Topico 8: Regressão Linear

Ben Dêivide

6 de outubro de 2021

Um problema comum na estatística é tentar verificar a associação entre duas variáveis X e Y. Após isto, estamos interessados na forma como ocorre esta associação. Procuramos então uma relação funcional entre essas variáveis, tal que

$$(X, Y) \to Y \simeq f(X). \tag{1}$$

A relação em (1) não é perfeita. Os pontos não se situam perfeitamente sobre a função que relaciona as duas variáveis. Mesmo se existisse uma relação exata entre as variáveis como temperatura e pressão, flutuações em torno da curva aparecerão devido a erros de medidas. Frequentemente, o tipo de curva a ser ajustada é sugerido por evidência empírica ou por argumentos teóricos. O modelo a ser adotado depende de vários fatores, por exemplo, natureza das variáveis, relação linear ou não, homogeneidade de variâncias ou não, tipos de erros, independência dos erros etc.

Diremos que a variável Y é aleatória sendo chamada de variável resposta ou dependente. A variável X é fixa, sendo chamada de variável explicativa, regressora ou independente. Os valores da variável X são selecionados pelo pesquisador, não havendo variação aleatória associada. A seleção dos Xs pode envolver um conjunto específico de valores ou valores que estão simplesmente dentro de uma amplitude de variação.

Portanto, na estatística é a Teoria de Regressão que tenta encontrar a relação funcional em (1), que chamaremos modelo de regressão. Essa teoria teve origem no século XIX com Francis Galton. Em um de seus trabalhos, Galton estudou a relação da altura dos pais (X_i) e dos filhos (Y_i), procurando saber como a altura dos pais influenciava a altura dos filhos. Ele notou que se o pai fosse muito alto ou muito baixo, o filho teria uma altura tendendo à média, isto é, existia uma tendências dos dados regredirem à média. Por isso do nome "regressão". Os objetivos desse modelo são:

- **Predição:** Espera-se que a maior variação ocorrida na variável Y seja explicada por X. Assim, por meio do modelo de regressão podemos obter valores de Y correspondente aos valores de X que não estavam entre os dados. Em geral, sempre se seleciona valores de X dentro do intervalo das observações de X utilizadas para ajustar o modelo. Muito embora possa ser possível utilizar valores de X fora desse intervalo, dizemos pois que ocorreu uma extrapolação nos dados. A extrapolação deve ser utilizada com muito cuidado, pois o modelo adotado não garante um valor correto para Y fora do intervalo estudado para X. A predição, talvez, seja o uso mais comum para o estudo dos modelos de regressão.

- **Seleção de variáveis:** Muitas vezes a variável Y pode ser explicada por mais de uma variável X, e a relação funcional (1) pode ser expressa por: $Y \simeq f(X_1, X_2, \ldots, X_p)$.

Ou seja, Y é explicada por p variáveis explicativas. Entretanto, nem sempre essas p variáveis apresentam tanta influência na explicação dos valores para Y. Dessa forma, a análise de regressão poderá auxiliar no processo de seleção das p variáveis X, eliminando aquelas que não sejam importantes.

- **Estimação de parâmetros:** Em (1), f é dependente de parâmetros, e estes são desconhecidos. Então dada uma amostra e um modelo, a estimação pontual tenta encontrar, por meio de algum método, estimativas que possam representar os parâmetros desconhecidos. Após substituí-los, temos um modelo ajustado.

- **Inferência:** Após realizarmos a estimação pontual, as estimativas mudam para cada amostra realizada. Por isso, precisamos de uma confiança sobre estas estimativas. Daí fazemos a inferência sobre os parâmetros, por meio dessas estimativas, realizando testes de hipóteses e intervalos de confiança.

Definição 1 (Modelo de Regressão)**.** *Seja uma variável aleatória Y e $X_1, X_2, \ldots, X_p$ um conjunto de variáveis regressoras, então a relação*

$$Y = f(X_1, X_2, \ldots, X_p; \beta_0, \beta_1, \ldots, \beta_p) + \varepsilon, \tag{2}$$

é chamada de Modelo de Regressão, sendo $\beta_j \in \mathbb{R}$, para $j = 0, 1, 2, \ldots, p$, os $p' = p + 1$ parâmetros desconhecidos da função $f(.)$, e ε é o erro aleatório não observável. □

Nesse estudo sempre consideraremos as variáveis regressoras fixas. A notação usada para esse estudo será:

a) a variável aleatória bem como as variáveis fixas serão representadas por letras maiúsculas;

b) A realização dessas variáveis será representado por letra minúscula;

c) os parâmetros dos modelos serão simbolizados por letras gregas e seus estimadores identificados por um acento circunflexo em cima das letras. Por exemplo, seja β um parâmetro, então $\hat{\beta}$ é seu estimador.

Definição 2 (Modelo de Regressão Linear)**.** *Um modelo de regressão, expresso em* (2)*, tal que*

$$\frac{\partial f}{\partial \beta_j} = h(X_1, X_2, \ldots, X_p),$$

para $j = 0, 1, 2, \ldots, p$, e $h(.)$ uma função qualquer que dependa apenas das variáveis regressoras, é chamado de modelo de regressão linear. □

Se pelo menos uma das derivadas parciais $\partial f / \partial \beta_j$ depender de algum dos parâmetros, então $f(.)$ é dita função ou modelo de regressão não-linear. Esse ponto não fará parte desse estudo.

Exemplo 1 (Regressão linear e não-linear)**.** *Exemplos de modelos de regressão linear:*

- $f(X, \beta_0) = \beta_0$, *pois* $\frac{df}{d\beta_0} = 1$;

- $Y = f(X, \beta_0, \beta_1) = \beta_0 + \beta_1 X$, *pois* $\frac{\partial f}{\partial \beta_0} = 1$ *e* $\frac{\partial f}{\partial \beta_1} = X$.

Exemplos de modelo de regressão não-linear:

- $f(X, \beta_0, \beta_1, \beta_2) = \frac{\beta_0 + \beta_1 X}{1 + \beta_2 X}$, *pois* $\frac{\partial f}{\partial \beta_0} = \frac{1}{1+\beta_2}$, $\frac{\partial f}{\partial \beta_1} = \frac{X}{1+\beta_2}$, *e* $\frac{\partial f}{\partial \beta_2} = -\frac{(\beta_0 + \beta_1 X)X}{(1+\beta_2)^2}$;
- $f(X, \beta_0, \beta_1, \beta_2) = \beta_0 + \beta_1 e^{\beta_2 X}$, *pois* $\frac{\partial f}{\partial \beta_0} = 1$, $\frac{\partial f}{\partial \beta_1} = e^{\beta_2 X}$, *e* $\frac{\partial f}{\partial \beta_2} = \beta_1 X e^{\beta_2 X}$.

Definição 3 (Modelo de Regressão Linear Múltipla - MRLM)**.** *Considerando as Definições 1 e 2, um modelo de regressão linear múltipla para n observações independentes de Y associados com os Xs pode ser definido por*

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip} + \varepsilon_i, \tag{3}$$

sendo $i = 1, 2, \ldots, n$, *supondo* $E[\varepsilon_i] = 0$, $var[\varepsilon_i] = \sigma^2$ *e* $cov[\varepsilon_i, \varepsilon_j] = 0$ *para todo* $i \neq j$. □

Esse modelo descreve um hiperplano no espaço p-dimensional de variáveis regressoras. Podemos interpretar β_j, para $j = 1, 2, \ldots, p$, como a mudança esperada em Y_i devido ao aumento de uma unidade em X_j, estando as outras variáveis regressoras constantes. Por consequência de $E[\varepsilon_i] = 0$ e $var[\varepsilon_i] = \sigma^2$, temos que $E[Y_i] = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip}$ e $var[Y_i] = \sigma^2$.

Definição 4 (Modelo de Regressão Linear Simples - MRLS)**.** *Considerando a Definição 3 para* $p = 1$, *o modelo expresso em* (3), *reduz-se a*

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad i = 1, 2, \ldots, n, \tag{4}$$

também conhecido como modelo de regressão linear simples. □

Para a expressão (4) o parâmetro β_0 representa o coeficiente linear e o β_1 o coeficiente angular para o modelo.

Percebe-se que a Definição 4 é um caso particular da Definição 3. Assim, quando falarmos sobre regressão linear, sempre estaremos nos referindo ao caso geral, Definição 3, que englobará todos os outros modelos de regressão linear.

Inicialmente, pensemos num conjunto de pares (X_i, Y_i), $i = 1, 2, \ldots, n$ de duas variáveis e plotemos um gráfico de dispersão para que possamos obter alguma ideia sobre a forma de associação entre X e Y, Figura 1. Esse gráfico também ajuda a detectar pontos discrepantes (*Outliers*). Contudo, o gráfico de dispersão deve ser olhado com muito cuidado, uma vez que este não leva em consideração a correlação entre duas ou mais variáveis regressoras, caso exista.

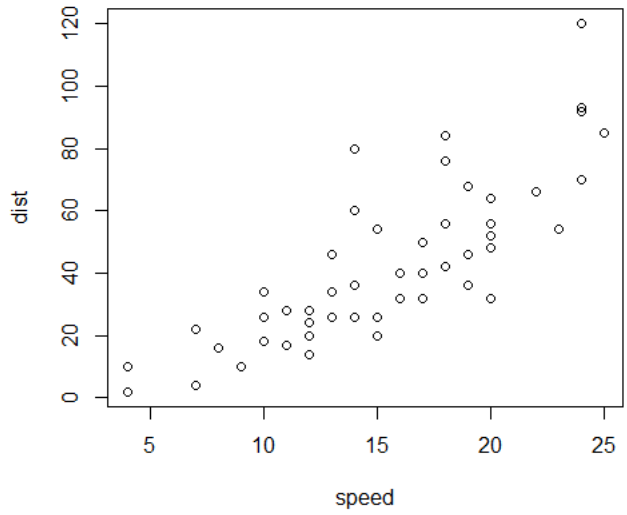

Figura 1: Gráfico de dispersão.

Após averiguar essas informações iniciais, sabemos que mesmo tendo indicativos de um modelo preliminar, a equação (3) constitui uma família de funções de parâmetros $\beta_j s$. Assim, o objetivo é encontrar aquele que melhor descreva o comportamento de Y.

Para isto devemos estimar os parâmetros desconhecidos. Para estimar os $\beta_j s$ em (3), usaremos a notação matricial para facilitar a compreensão. O modelo de regressão em (3) matricialmente é dado por:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\theta} + \boldsymbol{\varepsilon}, \tag{5}$$

em que $\boldsymbol{Y}$ é um vetor de dimensões $n \times 1$ da variável aleatória Y; $\boldsymbol{X}$ é a matriz de dimensões $n \times p'$ conhecida do delineamento, assumindo que $n > p'$ e que $\boldsymbol{X}$ é de posto completo p'; $\boldsymbol{\theta}$ é o vetor de dimensões $p' \times 1$; $\boldsymbol{\varepsilon}$ é o vetor de dimensões $n \times 1$ dos erros aleatórios. Assim, estes podem ser expressos da seguinte forma:

$$\boldsymbol{Y}_{n\times 1} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \boldsymbol{X}_{n\times p'} = \begin{bmatrix} 1 & X_{11} & X_{12} & \dots & X_{1p} \\ 1 & X_{21} & X_{22} & \dots & X_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & X_{n2} & \dots & X_{np} \end{bmatrix}, \boldsymbol{\theta}_{p'\times 1} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} \text{ e}$$

$$\boldsymbol{\varepsilon}_{n\times 1} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}.$$

As pressuposições para esse modelo são as mesmas usadas em (3), isto é:

1. $E[\boldsymbol{\varepsilon}] = \boldsymbol{0}$, sendo $\boldsymbol{0}$ um vetor de zeros de dimensão $n \times 1$, ou equivalentemente $E[\boldsymbol{Y}] = \boldsymbol{X}\boldsymbol{\theta}$;

2. $cov[\boldsymbol{\varepsilon}] = \boldsymbol{I}_n\sigma^2$, sendo $\boldsymbol{I}_n$ uma matriz identidade de dimensão $n \times n$, ou equivalentemente $cov[\boldsymbol{Y}] = \boldsymbol{I}_n\sigma^2$, sendo $0 < \sigma^2 < \infty$.

3. $cov[\varepsilon_i, \varepsilon_j] = 0$ para todo $i \neq j \in \{1, 2, \dots, n\}$, ou equivalentemente, $cov[Y_i, Y_j] = 0$.

Como sabemos o vetor de parâmetros $\boldsymbol{\theta}$ é desconhecido e precisa ser estimado. O número de parâmetros a ser estimado é $p' = p + 1$. Se existirem apenas p' observações, a estimação dos parâmetros reduz-se a um problema matemático de resolução de um sistema de p' equações a p' incógnitas, não deixando nada para ser explicado como variabilidade natural. Deve-se portanto, ter $p' < n$.

Um método muito utilizado para estimar os parâmetros de um modelo de regressão é o método dos mínimos quadrados. Se o modelo adotado é dado por (5), então, o vetor de erros, $\boldsymbol{\epsilon}$, pode ser expresso por:

$$\boldsymbol{\epsilon} = \boldsymbol{Y} - \boldsymbol{X\theta}$$

De uma maneira geral, tem-se que tanto melhor será o modelo quanto menor for o comprimento de $\boldsymbol{\epsilon}$, isto é, o ajuste de mínimos quadrados requer encontrar os valores de $\boldsymbol{\theta}$ que minimizem a soma de quadrados das diferenças $\boldsymbol{Y} - \boldsymbol{X\theta}$, isto é, $\min[(\boldsymbol{Y} - \boldsymbol{X\theta})'(\boldsymbol{Y} - \boldsymbol{X\theta})]$. Assim, apresentamos o seguinte teorema,

Teorema 1 (Estimador de mínimos quadrados de $\boldsymbol{\theta}$)**.** *Se $\boldsymbol{Y} = \boldsymbol{X\theta} + \boldsymbol{\varepsilon}$, em que $\boldsymbol{X}$ é $n \times p'$ de posto $p' < n$, então o valor de $\hat{\boldsymbol{\theta}} = [\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p]'$ que minimiza $\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}$ é igual a*

$$\hat{\boldsymbol{\theta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}. \tag{6}$$

Logo (6) *é o estimador de mínimos quadrados de $\boldsymbol{\theta}$.* □

Demonstração. Podemos escrever $\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}$ como

$$\begin{aligned} \boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} = & \quad (\boldsymbol{Y} - \boldsymbol{X\theta})'(\boldsymbol{Y} - \boldsymbol{X\theta}) \\ = & \quad \boldsymbol{Y}'\boldsymbol{Y} - 2\boldsymbol{\theta}'\boldsymbol{X}'\boldsymbol{Y} + \boldsymbol{\theta}'\boldsymbol{X}'\boldsymbol{X\theta}. \end{aligned}$$

Para encontrarmos $\hat{\boldsymbol{\theta}}$ que minimiza $\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}$, calculamos a diferencial $\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}$ em relação a $\boldsymbol{\theta}$:

$$\frac{\partial \boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}}{\partial \boldsymbol{\theta}} = \quad 0 - 2\boldsymbol{X}'\boldsymbol{Y} + 2\boldsymbol{X}'\boldsymbol{X\theta}.$$

Igualando a zero, obtemos o sistema de equações normais (SEN):

$$\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\theta}} = \boldsymbol{X}'\boldsymbol{Y}. \tag{7}$$

Como $\boldsymbol{X}$ tem posto completo, $\boldsymbol{X}'\boldsymbol{X}$ é não singular e portanto invertível. Assim, a solução (7) é (6). Como

$$\frac{\partial^2 \boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}}{\partial \boldsymbol{\theta}^2} = \quad 2\boldsymbol{X}'\boldsymbol{X} = \boldsymbol{Q},$$

é positiva definida, tal que $\boldsymbol{Z}'\boldsymbol{Q}\boldsymbol{Z} > 0$, sendo $\boldsymbol{Z}$ um vetor $p' \times 1$ para todo $\boldsymbol{Z} \neq \boldsymbol{0}$, em que $\boldsymbol{0}$ é um vetor de zeros $p' \times 1$, logo $\hat{\boldsymbol{\theta}}$ é um estimador de mínimos quadrados para $\boldsymbol{\theta}$. □

Exemplo 2. *Seja um modelo de regressão linear simples do tipo $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$ para $i = 1, 2, \ldots, n$. De modo matricial, temos*

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Os termos matriciais podem ser expressos:

$$\underset{\sim}{X}'\underset{\sim}{Y} = \begin{bmatrix} \sum_{i=1}^n Y_i \\ \sum_{i=1}^n X_i Y_i \end{bmatrix}, \ (\underset{\sim}{X}'\underset{\sim}{X})^{-1} = \frac{1}{n\sum_{i=1}^n X_i^2 - (\sum_{i=1}^n X_i)^2} \begin{bmatrix} \sum_{i=1}^n X_i^2 & -\sum_{i=1}^n X_i \\ -\sum_{i=1}^n X_i & n \end{bmatrix} \Rightarrow$$

$$\underset{\sim}{\hat{\theta}} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} = \frac{1}{n\sum_{i=1}^n X_i^2 - (\sum_{i=1}^n X_i)^2} \begin{bmatrix} \sum_{i=1}^n X_i^2 \sum_{i=1}^n Y_i - \sum_{i=1}^n X_i \sum_{i=1}^n X_i Y_i \\ -\sum_{i=1}^n X_i \sum_{i=1}^n Y_i + n\sum_{i=1}^n X_i Y_i \end{bmatrix}$$

Assim, os estimadores de mínimos quadrados podem ser dados por

$$\begin{aligned} \hat{\beta}_1 &= \frac{-\sum_{i=1}^n X_i \sum_{i=1}^n Y_i + n\sum_{i=1}^n X_i Y_i}{n\sum_{i=1}^n X_i^2 - (\sum_{i=1}^n X_i)^2} \times \frac{n}{n} \\ &= \frac{\sum_{i=1}^n X_i Y_i - \frac{\sum_{i=1}^n X_i \sum_{i=1}^n Y_i}{n}}{\sum_{i=1}^n X_i^2 - \frac{(\sum_{i=1}^n X_i)^2}{n}} = \frac{SPXY}{SQX}. \end{aligned}$$

$$\begin{aligned} \hat{\beta}_0 &= \frac{\sum_{i=1}^n X_i^2 \sum_{i=1}^n Y_i - \sum_{i=1}^n X_i \sum_{i=1}^n X_i Y_i}{n\sum_{i=1}^n X_i^2 - (\sum_{i=1}^n X_i)^2} \times \frac{n}{n} \\ &= \frac{n\sum_{i=1}^n X_i^2 \sum_{i=1}^n Y_i - n\sum_{i=1}^n X_i \sum_{i=1}^n X_i Y_i}{n^2\sum_{i=1}^n X_i^2 - n(\sum_{i=1}^n X_i)^2} \\ &= \frac{n\sum_{i=1}^n X_i^2 \sum_{i=1}^n Y_i - \sum_{i=1}^n Y_i(\sum_{i=1}^n X_i)^2 + \sum_{i=1}^n Y_i(\sum_{i=1}^n X_i)^2 - n\sum_{i=1}^n X_i \sum_{i=1}^n X_i Y_i}{n^2\sum_{i=1}^n X_i^2 - n(\sum_{i=1}^n X_i)^2} \\ &= \frac{\sum_{i=1}^n Y_i[n\sum_{i=1}^n X_i^2 - (\sum_{i=1}^n X_i)^2] + \sum_{i=1}^n Y_i(\sum_{i=1}^n X_i)^2 - n\sum_{i=1}^n X_i \sum_{i=1}^n X_i Y_i}{n^2\sum_{i=1}^n X_i^2 - n(\sum_{i=1}^n X_i)^2} \\ &= \frac{\sum_{i=1}^n Y_i[n\sum_{i=1}^n X_i^2 - (\sum_{i=1}^n X_i)^2]}{n[n\sum_{i=1}^n X_i^2 - (\sum_{i=1}^n X_i)^2]} + \frac{\sum_{i=1}^n Y_i(\sum_{i=1}^n X_i)^2 - n\sum_{i=1}^n X_i \sum_{i=1}^n X_i Y_i}{n^2\sum_{i=1}^n X_i^2 - n(\sum_{i=1}^n X_i)^2} \\ &= \frac{\sum_{i=1}^n Y_i[n\sum_{i=1}^n X_i^2 - (\sum_{i=1}^n X_i)^2]}{n[n\sum_{i=1}^n X_i^2 - (\sum_{i=1}^n X_i)^2]} + \frac{\sum_{i=1}^n Y_i(\sum_{i=1}^n X_i) - n\sum_{i=1}^n X_i Y_i}{n\sum_{i=1}^n X_i^2 - (\sum_{i=1}^n X_i)^2} \times \frac{\sum_{i=1}^n X_i}{n} \\ &= \frac{\sum_{i=1}^n Y_i}{n} - \frac{-\sum_{i=1}^n Y_i(\sum_{i=1}^n X_i) + n\sum_{i=1}^n X_i Y_i}{n\sum_{i=1}^n X_i^2 - (\sum_{i=1}^n X_i)^2} \times \frac{\sum_{i=1}^n X_i}{n} \\ &= \bar{Y} - \hat{\beta}_1 \bar{X}. \end{aligned}$$

Exemplo 3 (Estimador de mínimos quadrados no MRLS). *Considerando o modelo de regressão linear simples em* (4)*, podemos expressar* $\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}$ *como:*

$$\begin{aligned} \boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} &= \sum_{i=1}^n \varepsilon_i^2 \\ &= \sum_{i=1}^n (Y_i - \beta_0 - \beta_1 X)^2. \end{aligned} \tag{8}$$

Para o ajuste de mínimos quadrados, devemos encontrar os valores para β_0 *e* β_1 *que minimizam*

(8). *Para isso, obtemos as seguintes derivadas parciais:*

$$\frac{\partial}{\partial \beta_0} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2 = -2 \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i), \tag{9}$$

$$\frac{\partial}{\partial \beta_1} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2 = -2 \sum_{i=1}^{n} X_i (Y_i - \beta_0 - \beta_1 X_i). \tag{10}$$

Igualando (9) *e* (10) *a zero, determinamos os estimadores* $\hat{\beta}_0$ *e* $\hat{\beta}_1$ *de* β_0 *e* β_1*, respectivamente. Assim, temos o chamado sistema de equações normais (SEN), dado por:*

$$SEN = \begin{cases} \sum_{i=1}^{n} Y_i - \hat{\beta}_1 \sum_{i=1}^{n} X_i & = n\hat{\beta}_0 \\ \sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} X_i \hat{\beta}_0 & = \hat{\beta}_1 \sum_{i=1}^{n} X_i^2 \end{cases}$$

Agora basta resolver o sistema. Para a primeira equação do SEN, temos que

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}, \tag{11}$$

sendo $\bar{Y} = \sum_{i=1}^{n} Y_i / n$ *e* $\bar{X} = \sum_{i=1}^{n} X_i / n$*. Substituindo* (11) *na segunda equação do SEN, temos que*

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} X_i Y_i - \dfrac{\sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{n}}{\sum_{i=1}^{n} X_i^2 - \dfrac{\left(\sum_{i=1}^{n} X_i\right)^2}{n}}. \tag{12}$$

Vejamos algumas propriedades para $\hat{\boldsymbol{\theta}}$,

Teorema 2 ($\hat{\boldsymbol{\theta}}$ como combinação linear do vetor $\boldsymbol{Y}$)**.** *Os elementos do vetor* $\hat{\boldsymbol{\theta}} = [\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p]'$*, estimador de mínimos quadrados de* $\boldsymbol{\theta}$*, são combinações lineares dos* Y_i*, isto é,* $\hat{\beta}_j$*,* $j = 0, 1, \ldots, p$*, pode ser expresso como:*

$$\hat{\beta}_j = \sum_{i=1}^{n} c_{ij} Y_i, \tag{13}$$

sendo, $c_{ij} = (\boldsymbol{x}'\boldsymbol{x}_j)\boldsymbol{x}_i'$*,* $(\boldsymbol{X}'\boldsymbol{X})^{-1} = \begin{bmatrix} \boldsymbol{x}'\boldsymbol{x}_1 \\ \boldsymbol{x}'\boldsymbol{x}_2 \\ \vdots \\ \boldsymbol{x}'\boldsymbol{x}_p \end{bmatrix}$*,* $\boldsymbol{x}'\boldsymbol{x}_j = [x_{j1}^*, x_{j2}^*, \ldots, x_{jp}^*]$*,*

$\boldsymbol{X}' = [\boldsymbol{x}_1', \boldsymbol{x}_2', \ldots, \boldsymbol{x}_n']$ *e* $\boldsymbol{x}_i' = \begin{bmatrix} x_{1i} \\ x_{2i} \\ \vdots \\ x_{pi} \end{bmatrix}$*.* □

Teorema 3 (Estimador não viesado de $\boldsymbol{\theta}$)**.** *Se* $E[\hat{\boldsymbol{\theta}}] = \boldsymbol{\theta}$ *então* $\hat{\boldsymbol{\theta}}$ *é não viesado de* $\boldsymbol{\theta}$*.* □

Demonstração.

$$E[\hat{\boldsymbol{\theta}}] = E[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}] = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'E[\boldsymbol{Y}] = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\theta} = \boldsymbol{\theta}.$$

□

Teorema 4 (Matriz de covariância de $\hat{\boldsymbol{\theta}}$)**.** *Se* $cov(\boldsymbol{Y}) = \boldsymbol{I}_n\sigma^2$*, a matriz de covariâncias de* $\hat{\boldsymbol{\theta}}$ *é*

$$cov[\hat{\boldsymbol{\theta}}] = \ \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}. \tag{14}$$

□

Demonstração.

$$\begin{aligned} cov[\hat{\boldsymbol{\theta}}] &= cov[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}] \\ &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'cov[\boldsymbol{Y}][(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}']' \\ &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{I}_n\sigma^2[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}']' \\ &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\sigma^2 \\ &= \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}. \end{aligned}$$

□

O teorema a seguir mostra que nenhum outro estimador não viesado apresenta menor variância que $\hat{\boldsymbol{\theta}}$, dentre os estimadores lineares não viesados.

Teorema 5 (Teorema de Gauss Markov)**.** *Considere o modelo linear em* (5) *com as suposições* $E[\boldsymbol{\varepsilon}] = \boldsymbol{0}$ *e* $cov[\boldsymbol{\varepsilon}] = \boldsymbol{I}_n\sigma^2$*. Considere ainda que* $\hat{\boldsymbol{\theta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$ *é o estimador de mínimos quadrados. Se* $\boldsymbol{A}'\boldsymbol{Y}$ *é um outro estimador de* $\boldsymbol{\theta}$ *tal que* $E[\boldsymbol{A}'\boldsymbol{Y}] = \boldsymbol{\theta}$*, em que* $\boldsymbol{A}$ *é uma matriz* $n \times p'$*, então:*

$$\boldsymbol{Z}'cov[\boldsymbol{A}\boldsymbol{Y}]\boldsymbol{Z} \geq \boldsymbol{Z}'cov[\hat{\boldsymbol{\theta}}]\boldsymbol{Z},\ \forall\ \boldsymbol{Z} \in \mathbb{R}^{p'}.$$

Demonstração. Sabemos que $cov[\boldsymbol{\varepsilon}] = \boldsymbol{I}_n\sigma^2$ e por consequência

$$cov[\boldsymbol{Y}] = \boldsymbol{I}_n\sigma^2. \tag{15}$$

Pelo Teorema 4 $cov[\hat{\boldsymbol{\theta}}] = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$. Por outro lado, usando (15) temos

$$cov[\boldsymbol{A}'\boldsymbol{Y}] = \boldsymbol{A}'cov[\boldsymbol{Y}]\boldsymbol{A} = \sigma^2(\boldsymbol{A}'\boldsymbol{A}). \tag{16}$$

Por hipótese $E[\boldsymbol{A}'\boldsymbol{Y}] = \boldsymbol{\theta}$, segue que

$$\boldsymbol{\theta} = \ \ E[\boldsymbol{A}'\boldsymbol{Y}] = \boldsymbol{A}'E[\boldsymbol{Y}] = \boldsymbol{A}'\boldsymbol{X}\boldsymbol{\theta},\ \forall\boldsymbol{\theta} \in \mathbb{R}^{p'}.$$

O que implica que

$$\boldsymbol{I}_{p'} = \boldsymbol{A}'\boldsymbol{X}. \tag{17}$$

Assim, usando (14), (16) e (17) segue que

$$\begin{aligned} cov[\boldsymbol{A}'\boldsymbol{Y}] - cov[\hat{\boldsymbol{\theta}}] =& \sigma^2(\boldsymbol{A}'\boldsymbol{A}) - \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1} \\ =& \sigma^2[\boldsymbol{A}'\boldsymbol{A} - \underbrace{\boldsymbol{A}'\boldsymbol{X}}_{\boldsymbol{I}_p}(\boldsymbol{X}'\boldsymbol{X})^{-1}\underbrace{\boldsymbol{X}'\boldsymbol{A}}_{\boldsymbol{I}_p}] \\ =& \sigma^2\boldsymbol{A}'[\boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}']\boldsymbol{A}. \end{aligned} \tag{18}$$

A matriz $\boldsymbol{M} = \boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ é simétrica e tal que $\boldsymbol{M}^2 = \boldsymbol{M}$ (idempotente). Assim,

$$\begin{aligned}\boldsymbol{Z}'cov[\boldsymbol{A}'\boldsymbol{Y}]\boldsymbol{Z} - \boldsymbol{Z}'cov[\hat{\boldsymbol{\theta}}]\boldsymbol{Z} =& \boldsymbol{Z}'[cov[\boldsymbol{A}'\boldsymbol{Y}] - cov[\hat{\boldsymbol{\theta}}]]\boldsymbol{Z} \\ =& \sigma^2\boldsymbol{Z}'[\boldsymbol{A}'\boldsymbol{M}\boldsymbol{A}]\boldsymbol{Z} \\ =& \sigma^2[\boldsymbol{M}'\boldsymbol{A}\boldsymbol{Z}]'[\boldsymbol{Z}'\boldsymbol{A}'\boldsymbol{M}] \\ =& \sigma^2||\boldsymbol{M}'\boldsymbol{A}\boldsymbol{Z}||^2 \geq 0,\end{aligned}$$

o que prova o resultado. □

Outra forma de apresentar o Teorema de Gauss-Markov é da seguinte forma:

Teorema 6 (Teorema de Gauss Markov)**.** *Se $E[\boldsymbol{Y}] = \boldsymbol{X}\boldsymbol{\theta}$ e $cov[\boldsymbol{Y}] = \boldsymbol{I}_n\sigma^2$, o vetor de estimadores de mínimos quadrados $\hat{\boldsymbol{\theta}}$ tem variância mínima dentre todos os outros vetores de estimadores lineares não viesados.* □

Demonstração. Considerando um estimador linear de $\boldsymbol{\theta}$ igual a $\boldsymbol{A}\boldsymbol{Y}$, sendo $\boldsymbol{A}$ uma matriz. Vamos identificar a matriz $\boldsymbol{A}$, para que $\boldsymbol{A}\boldsymbol{Y}$ seja um estimador linear não viesado de variância mínima. Para $\boldsymbol{A}\boldsymbol{Y}$ ser não viesado temos que $E[\boldsymbol{A}\boldsymbol{Y}] = \boldsymbol{\beta}$. Assim,

$$\begin{aligned}E[\boldsymbol{A}\boldsymbol{Y}] &= \boldsymbol{A}E[\boldsymbol{Y}] \\ &= \boldsymbol{A}\boldsymbol{X}\boldsymbol{\beta} \\ &= \boldsymbol{\beta}.\end{aligned}$$

Isso só será possível se $\boldsymbol{A}\boldsymbol{X} = \boldsymbol{I}$. A matriz de covariância de $\boldsymbol{A}\boldsymbol{X}$ é dada por

$$cov(\boldsymbol{A}\boldsymbol{Y}) = \boldsymbol{A}\sigma^2\boldsymbol{I}\boldsymbol{A}' = \sigma^2\boldsymbol{A}\boldsymbol{A}'.$$

As variâncias de $\hat{\boldsymbol{\theta}}$ estão na diagonal de $\sigma^2\boldsymbol{A}\boldsymbol{A}'$. Precisamos assim, determinar $\boldsymbol{A}$, sabendo que $\boldsymbol{A}\boldsymbol{X} = \boldsymbol{I}$, de tal modo que os elementos da diagonal sejam mínimos. Dessa forma, temos

$$\begin{aligned}\boldsymbol{A}\boldsymbol{A}' &= [\boldsymbol{A} - \underbrace{(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'}_{=0}][\boldsymbol{A} - \underbrace{(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'}_{=0}]' \\ &= [\boldsymbol{A} - (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'][\boldsymbol{A} - (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}']' + [\boldsymbol{A} - (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'][(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}']' + \\ &\quad + [(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'][\boldsymbol{A} - (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}']' + [(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}]' \\ &= [\boldsymbol{A} - (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'][\boldsymbol{A} - (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}']' + \underbrace{\boldsymbol{A}\boldsymbol{X}}_{\boldsymbol{I}}(\boldsymbol{X}'\boldsymbol{X})^{-1} - (\boldsymbol{X}'\boldsymbol{X})^{-1} + \\ &\quad (\boldsymbol{X}'\boldsymbol{X})^{-1}\underbrace{\boldsymbol{X}'\boldsymbol{A}'}_{\boldsymbol{I}} - (\boldsymbol{X}'\boldsymbol{X})^{-1} + (\boldsymbol{X}'\boldsymbol{X})^{-1} \\ &= [\boldsymbol{A} - (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'][\boldsymbol{A} - (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}']' + (\boldsymbol{X}'\boldsymbol{X})^{-1}.\end{aligned}$$

Como a matriz $[\boldsymbol{A} - (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'][\boldsymbol{A} - (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}']'$ é positiva semi-definida, os elementos da sua diagonal são maiores ou iguais a zero. O elementos da diagonal podem ser iguais a zero se escolhermos $\boldsymbol{A} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$. Desse modo, o estimador de variância mínima de $\boldsymbol{\beta}$ é

$$\boldsymbol{\theta} = \boldsymbol{A}\boldsymbol{Y} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y},$$

que é igual ao estimador de mínimos quadrados de $\boldsymbol{\beta}$. □

Adicionando ao modelo expresso em (5) a pressuposição $\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \boldsymbol{I}_n\sigma^2)$ ou equivalentemente $\boldsymbol{Y} \sim N_n(\boldsymbol{X\theta}, \boldsymbol{I}_n\sigma^2)$, isto é, que o vetor de erros $\boldsymbol{\varepsilon}$ tem distribuição normal multivariada n dimensional com o vetor de médias igual a $\mathbf{0}$ e matriz de covariância igual a $\boldsymbol{I}_n\sigma^2$, podemos apresentar um outro método de estimação no seguinte teorema,

Teorema 7 (Estimadores de máxima verossimilhança de $\boldsymbol{\theta}$ e σ^2)**.** *Considere o modelo linear em* (5)*. Se* $\boldsymbol{Y} \sim N_n(\boldsymbol{X\theta}, \boldsymbol{I}\sigma^2)$*, em que* $\boldsymbol{X}$ *é* $n \times p'$ *de posto* $p' < n$*, os estimadores de máxima verossimilhança de* $\boldsymbol{\theta}$ *e* σ^2 *são*

$$\hat{\boldsymbol{\theta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}. \tag{19}$$

e

$$\hat{\sigma}^2 = \frac{1}{n}(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\theta}})'(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\theta}}), \tag{20}$$

respectivamente. □

Demonstração. A função de verossimilhança é dado pela função densidade conjunta dos Ys que denotamos por $L(\boldsymbol{\theta}, \sigma^2|\boldsymbol{Y}) = f_{\boldsymbol{Y}}(\boldsymbol{Y}; \boldsymbol{\theta}, \sigma^2)$. Como os elementos do vetor $\boldsymbol{Y}$ são amostrados independentemente e considerando σ^2 conhecido, então

$$\begin{aligned} L(\boldsymbol{\theta}, \sigma^2|\boldsymbol{Y}) = & \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2}\left(\frac{y_i - \boldsymbol{x}_i'\boldsymbol{\theta}}{\sigma}\right)^2\right\} \\ = & \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - \boldsymbol{x}_i'\boldsymbol{\theta}\right)\right\} \\ = & \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\theta}})'(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\theta})\right\}, \end{aligned} \tag{21}$$

sendo $\boldsymbol{x}_i'$ a i-ésima linha da matriz. Aplicando o logaritmo neperiano em (21), temos a função suporte dada por:

$$\begin{aligned} \ell(\boldsymbol{\theta}|\boldsymbol{Y}, \sigma^2) = & \ln(1) - \frac{n}{2}[\ln(2\pi) + \ln(\sigma^2)] - \frac{1}{2\sigma^2}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\theta})'(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\theta}) \\ = & -\frac{n}{2}[\ln(2\pi) + \ln(\sigma^2)] - \frac{1}{2\sigma^2}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\theta})'(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\theta}). \end{aligned} \tag{22}$$

Em busca de maximizar (22) em relação a $\boldsymbol{\theta}$, percebemos que uma vez que σ^2 é conhecido, o primeiro termo e o denominador do segundo termo são fixos e ambos negativos, então basta minimizar o numerador $(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\theta})'(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\theta})$ e derivar em relação a $\boldsymbol{\theta}$ (**??**, p. 77). Posteriormente igualamos a zero para obter o estimador de máxima verossimilhança. Este procedimento é igual ao que foi utilizado no Teorema 1, e assim, concluímos que $\hat{\boldsymbol{\theta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$. Para estimar σ^2, vamos substituir $\boldsymbol{\theta}$ por $\hat{\boldsymbol{\theta}}$ e novamente usar a expressão (22), isto é,

$$\ell(\sigma^2|\boldsymbol{Y}, \hat{\boldsymbol{\theta}}) = -\frac{n}{2}[\ln(2\pi) + \ln(\sigma^2)] - \frac{1}{2\sigma^2}(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\theta}})'(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\theta}}). \tag{23}$$

Para maximizar (23) em relação a σ^2, vamos derivá-la em relação a σ^2 fazendo

$$\frac{d\ell(\sigma^2|\boldsymbol{Y}, \hat{\boldsymbol{\theta}})}{d\sigma^2} = -\frac{n}{2\sigma^2} + \frac{(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\theta}})'(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\theta}})}{2(\sigma^2)^2} \tag{24}$$

Igualando a expressão (24) a zero, temos

$$\begin{aligned}
-\frac{n}{2\hat{\sigma^2}} + \frac{(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\theta}})'(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\theta}})}{2(\hat{\sigma^2})^2} &= 0 \\
\frac{-n\hat{\sigma^2} + (\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\theta}})'(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\theta}})}{2(\hat{\sigma^2})^2} &= 0 \\
-n\hat{\sigma^2} + (\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\theta}})'(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\theta}}) &= 0 \\
n\hat{\sigma^2} &= (\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\theta}})'(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\theta}}) \\
\hat{\sigma^2} &= \frac{(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\theta}})'(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\theta}})}{n},
\end{aligned}$$

o que prova o resultado. □

Sabemos que $\hat{\boldsymbol{\theta}}$ é um estimador não viesado para $\boldsymbol{\theta}$, Teorema 3. Entretanto, não podemos afirmar o mesmo para $\hat{\sigma}^2$. Um estimador não viesado para σ^2 é dado por:

$$S^2 = QME = \frac{1}{n - p'}(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\theta}})'(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\theta}}), \tag{25}$$

sendo p' o número de Xs e QME também conhecido como o quadrado médio do erro. como pode ser apresentado nos próximos teoremas.

Teorema 8 (Teorema do valor esperado de uma forma quadrática)**.** *Se $\boldsymbol{Y}$ é um vetor aleatório com média $\boldsymbol{\mu}$ e matriz de covariâncias $\boldsymbol{\Sigma}$ e $\boldsymbol{A}$ é uma matriz simétrica de constantes, então*

$$E[\boldsymbol{Y'AY}] = tr(\boldsymbol{A\Sigma}) + \boldsymbol{\mu}'\boldsymbol{A\mu}. \tag{26}$$

Demonstração. Sabemos que $\boldsymbol{\Sigma} = E[\boldsymbol{YY}'] - \boldsymbol{\mu\mu}'$ que pode ser escrita como

$$E[\boldsymbol{YY}'] = \boldsymbol{\Sigma} + \boldsymbol{\mu\mu}'.$$

Se considerarmos $\boldsymbol{Y'AY}$ como um escalar, este é igual ao seu traço. Assim,

$$\begin{aligned}
E[\boldsymbol{Y'AY}] &= E[tr(\boldsymbol{Y'AY})] \\
&= E[tr(\boldsymbol{AYY'})] \\
&= tr(E[\boldsymbol{AYY'}]) \\
&= tr(\boldsymbol{A}E[\boldsymbol{YY'}]) \\
&= tr(\boldsymbol{A}[\boldsymbol{\Sigma} + \boldsymbol{\mu\mu}']) \\
&= tr(\boldsymbol{A\Sigma} + \boldsymbol{A\mu\mu}') \\
&= tr(\boldsymbol{A\Sigma}) + \boldsymbol{\mu}'\boldsymbol{A\mu}.
\end{aligned}$$

□

Teorema 9 (Estimador não viesado de σ^2)**.** *Se $S^2 = QME = \frac{1}{n-p'}(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\theta}})'(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\theta}})$, e se $E[\boldsymbol{Y}] = \boldsymbol{X\theta}$ e $cov(\boldsymbol{Y}) = \sigma^2\boldsymbol{I}_n$, então*

$$E[S^2] = \sigma^2, \tag{27}$$

isto é, S^2 é um estimador não viesado de σ^2.

Demonstração. Podemos escrever $(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\theta}})'(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\theta}})$ como uma forma quadrática da seguinte forma:

$$\begin{aligned}(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\theta}})'(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\theta}}) &= \boldsymbol{Y}'\boldsymbol{Y} - \hat{\boldsymbol{\theta}}'\boldsymbol{X}'\boldsymbol{Y} \\ &= \boldsymbol{Y}'\boldsymbol{Y} - [(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}]'\boldsymbol{X}'\boldsymbol{Y} \\ &= \boldsymbol{Y}'\boldsymbol{Y} - \boldsymbol{Y}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y} \\ &= \boldsymbol{Y}'[\boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}']\boldsymbol{Y}'\end{aligned}$$

Pelo Teorema 8, temos que

$$\begin{aligned}E[(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\theta}})'(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\theta}})] &= tr([\boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}']\sigma^2\boldsymbol{I}_n) + E[\boldsymbol{Y}'][\boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}']E[\boldsymbol{Y}] \\ &= \sigma^2 tr([\boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}']) + \boldsymbol{\theta}'\boldsymbol{X}'[\boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}']\boldsymbol{X}\boldsymbol{\theta} \\ &= \sigma^2[n - tr(\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}')] + \boldsymbol{\theta}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\theta} - \boldsymbol{\theta}'\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\theta} \\ &= \sigma^2[n - tr\{(\boldsymbol{X}'\boldsymbol{X})^{-1}\}] + \boldsymbol{\theta}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\theta} - \boldsymbol{\theta}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\theta} \\ &= \sigma^2(n - p').\end{aligned}$$

e então, $E[S^2] = E\left[\frac{1}{n-p'}(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\theta}})'(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\theta}})\right] = \frac{1}{n-p'}E\left[\sigma^2(n-p')\right] = \sigma^2$. □

Corolário 1. *Um estimador não viesado de $cov(\boldsymbol{\theta})$ é dado por*

$$\hat{cov}(\hat{\beta}) = S^2(\boldsymbol{X}'\boldsymbol{X})^{-1}. \tag{28}$$

As propriedades dos estimadores de máxima verossimilhança podem ser apresentadas da seguinte forma,

Teorema 10 (Propriedades de $\hat{\boldsymbol{\theta}}$ e $\hat{\sigma}^2$)**.** *Supondo que $Y \sim N_n(\boldsymbol{X}\boldsymbol{\theta}, I\sigma^2)$, em que $\boldsymbol{X}$ é $n \times p'$ de posto $p' < n$ e $\boldsymbol{\theta} = [\beta_0, \beta_1, \ldots, \beta_p]'$. Então os estimadores de máxima verossimilhança $\hat{\boldsymbol{\theta}}$ e $\hat{\sigma}^2$, apresentados no teorema 7, têm as seguintes propriedade distribucionais:*

(i) $\hat{\boldsymbol{\theta}} \sim N_{p'}[\boldsymbol{\theta}, \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}]$;

(ii) $n\hat{\sigma}^2/\sigma^2 \sim \chi^2_{n-p'}$, *ou* $(n-p')QME/\sigma^2 \sim \chi^2_{n-p'}$;

(iii) $\hat{\boldsymbol{\theta}}$ *e* $\hat{\sigma}^2$ *são independentes.* □

Uma outra propriedade interessante dos estimadores de máxima verossimilhança, sob a suposição de normalidade é que estes conseguem capturar toda a informação sobre os parâmetros contidos na amostra. Esta característica é chamada de suficiência.

Teorema 11 (Suficiência dos estimadores $\hat{\boldsymbol{\theta}}$ e $\hat{\sigma}^2$)**.** *Se $Y \sim N_n(\boldsymbol{X}\boldsymbol{\theta}, I\sigma^2)$, então $\hat{\boldsymbol{\theta}}$ e $\hat{\sigma}^2$ são conjuntamente suficientes para $\boldsymbol{\theta}$ e σ^2.* □

Demonstração. Prova em Rencher p. 144. □

Teorema 12 (BLUE)**.** *Se $Y \sim N_n(\boldsymbol{X}\boldsymbol{\theta}, I\sigma^2)$, então $\hat{\boldsymbol{\theta}}$ e s^2 têm menor variância dentre todos os estimadores não viesados de $\boldsymbol{\theta}$ e σ^2.* □

Demonstração. Graybill 176. □

Depois de estimado os parâmetros do modelo, queremos avaliar o ajuste do modelo. Pensando em um ajuste para um modelo de regressão linear simples, para cada observação Y_i existe uma estimativa $\hat{Y}_i$ que pertence a reta de regressão estimada e um valor fixo $\bar{X}$ que é a média das observações de $\boldsymbol{Y}$. A representação gráfica pode ser apresentada na Figura 2. Nada mais intuitivo e avaliarmos:

$$\underbrace{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}_{SQT} = \underbrace{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}_{SQReg} + \underbrace{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}_{SQErro}. \tag{29}$$

Assim, a variação total dos dados pode ser explicada pelo modelo (SQReg) ou não (Não

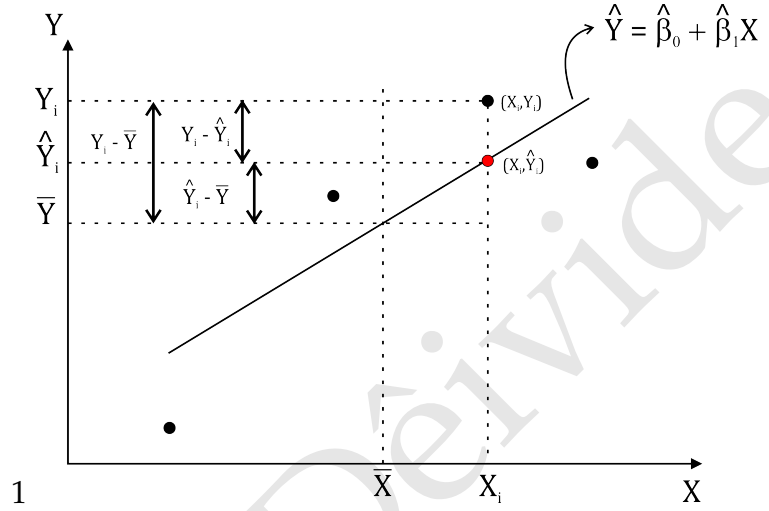

Figura 2: Somas de quadrados de um modelo de regressão.

sendo, implica que SQErro é maior do que SQReg estatisticamente). Então, poderíamos as somas de quadrados da seguinte forma:

- Soma de quadrado total:

$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n} Y_i^2 - \frac{(\sum_{i=1}^{n} Y_i)^2}{n}$$

- Soma de quadrado da regressão:

$$\begin{aligned}\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 &= \sum_{i=1}^{n}(\hat{\beta}_0 + \hat{\beta}_1 X_i - \bar{Y})^2 \\ &= \sum_{i=1}^{n}(\bar{Y} - \hat{\beta}_1 \bar{X} + \hat{\beta}_1 X_i - \bar{Y})^2 \\ &= \hat{\beta}_1^2 \sum_{i=1}^{n}(X_i - \bar{X})^2\end{aligned}$$

- Soma de quadrado do erro:

$$\begin{aligned}
\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 &= \sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2 \\
&= \sum_{i=1}^{n}(Y_i - \bar{Y} + \hat{\beta}_1\bar{X} - \hat{\beta}_1 X_i)^2 \\
&= \sum_{i=1}^{n}\left\{(Y_i - \bar{Y}) - \hat{\beta}_1(X_i - \bar{X})\right\}^2 \\
&= \sum_{i=1}^{n}(Y_i - \bar{Y})^2 + \hat{\beta}_1^2\sum_{i=1}^{n}(X_i - \bar{X})^2 - 2\hat{\beta}_1\sum_{i=1}^{n}(Y_i - \bar{Y})(X_i - \bar{X}) \\
&= SQY + \hat{\beta}_1^2 SQX - 2\hat{\beta}_1 SPXY \\
&= SQY + \hat{\beta}_1^2 SQX - 2\hat{\beta}_1 SPXY\frac{SQX}{SQX} \\
&= SQY + \hat{\beta}_1^2 SQX - 2\hat{\beta}_1^2 SQX \\
&= \sum_{i=1}^{n}(Y_i - \bar{Y})^2 - \hat{\beta}_1^2\sum_{i=1}^{n}(X_i - \bar{X})^2 \\
&= SQT - SQReg.
\end{aligned}$$

Pensando em um modelo de regressão linear múltipla, expressa em (5), podemos obter as somas de quadrados da seguinte forma:

- Soma de quadrado total:

$$\begin{aligned}
SQT &= \sum_{i=1}^{n} Y_i^2 - \frac{(\sum_{i=1}^{n} Y_i)^2}{n} \\
&= \boldsymbol{Y}'\boldsymbol{Y} - \boldsymbol{Y}'\left(\frac{1}{n}\boldsymbol{J}\right)\boldsymbol{Y}, \quad \text{(Rencher port. 74)}
\end{aligned}$$

  sendo $\boldsymbol{J}$ uma matriz simétrica de tamanho n de 1's.

- Soma de quadrado de erro:

$$\begin{aligned}
SQErro &= (\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\theta}})'(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\theta}}) \\
&= \boldsymbol{Y}'\boldsymbol{Y} - 2\hat{\boldsymbol{\theta}}'\boldsymbol{X}'\boldsymbol{Y} + \hat{\boldsymbol{\theta}}'\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\theta}} \quad (SEN:\ \boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\theta}} = \boldsymbol{X}'\boldsymbol{Y}) \\
&= \boldsymbol{Y}'\boldsymbol{Y} - \hat{\boldsymbol{\theta}}'\boldsymbol{X}'\boldsymbol{Y}
\end{aligned}$$

- Soma de quadrado da Regressão:

$$\begin{aligned}
SQReg &= SQT - SQErro \\
&= \boldsymbol{Y}'\boldsymbol{Y} - \boldsymbol{Y}'\left(\frac{1}{n}\boldsymbol{J}\right)\boldsymbol{Y} - [\boldsymbol{Y}'\boldsymbol{Y} - \hat{\boldsymbol{\theta}}'\boldsymbol{X}'\boldsymbol{Y}] \\
&= \hat{\boldsymbol{\theta}}'\boldsymbol{X}'\boldsymbol{Y} - \boldsymbol{Y}'\left(\frac{1}{n}\boldsymbol{J}\right)\boldsymbol{Y}
\end{aligned}$$

Vejamos algumas provas interessante sobre as somas de quadrados.

Teorema 13. *Seja as v.a. iid com* $X_i \sim N\left(\mu_i; \sigma_i^2\right)$, *então a v.a.* $U = \sum_{i=1}^{n} \left(\frac{X_i - \mu_i}{\sigma_i}\right)^2$ *tem distribuição qui-quadrado com n graus de liberdade.*

Demonstração. Seja $Z_i = \frac{X_i - \mu_i}{\sigma_i}$. Logo $Z_i \sim N(0,1)$. Agora,

$$m_U(t) = E\left[e^{tU}\right] = E\left[e^{t\sum_{i=1}^{n} Z_i^2}\right] = E\left[\prod_{i=1}^{n} e^{tZ_i^2}\right] = \prod_{i=1}^{n} E\left[e^{tZ_i^2}\right].$$

Mas,

$$\begin{aligned}
E\left[e^{tZ^2}\right] &= \int_{-\infty}^{\infty} e^{tZ^2} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}Z^2} dz = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{\left(t-\frac{1}{2}\right)Z^2} dz \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(1-2t)Z^2} dz = \frac{\sqrt{1-2t}}{\sqrt{1-2t}} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(1-2t)Z^2} dz \\
&= \frac{1}{\sqrt{1-2t}} \int_{-\infty}^{\infty} \frac{\sqrt{1-2t}}{\sqrt{2\pi}} e^{-\frac{1}{2}(1-2t)Z^2} dz = \frac{1}{\sqrt{1-2t}} \underbrace{\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\frac{1}{(1-2t)}}} e^{-\frac{1}{2}\left(\frac{Z}{\sqrt{\frac{1}{1-2t}}}\right)^2} dz}_{=1} \\
&= \frac{1}{\sqrt{1-2t}} = \left(\frac{1}{1-2t}\right)^{\frac{1}{2}} = \left(\frac{\frac{1}{2}}{\frac{1}{2}-t}\right)^{\frac{1}{2}},
\end{aligned}$$

em que a integral resulta em 1, pois é a densidade de uma Normal padrão com média 0 e variância $\frac{1}{1-2t}$. Portanto,

$$m_U(t) = \prod_{i=1}^{n} \left(\frac{\frac{1}{2}}{\frac{1}{2}-t}\right)^{\frac{1}{2}} = \left[\left(\frac{\frac{1}{2}}{\frac{1}{2}-t}\right)^{\frac{1}{2}}\right]^n = \left(\frac{\frac{1}{2}}{\frac{1}{2}-t}\right)^{\frac{n}{2}}, \; t < \frac{1}{2},$$

que é f.g.m. de uma distribuição qui-quadrado com n graus de liberdade. □

O próximo teorema também é necessário.

Teorema 14. *Seja* $Z_1, \ldots, Z_n$ *uma a.a. e* $Z_i \sim N(0,1)$. *Considere que* $\bar{Z}$ *e* $\sum_{i=1}^{n} (Z_i - \bar{Z})^2$ *são independentes. Seja* $W = \sum_{i=1}^{n} (Z_i - \bar{Z})^2$. *Então* $W \sim \chi^2(n-1)$.

Demonstração. Note que

$$\begin{aligned}
\sum Z_i^2 &= \sum (Z_i - \bar{Z} + \bar{Z})^2 = \sum \left[(Z_i - \bar{Z})^2 + 2(Z_i - \bar{Z})\bar{Z} + \bar{Z}^2\right] \\
&= \sum (Z_i - \bar{Z})^2 + 2\bar{Z}\sum (Z_i - \bar{Z}) + n\bar{Z}^2 = \sum (Z_i - \bar{Z})^2 + n\bar{Z}^2.
\end{aligned}$$

Como $\sum (Z_i - \bar{Z})^2$ e $\bar{Z}$ são independentes , temos que

$$m_{\sum Z_i^2}(t) = m_{\sum (Z_i - \bar{Z})^2}(t)\, m_{n\bar{Z}^2}(t).$$

Daí,

$$m_{\sum (Z_i - \bar{Z})^2}(t) = \frac{m_{\sum Z_i^2}(t)}{m_{n\bar{Z}^2}(t)} = \frac{\left(\frac{\frac{1}{2}}{\frac{1}{2}-t}\right)^{\frac{n}{2}}}{\left(\frac{\frac{1}{2}}{\frac{1}{2}-t}\right)^{\frac{1}{2}}} = \left(\frac{\frac{1}{2}}{\frac{1}{2}-t}\right)^{\frac{(n-1)}{2}}, \; t < \frac{1}{2},$$

que é a f.g.m. de uma $\chi^2(n-1)$. Observe ainda que se $Z_i \sim N(0,1)$, $Z_1 + \ldots + Z_n \sim N(0,n)$, então $\bar{Z}_n \sim N\left(0, \frac{1}{n}\right)$. Logo

$$\sqrt{n}\bar{Z}_n \sim N(0,1) \Rightarrow \left(\sqrt{n}\bar{Z}_n\right)^2 = n\bar{Z}_n^2 \sim \chi^2(1).$$

□

Com os dois teoremas anteriores, temos condições de determinar a distribuição amostral para S^2 e será pelo teorema a seguir. Antes disso, note as seguintes consequências dos teoremas vistos:

(i) Como $\sum (Z_i - \bar{Z})^2$ e $\bar{Z}$ são independentes, implica que $\bar{X}$ e $\sum (X_i - \bar{X})^2$ também o são.

(ii) Como $\sum (Z_i - \bar{Z})^2 \sim \chi^2(n-1)$, implica que $\sum \frac{(X_i - \bar{X})^2}{\sigma^2}$ também tem distribuição $\chi^2(n-1)$.

(iii) $\sum (Z_i - \bar{Z})^2 = \sum \left(\frac{X_i - \mu}{\sigma} - \frac{\bar{X} - \mu}{\sigma}\right)^2 = \sum \left(\frac{X_i - \bar{X}}{\sigma}\right)^2$.

Podemos expressar as somas de quadrados em termos de formas quadráticas:

- Forma quadrática de SQT:

$$\begin{aligned} SQT &= \boldsymbol{Y}'\boldsymbol{Y} - \boldsymbol{Y}'\left(\frac{1}{n}\boldsymbol{J}\right)\boldsymbol{Y} \\ &= \boldsymbol{Y}'\left(\boldsymbol{I}_n - \frac{1}{n}\boldsymbol{J}\right)\boldsymbol{Y}, \end{aligned}$$

  em que $\boldsymbol{I}_n - \frac{1}{n}\boldsymbol{J}$ é a matriz (núcleo) da forma quadrática, sendo esta simétrica, idempotente de posto (n-1).

- Forma quadrática de SQReg:

$$\begin{aligned} SQReg &= \hat{\boldsymbol{\theta}}'\boldsymbol{X}'\boldsymbol{Y} - \boldsymbol{Y}'\left(\frac{1}{n}\boldsymbol{J}\right)\boldsymbol{Y} \\ &= [(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}]'\boldsymbol{X}'\boldsymbol{Y} - \boldsymbol{Y}'\left(\frac{1}{n}\boldsymbol{J}\right)\boldsymbol{Y} \\ &= \boldsymbol{Y}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y} - \boldsymbol{Y}'\left(\frac{1}{n}\boldsymbol{J}\right)\boldsymbol{Y} \\ &= \boldsymbol{Y}'\left[\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' - \left(\frac{1}{n}\boldsymbol{J}\right)\right]\boldsymbol{Y} \\ &= \boldsymbol{Y}'\left[\boldsymbol{H} - \left(\frac{1}{n}\boldsymbol{J}\right)\right]\boldsymbol{Y}, \end{aligned}$$

sendo $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'^{'}$ e que $\boldsymbol{H} - \left(\frac{1}{n}\boldsymbol{J}\right)$ é a matriz (núcleo), sendo esta simétrica e idempotente de posto .

- Forma quadrática de SQErro:

$$\begin{aligned} SQErro &= \boldsymbol{Y}'\boldsymbol{Y} - \hat{\boldsymbol{\theta}}'\boldsymbol{X}'\boldsymbol{Y} \\ &= \boldsymbol{Y}'\boldsymbol{Y} - [(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}]'\boldsymbol{X}'\boldsymbol{Y} \\ &= \boldsymbol{Y}'\boldsymbol{Y} - \boldsymbol{Y}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y} \\ &= \boldsymbol{Y}'\left[\boldsymbol{I}_n - \boldsymbol{H}\right]\boldsymbol{Y}, \end{aligned}$$

sendo $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'^{'}$ e $\boldsymbol{I}_n - \boldsymbol{H}$ a matriz (núcleo), sendo esta simétrica.

Pelo Teorema 8 podemos calcular a esperança das somas de quadrados da seguinte forma:

- Esperança de SQT:

$$\begin{aligned} E[SQT] &= E\left[\boldsymbol{Y}'\left(\boldsymbol{I}_n - \frac{1}{n}\boldsymbol{J}\right)\boldsymbol{Y}\right] \\ &= tr\left\{\left(\boldsymbol{I}_n - \frac{1}{n}\boldsymbol{J}\right)\sigma^2\right\} + \hat{\boldsymbol{\theta}}'\boldsymbol{X}'\left(\boldsymbol{I}_n - \frac{1}{n}\boldsymbol{J}\right)\boldsymbol{X}\hat{\boldsymbol{\theta}} \\ &= \sigma^2\left[tr\left\{\boldsymbol{I}_n\right\} - tr\left\{\frac{1}{n}\boldsymbol{J}\right\}\right] + \underbrace{\hat{\boldsymbol{\beta}}'\boldsymbol{X}_1'\boldsymbol{X}_1'\hat{\boldsymbol{\beta}}}_{\text{Ver Demetrio p. 80, Rencher port. 126}} \\ &= \sigma^2(n-1) + \hat{\boldsymbol{\beta}}'\boldsymbol{X}_1'\boldsymbol{X}_1\hat{\boldsymbol{\beta}}, \end{aligned}$$

sendo $\boldsymbol{X}_1$ a matriz X sem a coluna de 1's e $\hat{\boldsymbol{\beta}} = [\beta_1, \beta_2, \ldots, \beta_p]'$.

- Esperança de SQErro:

$$\begin{aligned} E[SQErro] &= E\left[\boldsymbol{Y}'\left[\boldsymbol{I}_n - \boldsymbol{H}\right]\boldsymbol{Y}\right] \\ &= tr\left\{\left(\boldsymbol{I}_n - \boldsymbol{H}\right)\sigma^2\right\} + \hat{\boldsymbol{\theta}}'\boldsymbol{X}'\left(\boldsymbol{I}_n - \boldsymbol{H}\right)\boldsymbol{X}\hat{\boldsymbol{\theta}} \\ &= \sigma^2[tr\left\{\boldsymbol{I}_n\right\} - tr\left\{\boldsymbol{H}\right\}] + \hat{\boldsymbol{\theta}}'\boldsymbol{X}'\left(\boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\right)\boldsymbol{X}\hat{\boldsymbol{\theta}} \\ &= \sigma^2(n-p') + \hat{\boldsymbol{\theta}}'\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}'\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\theta}} \\ &= \sigma^2(n-p'). \end{aligned}$$

- Esperança de SQReg:

$$\begin{aligned} E[SQReg] &= E\left[SQT - SQE\right] \\ &= \sigma^2(n-1) + \hat{\boldsymbol{\beta}}'\boldsymbol{X}_1'\boldsymbol{X}_1\hat{\boldsymbol{\beta}} - \sigma^2(n-p') \\ &= \sigma^2(n-1-n+p') + \hat{\boldsymbol{\beta}}'\boldsymbol{X}_1'\boldsymbol{X}_1\hat{\boldsymbol{\beta}} \\ &= \sigma^2(p'-1) + \hat{\boldsymbol{\beta}}'\boldsymbol{X}_1'\boldsymbol{X}_1\hat{\boldsymbol{\beta}}. \end{aligned}$$

Dizemos que os graus de liberdade associados a uma forma quadrática $\boldsymbol{Y}'\boldsymbol{A}\boldsymbol{Y}$, é dado pela característica$(\boldsymbol{A})$, e considerando esta idempotente, então característica$(\boldsymbol{A}) =$

$tr(\boldsymbol{A})$. Assim, os graus de liberdade para SQT, SQReg e SQErro são (n-1), (p'-1) e (n - p'), respectivamente.

Definimos o quadrado médio (QM) como a razão entre a soma de quadrado médios e o seu respectivo grau de liberdade. Assim, temos:

- Quadrado médio da regressão:

$$QMReg = \frac{SQReg}{p' - 1},$$

- Quadrado médio do resíduo:

$$QMErro = \frac{SQErro}{n - p'}$$

Aplicando a esperança nesses quadrados médios, temos:

- Esperança do QMReg:

$$E[QMReg] = \sigma^2 + \frac{\hat{\boldsymbol{\beta}}' \boldsymbol{X}_1' \boldsymbol{X}_1 \hat{\boldsymbol{\beta}}}{p' - 1},$$

- Esperança de QMErro:

$$E[QMErro] = \sigma^2.$$

Os dois QMs praticamente estimam o valor de σ^2, a menos do primeiro que há um acréscimo de uma quantidade que depende do vetor $\hat{\boldsymbol{\beta}}$, Se considerássemos o estimador $\hat{\boldsymbol{\beta}}$ estatisticamente igual a um vetor de 0's, logo $E[QMReg] = \sigma^2$. Isso reflete na contribuição que o modelo consegue explicar a variação da variável resposta, que nessa situação, não passa da própria variação devida ao acaso, uma vez que $Var[\epsilon_i] = \sigma^2$. Caso algum dos β_i's (exceto o β_0) seja significativamente diferente de 0, o $E[QMReg] > \sigma^2$, e isso implicará que o modelo estimado consegue explicar a variação da variável resposta além do que a própria variação devida ao acaso.

Para desenvolvermos um teste de hipóteses, vamos descobrir a distribuição das somas de quadrados.

Teorema 15 (Distribuição de uma forma quadrática)**.** *Se* $\boldsymbol{Y} \sim N_{p'}(\boldsymbol{\mu}, \sigma^2 \boldsymbol{I}_n)$*, então* $\frac{\boldsymbol{Y}'\boldsymbol{A}\boldsymbol{Y}}{\sigma^2} \sim \chi^2_{(r,\ \boldsymbol{\mu}'\boldsymbol{A}\boldsymbol{\mu}/2\sigma^2)}$*, se e somente se* $\boldsymbol{A}$ *é idempotente de posto r.* □

Podemos mostrar algumas informações relevantes sobre a distribuição de qui-quadrado. Se temos uma amostra aleatória $Y_1, Y_2, \ldots, Y_n$ de uma distribuição normal padrão, o vetor $\boldsymbol{Y} \sim N_n(\boldsymbol{0}, \boldsymbol{I})$. Por definição

$$\sum_{i=1}^{n} Y_i^2 = \boldsymbol{Y}'\boldsymbol{Y} \sim \chi^2_{(n)}, \tag{30}$$

isto é, a soma de quadrados de n variáveis aleatórias independentes e com distribuição normal padrão tem distribuição qui-quadrado (central) com n graus de liberdade.

Suponha agora que $\boldsymbol{Y} \sim N_n(\boldsymbol{\mu}, \boldsymbol{I})$. Agora, $\sum_{i=1}^{n} Y_i^2$ não tem distribuição qui-quadrado central, mas $\sum_{i=1}^{n}(Y_i - \mu_i)^2 \sim \chi^2_{(n)}$, já que $(Y_i - \mu_i) \sim N(0,1)$. Dizemos que $\sum_{i=1}^{n} Y_i^2$ tem distribuição qui-quadrado não central, denotada por $\chi^2_{(n,\lambda)}$. O parâmetro de não centralidade é definido como:

$$\lambda = \frac{1}{2}\boldsymbol{\mu}'\boldsymbol{\mu}.$$

Se $V \sim \chi^2_{(n,\lambda)}$, então a esperança de V é dada por $E[V] = n + 2\lambda$.

O fato de $Y_i \sim N(\mu, \sigma^2)$ ter $E[Y_i] = \mu$ não será problema para a distribuição de $\sum_{i=1}^{n} Y_i^2$ não será problema, pois reflitirá no parâmetro de não centralidade, porém $Var[Y_i] = \sigma^2$ é um problema, uma vez que os graus de liberdade da distribuição de $\sum_{i=1}^{n} Y_i^2$ dependerá de σ^2, e isso será um problema quando pensarmos na razão de quadrados médios que terá distribuição F que dependerá dos graus de liberdade destes quadrados médios. Assim, podemos eliminar esse problema dividindo Y_i/σ e daí $Y_i/\sigma \sim N(\mu/\sigma, 1)$. Logo, $\sum_{i=1}^{n}\left(\frac{Y_i}{\sigma}\right)^2 \sim \chi^2_{(n,\lambda)}$, tal que $\lambda = \frac{1}{2\sigma^2}\boldsymbol{\mu}'\boldsymbol{\mu}$, e sua esperança pode ser expressa por $n + \frac{1}{\sigma^2}\boldsymbol{\mu}'\boldsymbol{\mu}$.

De um modo geral, diremos que a distribuição de uma forma quadrática sob as condições do Teorema 15 terá distribuição qui-quadrado central se $\lambda = 0$. Como as somas de quadrados apresentados anteriormente, são formas quadráticas, tal que $\boldsymbol{Y} \sim N_n(\boldsymbol{X\theta}, \sigma^2\boldsymbol{I}_n)$, então

$$\frac{SQReg}{\sigma^2} \sim \chi^2_{(p'-1,\lambda)}, \tag{31}$$

$$\frac{SQErro}{\sigma^2} \sim \chi^2_{(n-p')} \quad \text{(Central)}, \tag{32}$$

$$\frac{SQT}{\sigma^2} \sim \chi^2_{(n-1,\lambda)}, \tag{33}$$

sendo o parâmetro de não centralidade dado por $\lambda = \frac{\hat{\beta}'\boldsymbol{X}_1'\boldsymbol{X}_1\hat{\boldsymbol{\beta}}}{2\sigma^2}$.

Para concluir a proposta do teste para verificar a significância de algum dos elementos do vetor $\boldsymbol{\beta}$, se $U \sim \chi_{(p)}$ e $V \sim \chi_{(q)}$, sendo U e V independentes, então a razão

$$W = \frac{U/p}{V/q} \sim F_{(p,q)}, \tag{34}$$

isto é, W tem distribuição F central com p e q graus de liberdade.

Supondo $U \sim \chi_{(p,\lambda)}$ e $V \sim \chi_{(q)}$, sendo U e V independentes, então a razão

$$W = \frac{U/p}{V/q} \sim F_{(p,q,\lambda)}, \tag{35}$$

isto é, W tem distribuição F não central com p e q graus de liberdade e parâmetro de não centralidade λ.

Assim, testando a hipótese:

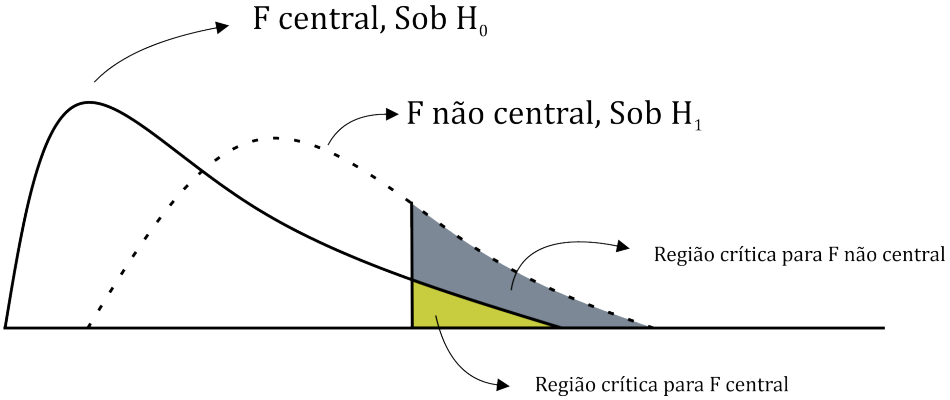

Figura 3: Decisão do teste F para ANAVA.

FV	GL	SQ	QM	E[QM]	F
Devida a $\boldsymbol{\beta}$	$p'-1$	$\boldsymbol{Y}'\left[\boldsymbol{H}-\left(\frac{1}{n}\boldsymbol{J}\right)\right]\boldsymbol{Y}$	$\frac{SQReg}{p'-1}$	$\sigma^2+\frac{\hat{\boldsymbol{\beta}}'\boldsymbol{X}_1'\boldsymbol{X}_1\hat{\boldsymbol{\beta}}}{\sigma^2}$	$\frac{QMReg}{QMErro}$
Erro	$n-p'$	$\boldsymbol{Y}'\left[\boldsymbol{I}_n-\boldsymbol{H}\right]\boldsymbol{Y}$	$\frac{SQErro}{n-p'}$	σ^2	

Tabela 1: Análise de variância para o teste F de $H_0: \hat{\boldsymbol{\beta}} = \mathbf{0}$.

$$H_0: \ \hat{\boldsymbol{\beta}} = \mathbf{0}, \text{ sendo } \hat{\boldsymbol{\beta}} = [\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_p]'$$
$$H_1 : \text{pelo menos um dos } \hat{\beta}\text{s é estatisticamente diferente de 0.}$$

Considerando que SQReg e SQErro são independentes (Ver Rencher port. p. 82, 83 e 127; Demetrio, p.84), então sob H_0, temos que

$$\frac{SQReg}{\sigma^2} \sim \chi^2_{(p'-1)}, \text{ (Central)} \tag{36}$$

uma vez que $\lambda = 0$. Logo,

$$F = \frac{\frac{\frac{SQReg}{\sigma^2}}{p'-1}}{\frac{\frac{SQErro}{\sigma^2}}{n-p'}} \sim F_{(p'-1,n-p')}\ (Central). \tag{37}$$

Se H_0 é falso então $F \sim F_{(p'-1,n-p',\lambda)}$, em que $\lambda = \frac{\hat{\beta}'\boldsymbol{X}_1'\boldsymbol{X}_1\hat{\boldsymbol{\beta}}}{2\sigma^2}$.

A decisão será rejeitarmos $H_0: \hat{\boldsymbol{\beta}} = \mathbf{0}$ se $F > F_{(\alpha,p'-1,n-p',\lambda)}$, em que $F_{(\alpha,p'-1,n-p',\lambda)}$ é o percentil de ordem 100α da distribuição F central.

Uma representação gráfica dessa decisão pode ser obtida:

Os resultados do teste podem ser apresentados resumidos na Tabela de análise de var. Se $H_0: \hat{\boldsymbol{\beta}} = \mathbf{0}$, as duas esperanças de quadrados médios são iguais a σ^2, e nós esperamos que o F seja próximo de 1. Se $H_0: \hat{\boldsymbol{\beta}} \neq \mathbf{0}$, então $E[SQReg/(p'-1)] > \sigma^2$, porque $\boldsymbol{X}_1'\boldsymbol{X}_1$ é positiva definida, e nós esperamos encontrar um F superior a 1. Só rejeitaremos H_0 para valores *muito grandes* de estatística F.

Após verificar a existência do modelo expresso em (5), parece natural que estejamos interessados em saber o quanto podemos confiar nesse modelo obtido.

Definição 5 (Coeficiente de determinação R^2)**.** *O coeficiente de determinação é definido por:*

$$R^2 = \frac{SQR}{SQT} = 1 - \frac{SQE}{SQT}, \tag{38}$$

sendo SQT a soma de quadrado total, SQR a soma de quadrado da regressão e SQE soma de quadrado do erro. □

O coeficiente de determinação indica a proporção da variação de Y que é explicada pela regressão. Note que $0 \leq R^2 \leq 1$. Entretanto, devemos ter algumas precauções sobre a interpretação do R^2:

- o R^2 aumenta com o número de observações na amostra diminui;
- o R^2 aumenta com a adição de uma nova variável ao modelo (isto não implica que o novo modelo é superior ao anterior);
- o R^2 aumenta com o aumento da amplitude de variação das variáveis regressoras;
- o R^2 somente é válido com o parâmetro β_0 incluído no modelo. A sua ausência inflaciona o R^2.

Assim, podemos observar que um valor grande de R^2 não implica necessariamente que o modelo está bem ajustado. Dessa forma, percebemos que o coeficiente de determinação não deve ser considerado sozinho, mas sempre aliado a outros diagnósticos do modelo. Pensando em corrigir os problemas de R^2, foi criado um outro coeficiente chamado de coeficiente de determinação ajustado.

Definição 6 (Coeficiente de determinação ajustado R^2_{aj})**.** *O coeficiente de determinação ajustado (R^2_{aj}) é definido por:*

$$R^2_{aj} = 1 - \frac{SQR}{SQT} = \frac{SQE}{SQT} = R^2 - \frac{1 - R^2}{n - p'}, \tag{39}$$

sendo SQT a soma de quadrado total, SQR a soma de quadrado da regressão, SQE soma de quadrado do erro, n o número de observações e p' o número de colunas da matriz $\boldsymbol{X}$. □

Características de R^2_{aj}:

- $R^2_{aj} < R^2$;
- pode ser negativo;
- não tem a interpretação prática de R^2;
- é mais justo para a comparação de modelos, isto é, quão mais próximo de 1, melhor o modelo.

A partir de um modelo ajustado, podemos está interessados em fazer previsões. Seja $\boldsymbol{X}_0 = [1, X_{01}, X_{02}, \ldots, X_{0p}]'$ uma escolha particular de $\boldsymbol{X} = [1, X_1, X_2, \ldots, X_p]'$, e Y_0 uma observação correspondente a $\boldsymbol{X}_0$. Então

$$Y_0 = \boldsymbol{X}_0'\boldsymbol{\theta} + \varepsilon \tag{40}$$

e

$$E[Y_0] = \boldsymbol{X}_0'\boldsymbol{\theta}. \tag{41}$$

Um estimado não viesado e de variância mínima para $E[Y_0]$ é $\widehat{E[Y_0]} = \boldsymbol{X}_0'\hat{\boldsymbol{\theta}}$. Dessa forma, um intervalo de confiança $(1-\alpha)100\%$ para $E[Y_0]$ é

$$\boldsymbol{X}_0'\hat{\boldsymbol{\theta}} \mp t_{\alpha/2,n-p'}[\boldsymbol{X}_0'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}_0 QME]^{1/2}. \tag{42}$$

Também é possível obter intervalos de confiança para uma observação futura Y_0 correspondente a $\boldsymbol{X}_0$. Este é chamado de intervalo de predição. Como Y_0 está expresso em (40), podemos predizer Y_0 por $\hat{Y}_0 = \boldsymbol{X}_0'\hat{\boldsymbol{\theta}}$, que também é um estimador de (41). Sabemos ainda que as variáveis aleatórias Y_0 e $\hat{Y}_0$ são independentes, pois Y_0 é uma observação futura a ser obtida independentemente das n observações usadas para computar $\hat{Y}_0 = \boldsymbol{X}_0'\hat{\boldsymbol{\theta}}$. Então, a variância de $Y_0 - \hat{Y}_0$ é

$$\begin{aligned}
Var[Y_0 - \hat{Y}_0] =& Var[Y_0] + Var[\hat{Y}_0] \\
=& \boldsymbol{I}\sigma^2 + Var[\boldsymbol{X}_0\hat{\boldsymbol{\theta}}] \\
=& \boldsymbol{I}\sigma^2 + \boldsymbol{X}_0 Var[\hat{\boldsymbol{\theta}}]\boldsymbol{X}_0' \\
=& \boldsymbol{I}\sigma^2 + \boldsymbol{X}_0\sigma^2(\boldsymbol{X}_0'\boldsymbol{X}_0)\boldsymbol{X}_0' \\
=& \sigma^2[1 + \boldsymbol{X}_0(\boldsymbol{X}_0'\boldsymbol{X}_0')] \\
=& \sigma^2[1 + \boldsymbol{X}_0(\boldsymbol{X}_0'\boldsymbol{X}_0')^{-1}\boldsymbol{X}'].
\end{aligned}$$

Pode ser mostrado que $E[Y_0 - \hat{Y}_0] = 0$ e que $\hat{\sigma}^2$ é independente de Y_0 e $\hat{Y}_0$. Assim, pode ser mostrado que

$$Y_0 - \hat{Y}_0 \sim N(0, \sigma^2[1 + \boldsymbol{X}_0(\boldsymbol{X}_0'\boldsymbol{X}_0')^{-1}\boldsymbol{X}'])$$

e

$$P\left(-Z_{\alpha/2}\sqrt{\sigma^2[1 + \boldsymbol{X}_0(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}_0']} \leq Y_0 - \hat{Y}_0 \leq Z_{\alpha/2}\sqrt{\sigma^2[1 + \boldsymbol{X}_0(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}_0']}\right) = 1 - \alpha,$$

sendo $Z_{\alpha/2}$ o quantil superior da distribuição normal padrão. Se considerarmos σ^2 desconhecido e substituirmos pelo seu estimador $\hat{\sigma}^2$, então

$$P\left(-t_{\alpha/2,n-p-1}\sqrt{\hat{\sigma}^2[1 + \boldsymbol{X}_0(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}_0']} \leq Y_0 - \hat{Y}_0 \leq t_{\alpha/2,n-p-1}\sqrt{\hat{\sigma}^2[1 + \boldsymbol{X}_0(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}_0']}\right) = 1 - \alpha,$$

sendo $t_{\alpha/2,n-p-1}$ o quantil superior da distribuição t com $n-p-1$ graus de liberdade. Portanto, um intervalo de predição para $\hat{Y}_0$ é

$$\hat{Y}_0 \mp t_{\alpha/2,n-p-1}\sqrt{\hat{\sigma}^2[1 + \boldsymbol{X}_0(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}_0']}.$$

Podemos observar que toda inferência realizada sobre o modelo expresso em (5) se baseou na pressuposição:

$$\boldsymbol{\varepsilon} \overset{\text{iid}}{\sim} N_n(\boldsymbol{0}, \boldsymbol{I}\sigma^2), \tag{43}$$

isso implica:

- a esperança dos ε_i é igual a zero;

- a variança dos ε_i é homocedástico;
- independência dos ε_i, para $i = 1, 2, \ldots, n$;
- os ε_i têm distribuição normal.

Entretanto, isso nem sempre é observado na prática. Quando isso ocorre, a análise resultante pode levar a conclusões duvidosas. Para verificar essas pressuposições, usamos um estimador para o erro ε_i, baseado no modelo estimado $\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\theta}}$, que é o resíduo observado.

Definição 7 (Resíduo ordinário)**.** *Seja o modelo expresso em* (5)*, o resíduo observado é dado por:*

$$\hat{\boldsymbol{\varepsilon}} = \boldsymbol{Y} - \hat{\boldsymbol{Y}}, \tag{44}$$

sendo $\hat{Y} = \boldsymbol{X}\hat{\boldsymbol{\theta}}$ *e* $\hat{\theta}$ *o estimador de mínimos quadrados para* θ*.* □

Entretanto, observe que

$$\begin{aligned} E[\hat{\boldsymbol{\varepsilon}}] =& E[\boldsymbol{Y}] - E[\hat{\boldsymbol{Y}}] \\ =& \boldsymbol{X}\boldsymbol{\theta} - E[\boldsymbol{X}\hat{\boldsymbol{\theta}}] \\ =& \boldsymbol{X}\boldsymbol{\theta} - E[\boldsymbol{X}\hat{\boldsymbol{\theta}}] \\ =& \boldsymbol{X}\boldsymbol{\theta} - E[\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}] \\ =& \boldsymbol{X}\boldsymbol{\theta} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'E[\boldsymbol{Y}] \\ =& \boldsymbol{X}\boldsymbol{\theta} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\theta} \\ =& \boldsymbol{X}\boldsymbol{\theta} - \boldsymbol{X}\boldsymbol{\theta} = \boldsymbol{0}. \end{aligned}$$

A matriz $\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ é chamada de matriz chapéu, denotada por $\boldsymbol{H}$. Esse nome é devido a essa matriz transformar $\boldsymbol{Y}$ em $\hat{\boldsymbol{Y}}$, isto é,

$$\begin{aligned} \hat{\boldsymbol{Y}} =& \boldsymbol{X}\hat{\boldsymbol{\theta}} \\ =& \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y} \\ =& \boldsymbol{H}\boldsymbol{Y}. \end{aligned}$$

Com o conhecimento da matriz $\boldsymbol{H}$, poderemos reescrever o resíduo observado como:

$$\hat{\boldsymbol{\varepsilon}} = \boldsymbol{Y} - \hat{\boldsymbol{Y}} = \boldsymbol{Y} - \boldsymbol{H}\boldsymbol{Y} = \boldsymbol{P}\boldsymbol{Y},$$

sendo $\boldsymbol{P} = \boldsymbol{I} - \boldsymbol{H}$ uma matriz simétrica e idempotente. A simetria é verificada da seguinte forma:

$$\begin{aligned} \boldsymbol{H}' =& [\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}']' \\ =& \boldsymbol{X}[(\boldsymbol{X}'\boldsymbol{X})']^{-1}\boldsymbol{X}' \\ =& \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' \\ =& \boldsymbol{H}. \end{aligned}$$

Assim,

$$\begin{aligned}\boldsymbol{P}' =&(\boldsymbol{I}-\boldsymbol{H})'\\ =&\boldsymbol{I}'-\boldsymbol{H}'\\ =&\boldsymbol{I}-\boldsymbol{H}\\ =&\boldsymbol{P}.\end{aligned}$$

Observe que a matriz $\boldsymbol{H}$ é idempotente, isto é,

$$\begin{aligned}\boldsymbol{H}^2 =&[\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'][\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}']'\\ =&\boldsymbol{X}\underbrace{(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}}_{\boldsymbol{I}}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\\ =&\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'=\boldsymbol{H}.\end{aligned}$$

Assim,

$$\begin{aligned}\boldsymbol{P}^2 =&(\boldsymbol{I}-\boldsymbol{H})(\boldsymbol{I}-\boldsymbol{H})'\\ =&\boldsymbol{I}-2\boldsymbol{H}+\boldsymbol{H}^2\\ =&\boldsymbol{I}-2\boldsymbol{H}+\boldsymbol{H}\\ =&\boldsymbol{I}-\boldsymbol{H}\\ =&\boldsymbol{P}.\end{aligned}$$

A matriz de covariâncias de $\hat{\boldsymbol{\varepsilon}}$ é dada por:

$$\begin{aligned}Cov[\hat{\boldsymbol{\varepsilon}}] =&Var[\boldsymbol{Y}]+Var[\hat{\boldsymbol{Y}}]-2Cov[\boldsymbol{Y},\hat{\boldsymbol{Y}}]\\ =&\boldsymbol{I}\sigma^2+Var[\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}]-2Cov[\boldsymbol{Y},\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}]\\ =&\boldsymbol{I}\sigma^2+\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'Var[\boldsymbol{Y}]-2\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'Cov[\boldsymbol{Y},\boldsymbol{Y}]\\ =&\boldsymbol{I}\sigma^2+\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\sigma^2-2\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'Var[\boldsymbol{Y}]\\ =&\boldsymbol{I}\sigma^2+\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\sigma^2-2\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\sigma^2\\ =&\boldsymbol{I}\sigma^2-\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\sigma^2\\ =&(\boldsymbol{I}-\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}')\sigma^2\\ =&(\boldsymbol{I}-\boldsymbol{H})\sigma^2.\end{aligned}$$

Portanto, apesar de $\hat{\boldsymbol{\varepsilon}}$ ter distribuição normal multivariada com o vetor de médias igual a 0, os elementos da diagonal h_{ii} (leverage) da matriz de covariâncias são diferentes entre si, isto é, $\hat{\boldsymbol{\varepsilon}} \sim N_n[\boldsymbol{0}, \boldsymbol{P}\sigma^2]$, e os elementos fora da diagonal são diferentes de zero, portanto os erros observados não são independentes.

Para resolver o problema da heterocedasticidade dos resíduos observados, foi usado a estudentização de cada erro observado, sendo que a variância dos resíduos foram estimados pelos elementos da diagonal de $\boldsymbol{P}S^2$.

Definição 8 (Resíduo estudentizado internamente)**.** *Seja o modelo expresso em* (5) *e o resíduo observado expresso em* (44)*, definição* (7)*, então o resíduo estudentizado internamente é definido por:*

$$r_i = \frac{\hat{\varepsilon}_i}{\sqrt{S^2(1-h_{ii})}}, \quad i = 1, 2, \ldots, n, \tag{45}$$

em que h_{ii} é o i-ésimo elemento da diagonal de $\boldsymbol{H}$, e S^2 expresso em (25)*.* □

(**??**) Observamos que r_i se comporta como uma variável aleatória que tem distribuição t exceto pelo fato que o numerador e o denominador não serem independentes. Como solução para esse problema, definimos outro tipo de erro.

Definição 9 (Resíduo estudentizado externamente)**.** *Seja o modelo expresso em* (5) *e o resíduo observado expresso em* (44)*, definição* (7)*, então o resíduo estudentizado externamente é definido por:*

$$r_i^* = \frac{\hat{\varepsilon}_i}{\sqrt{S^2_{(i)}(1-h_{ii})}}, \quad i = 1, 2, \ldots, n, \tag{46}$$

em que h_{ii} é o i-ésimo elemento da diagonal de $\boldsymbol{H}$, e $S^2_{(i)}$ é o estimador de σ^2, dado por:

$$S^2_{(i)} = \frac{(n-p')S^2 - \hat{\varepsilon}_i^2/(1-h_{ii})}{n-p'-1} = S^2\left(\frac{n-p'-1}{n-p'-r_i^2}\right). \tag{47}$$

O índice (i) indica que a i-ésimo observação será omitida para estimar σ^2. □

Cada resíduo r_i^* tem distribuição t com $n-p'-1$ graus de liberdade quando a normalidade de ε se mantém. Uma forma auxiliar para calcular $S^2_{(i)}$ e r_i^* sem retornar ao modelo de regressão com as observações omitidas, **??**, p. 20) menciona a relação:

$$r_i^* = r_i\left(\frac{n-p'-1}{n-p'-r_i^2}\right)^{1/2}, \tag{48}$$

em que:

- r_i é o i-ésimo resíduo estudentizado internamente;
- n é o número de observações;
- p' é o número de parâmetros.

Os erros r_i e r_i^* são chamados por alguns autores de resíduo padronizado e resíduo estudentizado. Para detalhes, ver **??**, p. 342).

Características desses três tipos de erros:

- os erros $\hat{\varepsilon}_i$ têm variâncias heterogêneas;
- os erros $\hat{\varepsilon}_i$, r_i e r_i^* não são independentes;
- os erros $\hat{\varepsilon}_i$, r_i e r_i^* podem inflacionar um possível *outlier*, isto é, um ponto atípico pode ter um pequeno um resíduo relativamente pequeno;
- mesmo que a pressuposição de normalidade não seja atendida, os resíduos $\hat{\varepsilon}_i$ de mínimos quadrados tendem a ter um melhor ajuste a distribuição normal do que os erros ε_i;
- testes exatos para os resíduos observados não estão disponíveis; aproximações e julgamentos subjetivos podem ser usados;

- os erros r_i e r_i^* são preferíveis para essa avaliação da homocedasticidade, principalmente este último, pois sob normalidade r_i^* tem distribuição t exata, sendo uma vantagem para avaliar a variância dos erros.

Mesmo com todos esses problemas, esses três tipos de erros têm provado ser úteis para detectar modelos inadequados e *outliers*.

Outra classe de resíduos são os resíduos recursivos, denotado por w_r, para $r = p' + 1, p' + 2, \ldots, n$, tal que

$$w_r \overset{\text{iid}}{\sim} N(0, \sigma^2).$$

Os resíduos recursivos são computados de uma sequência de regressões iniciada com uma base de p' observações (p' número de parâmetros a ser estimado) e posteriormente adicionado uma observação a cada passo. A equação de regressão computada em cada passo é usada para calcular o resíduo da próxima observação a ser adicionada. A sequência continua até que o último resíduo seja computado. Haverá $n - p'$ resíduos recursivos; os resíduos das p' primeiras observações serão iguais a zero.

Definição 10 (Resíduos recursivos)**.** *Considere o modelo expresso em* (5)*, sendo $\boldsymbol{y}_r$ e $\boldsymbol{x}_r'$ as r-ésimas linhas de $\boldsymbol{Y}$ e $\boldsymbol{X}$, respectivamente. Considere ainda $\boldsymbol{X}_r$ as primeiras r linhas de $\boldsymbol{X}$ e $\hat{\boldsymbol{\theta}}_r$ o estimador de quadrados mínimos usando as primeiras r observações. Então, o resíduo recursivo é definido por*

$$w_r = \frac{\boldsymbol{y}_r - \boldsymbol{x}_r'\hat{\boldsymbol{\theta}}_{r-1}}{[1 - \boldsymbol{x}_r'(\boldsymbol{X}_{r-1}'\boldsymbol{X}_{r-1})^{-1}\boldsymbol{x}_r]^{1/2}}, \quad r = p' + 1, \ldots, n. \tag{49}$$

□

1 Detecção de Pontos influentes ou outliers

Definição 11 (Ponto influente)**.** *Uma observação é um ponto influente se a sua exclusão causa uma mudança substancial nos valores ajustados do modelo de regressão.*

testes para avaliar: ⟨http://www.portalaction.com.br/analise-de-regressao/343-pontos-influentes

Definição 12. *Outlier Uma observação extrema cujo comportamento se apresenta diferente dos demais pontos.*

⟨http://www.portalaction.com.br/analise-de-regressao/34-diagnostico-de-outliers⟩

Além de diagnosticar heteroscedasticidade, o gráfico de resíduos versus valores ajustados também auxilia na detecção de pontos atípicos.

Se um outlier for influente, ele interfere sobre a função de regressão ajustada (a inclusão ou não do ponto modifica substancialmente os valores ajustados).

Mas uma observação ser considerada um outlier não quer dizer que consequentemente é um ponto influente. Por isso, um ponto pode ser um outlier em relação a Y ou aos X, e pode ou não ser um ponto influente.