subconjunto que exiba os fatores principais. Para entender o problema de forma geral estamos dispostos a sacrificar pequenos detalhes.

Seleção de Modelos

- Iremos discutir 3 métodos de seleção de modelos:
  - Forward Selection
  - Bacward Selection
  - Stepwise Selection

Akaike Information Criterion

1. A “Akaike information criterion” (AIC) é uma medida de qualidade ajuste para modelos estatísticos.
2. O AIC é baseado no conceito de entropia, que tenta medir a perda de informação dado um modelo para descrever a realidade.
3. Pode se dizer que o AIC é uma troca entre o viés e a variabilidade do modelo construído, ou de forma mais geral, uma troca entre a complexidade e a precisão do modelo.

$$AIC = 2p - 2\log L(\beta),$$

   onde $L(\beta)$ é a função de verossimilhança e p é o número de parâmetros no modelo.
4. Portanto, o AIC penaliza pela complexidade do modelo (p) e mede a precisão através da função de verossimilhança $L(\beta)$.

Forward Selection

1. Determine todos os fatores que serão usados, o modelo completo.
2. A primeira variável que entra no modelo é aquela no qual o AIC foi minimizado. Adiciona a variável caso o AIC seja inferior ao AIC do modelo com somente o intercepto.
3. Encontre os AIC referentes a adição de uma variável no presente modelo.
4. Se tiver algum fator que diminua o AIC, o inclua e retorne para 3. Caso contrário pare.

Backward Selection

1. Determine o modelo completo.
2. Encontre o AIC referentes a remoção de uma variável por vez no presente modelo.
3. Retire do modelo o fator no qual a remoção diminua mais o AIC e retorne para 2. Caso contrário pare.

Stepwise Selection

- O método stepwise é uma mistura entre os métodos `forward` e o `backward`. Basicamente o método permite que se adicione e/ou retire fatores passo a passo.

1. Defina o modelo completo
2. Determine se a remoção de algum fator reduz o AIC
3. Encontre o AIC para a remoção das variáveis presentes e o AIC para a adição das variáveis que já foram removidos.
4. Inclua ou retire o fator que reduz o AIC e volte para 3. Caso contrário pare.

- Esses modelos são gulosos e podem parecer sub-ótimos se comparado ao melhor subconjunto.
- Porém existem razões para escolher esses métodos
  1. Não é preciso determinar p.
  2. É computacionalmente eficiente se comparado ao método do subconjunto.
  3. Estatisticamente os métodos gulosos fazem uma busca mais restrita nos modelos; incluindo um viés maior mas reduzindo a variabilidade.

O critério BIC-Bayesian Information Criterion, é dado pela seguinte expressão:

$$BIC = p \log n - 2 \log L(\beta).$$

Obs.: O critério BIC penaliza mais fortemente modelos com um maior número de parâmetros do que o AIC tendendo, dessa forma, a selecionar modelos com um menor número de parâmetros.

Multicolinearidade

O que é colinearidade?

- Suponha que temos que a matriz experimental $\mathbf{X}$. A colinearidade é um fenômeno que ocorre quando alguns preditores de $\mathbf{X}$ são combinações lineares dos demais, i.e., quando a matriz $\mathbf{X}^\top\mathbf{X}$ é singular.
- Quando $\mathbf{X}^\top\mathbf{X}$ é singular, ou próxima de ser, o estimador dos mínimos quadrados para $\hat{\beta}$ não será único $(r(\mathbf{X}) < p)$. Neste caso, podemos dizer que há colinearidade (ou multicolinearidade).
- A presença da colinearidade causa sérios problemas com a estimação de $\hat{\beta}$ e outras quantidades associadas, como a interpretação.

- Suponha que temos dois preditores, x_1 e x_2. Então, x_1 e x_2 são ditos serem colineares se existe constantes c_0, c_1 e c_2 tal que:

$$c_1x_1 + c_2x_2 = c_0,$$

e aproximadamente colineares se

$$c_1x_1 + c_2x_2 \approx c_0.$$

- Essa ideia pode ser estendida para o caso em que temos $x_1, \cdots, x_k$ preditores. Logo, $x_1, \cdots, x_k$ são ditos serem colineares se temos,

$$c_1x_1 + \cdots + c_kx_k = c_0,$$

e aproximadamente colineares se

$$c_1x_1 + \cdots + c_kx_k \approx c_0.$$

Regressão Linear Múltipla

- No contexto do modelo de regressão linear $\mathbf{y} = \mathbf{X}\beta + \epsilon$. Então, o modelo é dito ser afetado pelo problema da colinearidade ou multicolinearidade, se pelo menos uma coluna de $\mathbf{X}$ é combinação linear de outras.
- Matematicamente temos que, a colinearidade ocorre se
$$|\mathbf{X}^\top \mathbf{X}| = 0.$$
- Porém, na prática tratamos o problema de multicolinearidade quando
$$|\mathbf{X}^\top \mathbf{X}| \approx 0.$$

Qual o problema de multicolinearidade?

- Vamos pensar no modelo simples com 2 covariáveis, $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$, com $\mathbb{E}(\epsilon) = 0$ e $\text{Var}(\epsilon) = \sigma^2$.
- Seja r_{12} a correlação amostral de Pearson x_1 e x_2, e seja $S^2_{x_j}$ a variância amostral de x_j. Portanto, se $\hat{\beta}_j$ é o estimador de mínimos quadrados de β_j. Pode se mostrar que:

$$\text{Var}\left(\hat{\beta}_j\right) = \sigma^2 \left(\frac{1}{1 - r^2_{12}}\right)\left(\frac{1}{(n-1)S^2_{x_j}}\right), \text{ para } j = 1, 2$$

- Assim, $\text{Var}\left(\hat{\beta}_j\right) \to \infty$ se $r^2_{12} \to 1$, e $\text{Var}\left(\hat{\beta}_j\right)$ é minimizada quando $r^2_{12} = 0$.

Regressão Linear Múltipla

- Lembre-se que o estimador de mínimos quadrados é da forma

$$\hat{\beta} = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{y},$$

com variância $\text{Var}\left(\hat{\beta}\right) = \sigma^2 \left(\mathbf{X}^\top \mathbf{X}\right)^{-1}$.

- A multicolinearidade gera uma matriz $\mathbf{X}^\top \mathbf{X}$ aproximadamente singular, o que pode causar estimativas pouco precisas de $\hat{\beta}$ com uma variância alta.

- Por conta disso, o teste t-Student para testar a significância de β_j, digamos $H_0 : \beta_j = 0$ pode ser enganoso.

Regressão Linear Múltipla

Sinais da presença de multicolinearidade:

- A correlação entre pares de preditores é surpreendentemente grande (`use a matriz de correlações`)
- Se alguns dos coeficientes forem estimados com o sinal trocado, baseando em algum conhecimento prévio
- Preditores conhecidos de serem importantes (através de conhecimento a priori) possuem uma estatística t pequena
- Remoção de uma linha ou coluna da matriz $\mathbf{X}$ produz uma mudança surpreendente no modelo ajustado.
- Examine os autovalores da matriz $\mathbf{X}^\top\mathbf{X}$, e denote por λ_1 o `maior autovalor` entre os demais, em ordem crescente. Em geral, autovalores menores indicam a presença de multicolinearidade (`facilmente obtido no R`).

- O fator de inflação da variância (VIF) para o j-ésimo preditor x_j é definido como

$$\text{VIF}_j = \frac{1}{1 - R_j^2},$$

onde R_j^2 é o coeficiente de determinação múltiplo, R^2 quando x_j é regressado nos outros $k-1$ regressores, para $j = 1, 2, \cdots, k$.

- Regra básica:

  VIF ≥ 10 é uma indicação de multicolinearidade severa

  VIF > 5 é uma indicação de multicolinearidade moderada, e deve ser investigada.

Remédios para multicolinearidade:

Essas técnica é particularmente útil para tirar a correlação entre os preditores e o intercepto, e também é muito útil em modelos que contém interação de alta ordem (interações maiores do que quadrática).

- Centrar e/ou escalonar e/ou padronizar empiricamente as variáveis preditoras. Ou seja, se $x_{1j}, x_{2j}, \cdots, x_{nj}$ são valores observados de x_j para cada indivíduo, defina $\bar{x}_j$ e S_{x_j} como sendo a média e desvio-padrão amostral dos x_{ij}'s.
  - Centrar: $x_{ij} - \bar{x}_j$
  - Escalonar: x_{ij}/S_{x_j}
  - Padronizar: $(x_{ij} - \bar{x}_j)/S_{x_j}$.

Regressão Linear Múltipla

- Remover um ou vários preditores (normalmente aquelas que possuem alto valores de VIF) do modelo
- Combinar alguns preditores lineares
- Quebrar o padrão de multicolinearidade adicionando novas observações, i.e., adicionando mais dados
- Usando algum tipo de regressão penalizada:
    - ridge regression
    - lasso regression

Aplicação do MRLM no R

Análise de Resíduos e Medidas Corretivas

Seguindo a mesma ideia apresentada no modelo de regressão linear simples (MRLS), a análise de resíduos objetiva:

- Avaliar o quão o modelo proposto se ajusta aos dados
- Verificar se existe(m) ou não valor(es) estimado(s) fora do padrão esperado (valores de alavancagem ou outliers)
- Verificar a presença ou não de multi(colinearidade) entre as variáveis explicativas
- Verificar a validade da suposição de homoscedasticidade, entre outros aspectos.

- No que diz respeito a análise dos resíduos, mas concretamente aos pontos influentes (`pontos de alavancagem`), considere a seguinte matriz experimental $\mathbf{X}_{n\times p}$. Assim, é possível mostrar que, os elementos da matriz projeção

$$\mathbf{H} = \mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top,$$

são tais que,

$$\begin{aligned} h_{ii} &= \frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}}, \quad \forall i = j \\ h_{ij} &= \frac{1}{n} + \frac{(x_i - \bar{x})(x_j - \bar{x})}{S_{xx}}, \quad \forall i \neq j \end{aligned}$$

Nota: Verifique assumindo que $\mathbf{X}$ é uma matriz $n \times 2$ quaisquer.

- Lembrando, temos que $e_i \sim \mathcal{N}\left(0, \sigma^2\left(1 - h_{ii}\right)\right)$. Assim,

$$\frac{e_i}{ep\left(e_i\right)} = \frac{e_i}{\sigma\sqrt{1 - h_{ii}}} \sim \mathcal{N}\left(0, 1\right).$$

Resíduo Interno Studentizado: O resíduo interno studentizado, é definido como:

$$r_i = \frac{e_i}{ep\left(e_i\right)} = \frac{e_i}{\hat{\sigma}\sqrt{1 - h_{ii}}} = \frac{e_i}{\sqrt{QME\left(1 - h_{ii}\right)}}.$$

- Apesar de assumirmos que ϵ_i devem ser independentes, eles apresentam uma pequena correlação, tal que, $\text{Cov}\left(r_i, r_j\right) = -h_{ij}\sigma^2 \ (i \neq j)$ e portanto,

$$Corr\left(r_i, r_j\right) = \frac{-h_{ij}}{\sqrt{\left(1 - h_{ii}\right)\left(1 - h_{jj}\right)}} \approx 0 \ (i \neq j).$$

- Sob a suposição de linearidade e normalidade, (Cook and Weisberg,1982) mostraram que,

$$\frac{r_i^2}{n-p} \sim \text{Beta}\left(\frac{1}{2}, \frac{n-p}{2}\right),$$

  Assim, $|r_i| \leq t_{(\alpha/2, n-p)}$

- Um gráfico de r_i versus $\hat{Y}_i$ ainda é útil para detectar discrepâncias das hipóteses do modelo. Espera-se ver um padrão aleatório e $|r_i| \leq 2$.

Regressão Linear Múltipla

- DFFITS: É uma medida que avalia a alteração provocada no valor ajustado $\hat{y}$, pela retirada da s-ésima observação no banco de dados. Por definição, temos que,

$$\text{DFFITS}_i = \frac{\hat{y}_i - \hat{y}_{i(s)}}{\sqrt{QME_{(s)} h_{ii}}},$$

  onde $\hat{y}_{i(s)}$ denota o valor predito para s-ésima observação obtida ao deletar a s-ésima observação para ajustar o modelo de regressão e $QME_{(s)}$ é o QME obtido ao deletar a s-ésima observação do modelo de regressão.

- Regra básica:
  Se $|\text{DFFITS}_i| \geq 1$ (para bancos de dados pequenos ou médios)
  Se $|\text{DFFITS}_i| \geq 2\sqrt{\frac{p}{n}}$, com $p = k + 1$ (para bancos de dados grandes), a i-ésima observação é considerada influente.

Influência sobre todos os valores ajustados - Distância de Cook:

- É uma medida de afastamento do vetor de estimativas provocada pela retirada da i-ésima observação.
- Sua expressão é muito semelhante com aquela usada no caso da influência nos valores ajustados (DFFITS) mas, que usa como estimativa de variância residual, aquela obtida com todas as n observações.
- Para começar, considere a seguinte formulação matemática. Seja $\hat{\beta} = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{y}$. Considere também, $\hat{\beta}_{(i)} = \left(\mathbf{X}_{(i)}^\top \mathbf{X}_{(i)}\right)^{-1} \mathbf{X}_{(i)}^\top \mathbf{y}_{(i)}$, onde $\mathbf{X}_{(i)}$ é a matriz $\mathbf{X}$ com a i-ésima linha deletada, e $\mathbf{y}_{(i)}$ é o vetor $\mathbf{y}$ com a i-ésima observação deletada.

- O objetivo é comparar $\hat{\beta}$ com $\hat{\beta}_{(i)}$ para determinar se a i-ésima observação é influente, e determinar sua magnitude.
- Para tal, Cook (1972) propôs que a influência da i-ésima observação deve ser medida em função da distância quadrática,

$$D_i = \frac{\left(\hat{\mathbf{y}} - \hat{\mathbf{y}}_{(i)}\right)^\top \left(\hat{\mathbf{y}} - \hat{\mathbf{y}}_{(i)}\right)}{p \times QME},$$

  onde $\hat{\mathbf{y}} = \mathbf{X}\hat{\beta}$, $\hat{\mathbf{y}}_{(i)} = \mathbf{X}\hat{\beta}_{(i)}$, p é o número de coeficientes de regressão, e QME é a estimativa da variância residual para todas as observações.
- Como $\hat{\mathbf{y}} - \hat{\mathbf{y}}_{(i)} = \mathbf{X}\left(\hat{\beta} - \hat{\beta}_{(i)}\right)$. Assim, a distância de Cook pode ser reescrita da seguinte forma:

$$D_i = \frac{\left(\hat{\beta} - \hat{\beta}_{(i)}\right)^\top \mathbf{X}^\top \mathbf{X} \left(\hat{\beta} - \hat{\beta}_{(i)}\right)}{p \times QME},$$

Regressão Linear Múltipla

- Outra forma de calcular a distância de Cook seria um desenvolvimento das expressões anteriores:

$$D_i = \left(\frac{e_i}{\sqrt{(1-h_{ii})QME}} \right)^2 \left(\frac{h_{ii}}{1-h_{ii}} \right) . \frac{1}{p} = \frac{r_i^2}{p} . \frac{h_{ii}}{1-h_{ii}},$$

  sendo r_i o resíduo internamente estudentizado.

- Quanto a magnitude de D_i, podemos concluir que: observações com grande valor de D_i são aquelas nos quais, a exclusão irá resultar em grandes mudanças na análise.

- Regra básica: Se $D_i \geq 0,80$, a i-ésima observação será considerada influente.

- DFBETAS: Mede a alteração no valor $\hat{\beta}_j$ ($j = 0, 1, \cdots, k$), quando a i-ésima observação é retirada das análises. A medida é dada pela diferença nas estimativas entre o coeficiente de regressão $\hat{\beta}_j$ (baseado em todas as n observações), e o coeficiente de regressão $\hat{\beta}_{j(i)}$ obtido quando o modelo é ajustado sem a observação i.
- A medida DFBETAS é definida como,

$$\text{DFBETAS}_{(i)} = \frac{\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)}}{\sqrt{QME_{(i)} c_{jj}}},$$

onde c_{jj} é o j-ésimo elemento da diagonal da matriz $\left(\mathbf{X}^\top \mathbf{X}\right)^{-1}$.
- É importante lembrar que, $\text{Var}\left(\hat{\boldsymbol{\beta}}\right) = \sigma^2 \left(\mathbf{X}^\top \mathbf{X}\right)^{-1}$ e, portanto, a variância de $\hat{\beta}_j$ é dada por:

$$\text{Var}\left(\hat{\beta}_j\right) = \sigma^2 c_{jj}.$$

Regressão Linear Múltipla

- Regra básica:
  A i-ésima observação é considerada ser `influente` se

$$|\text{DFBETAS}_{j(i)}| \geq \begin{cases} 1 & \text{para bancos pequenos ou médios} \\ \frac{2}{\sqrt{n}} & \text{para bancos de dados grandes.} \end{cases}$$

Regressão Linear Múltipla

- COVRATIO: É uma estatÍstica que mede a mudança no determinante na matriz de covariâcia das estimativas deletando a i- ésima observação.

$$\text{COVRATIO} = \frac{\det\left(QME_{(i)}\left(\mathbf{X}_{(i)}^{\top}\mathbf{X}_{(i)}\right)^{-1}\right)}{\det\left(QME\left(\mathbf{X}^{\top}\mathbf{X}\right)^{-1}\right)},$$

- onde $\mathbf{X}_{(i)}$ é a matriz $\mathbf{X}$ com a i-ésima linha deletada, e $QME_{(i)}$, é o QME obtido quando a i-ésima observação é omitida para ajustar o modelo de regressão.
- A estatística COVRATIO pode ser simplificada para:

$$\text{COVRATIO} = \left[\left(\frac{n-p-1}{n-p} + \frac{t_i}{n-p}\right)^k (1-h_{ii})\right]^{-1}.$$

- Na expressão anterior, t_i é o resíduo externo estudentizado (ou resíduo deletado). Tal que,

$$t_i = \frac{e_i}{\hat{\sigma}_{(i)}\sqrt{1 - h_{ii}}} = \frac{e_i}{\sqrt{QME_{(i)}\left(1 - h_{ii}\right)}}.$$

Tal que, $t \sim t_{n-p}$.

- Para o ponto de corte, Belsley, Kuh, and Welsch sugerem que observações com

$$|\text{COVRATIO} - 1| \geq \frac{3p}{n},$$

onde $p = k + 1$ o número de coeficientes no modelo.

Observação: No pacote estatístico R, a i-ésima observação é considerada influente se:

- $|DFBetaS_{j(i)}| > 1$ para n (pequeno e moderado), e $|DFBetaS_{j(i)}| > \frac{2}{\sqrt{n}}$, caso contrário.
- $|DFFITS_{j(i)}| > 1$ para n (pequeno e moderado), e $|DFFITS_{j(i)}| > 3\sqrt{\frac{p}{\sqrt{(n-p)}}}$, caso contrário.
- COVRATIO, se: $|1 - COVRATIO| > 3p/(n-p)$.
- Distância de Cook, se $D_{(i)} > F_{(0,5;p;n-p)}$.
- Pontos de Alavancagem, se $h_{ii} > 3p/n$.

Funções Estimáveis

- Conforme apresentado anteriormente, para o caso do MRLM, $\mathbf{y} = \mathbf{X}\beta + \epsilon$, a solução da equação $\hat{\beta} = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{y}$ é única quando $r(\mathbf{X}) = p$.

- Portando, quando $r(\mathbf{X}) < p$, existem infinitas soluções para $\hat{\beta} = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{y}$, neste caso, devemos nos restringir apenas algumas funções de β.

- Tais funções/soluções para as quais devemos nos restringir (ou escolher), são chamadas de estimáveis.

Definição 3: Seja $\mathbf{y}$ um vetor observado de $\mathbf{y}$, com distribuição conhecida. Uma função linear paramétrica $\mathbf{c}^\top\beta$ é dita ser estimável de β, se existir um vetor n-dimensional $\mathbf{t} = (t_1, t_2, \cdots, t_n)^\top$, tal que, a esperança da combinação linear $\mathbf{t}^\top\mathbf{y} = \mathbf{c}^\top\beta$ é,

$$\mathbb{E}\left(\mathbf{t}^\top\mathbf{y}\right) = \mathbf{t}^\top\mathbb{E}(\mathbf{y}) = \mathbf{c}^\top\beta, \text{ com } \beta \in \mathbb{R}^p, \text{ e}$$

$\mathbf{y} = \mathbf{X}\beta$, e $\mathbf{c} = (c_1, \cdots, c_p)^\top$.

- De forma geral, uma função $g(\beta)$ é dita ser estimável se existe uma função $h(\mathbf{y})$ tal que, $\mathbb{E}(h(\mathbf{y})) = g(\beta)$, para qualquer $\beta \in \mathbb{R}^p$, sendo então, $h(\mathbf{y}) = \mathbf{t}^\top\mathbf{y}$ e $g(\beta) = \mathbf{c}^\top\beta$.

Exemplo: Como $\mathbf{y} = \mathbf{X}\beta + \epsilon$ então, segue que:

$$\mathbb{E}\left(\mathbf{t}^\top \mathbf{y}\right) = \mathbf{t}^\top \mathbb{E}(\mathbf{y}) = \mathbf{t}^\top \mathbf{X}\beta = \mathbf{c}^\top \beta \Longrightarrow \mathbf{t}^\top \mathbf{X} = \mathbf{c}^\top.$$

- Se $r(\mathbf{X}) = p$ então, $\mathbf{c}^\top \beta$ é estimável, i.e., $\hat{\beta}$ tem solução única.
- Quando $r(\mathbf{X}) < p$, algumas funções são estimáveis.
- Para garantir que uma função $\mathbf{c}^\top \beta$ é estimável, devemos verificar se $\mathbf{t}^\top \mathbf{X} = \mathbf{c}^\top$, para algum vetor $\mathbf{t}$.
- O valor esperado de qualquer observação é estimável.
- Qualquer combinação linear de funções estimáveis é estimável.
- Dada uma função estimável $\mathbf{c}^\top \beta$, a quantidade $\mathbf{c}^\top \tilde{\beta}$ é invariante a escolha de $\tilde{\beta}$.

Teorema (Gauss-Markov)

Seja $\mathbf{c}^\top\beta$ uma função estimável de β, e seja $\tilde{\beta}$ uma solução qualquer para as equações normais.

(i) O teorema de Gauss-Markov afirma que, $\mathbf{c}^\top\tilde{\beta}$ é o melhor estimador não-viesado de $\mathbf{c}^\top\beta$ com variância,

$$\text{Var}\left(\mathbf{c}^\top\tilde{\beta}\right) = \mathbf{c}^\top\text{Var}\left(\tilde{\beta}\right)\mathbf{c} = \sigma^2\mathbf{c}^\top\mathbf{G}\mathbf{c}$$

(ii) Se $\tilde{\beta} = \hat{\beta}$, então,

$$\text{Var}\left(\mathbf{c}^\top\hat{\beta}\right) = \mathbf{c}^\top\text{Var}\left(\hat{\beta}\right)\mathbf{c} = \sigma^2\mathbf{c}^\top\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{c}.$$