1. Modelos de Regressão Linear

Modelos de regressão linear relacionam uma variável dependente ou variável de resposta, Y, a uma ou mais variáveis explicativas (também chamadas covariáveis ou variáveis independentes), X. A estimação destes modelos é geralmente feita por mínimos quadrados ordinários, e os estimadores obtidos por este algoritmo são ótimos sob certas condições, dadas pelo teorema de Gauss e Markov.

O caso mais simples, em que consideramos apenas uma variável explicativa, é estudado a seguir, e muitas das conclusões podem ser generalizadas para o caso de p covariáveis.

1.1. Regressão Linear Simples

Considere o modelo linear:

$$Y = \beta_0 + \beta_1.x + \varepsilon \qquad (1.1.1.)$$

onde ε representa uma variável aleatória (não necessariamente Normal) com média zero e variância constante σ^2 . Supomos que x é fixo (não aleatório) e β_0, β_1 representam parâmetros desconhecidos a serem estimados.

Dado um conjunto de n pares de observações, (x_1, y_1) , (x_2, y_2) , ..., (x_n, y_n) como ajustar o modelo (1.1.1.) a estes pares. Esta questão se resume, num nível operacional, a encontrar estimadores b_0 e b_1 dos parâmetros desconhecidos β_0 e β_1 em (1.1.1.)

A solução usual para este problema é a estimação por mínimos quadrados, onde os estimadores b_0 e b_1 de forma a minimizar:

$$S(b_0 , b_1) = \Sigma(Y_i - b_0 - b_1.x_i)^2 \qquad (1.1.2.)$$

Os estimadores b_0 e b_1 são obtidos por diferenciação da equação (1.1.2) com relação a b_0 e b_1, o que resulta em :

$$\frac{\partial S}{\partial b_0} = -2\sum_{i=1}^{n}\left(Y_i - b_0 - b_1.x_i\right) = 0$$

$$\frac{\partial S}{\partial b_1} = -2\sum_{i=1}^{n} x_i\left(Y_i - b_0 - b_1.x_i\right) = 0$$

Ao expandirmos estas duas expressões encontramos :

$$\sum_{i=1}^{n} Y_i = nb_0 + b_1.\sum_{i=1}^{n} x_i \Rightarrow n\overline{Y} = nb_0 + nb_1\overline{X} \Rightarrow \overline{Y} = b_0 + b_1\overline{X} \qquad (1.1.3.)$$

e

$$\sum_{i=1}^{n} x_i Y_i = b_0\sum_{i=1}^{n} x_i + b_1\sum_{i=1}^{n} x_i^2 \qquad (1.1.4.)$$

De (1.1.3.) encontramos:

$$\overline{Y} = b_0 + b_1\overline{X} \Rightarrow b_0 = \overline{Y} - b_1\overline{X}$$

e substituindo este resultado em (1.1.4.) leva a:

$$\sum_{i=1}^{n} x_i Y_i = \left(\overline{Y} - b_1 \overline{X}\right)\sum_{i=1}^{n} x_i + b_1 \sum_{i=1}^{n} x_i^2 \Rightarrow \sum_{i=1}^{n} x_i Y_i - \left(\overline{Y}\sum_{i=1}^{n} x_i\right) = b_1\left(\sum_{i=1}^{n} x_i^2 - \overline{X}\sum_{i=1}^{n} x_i\right)$$

$$\Rightarrow b_1 = \frac{\sum_{i=1}^{n} x_i Y_i - n\overline{XY}}{\left(\sum_{i=1}^{n} x_i^2 - \overline{X}\sum_{i=1}^{n} x_i\right)} = \frac{\sum_{i=1}^{n} x_i Y_i - n\overline{XY}}{\sum_{i=1}^{n} x_i^2 - n\overline{X}^2} = \frac{\sum_{i=1}^{n}\left(x_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sum_{i=1}^{n}\left(x_i - \overline{X}\right)^2} = \frac{SXY}{SXX} \qquad (1.1.5.)$$

Alguns resultados interessantes decorrem das expressões (1.1.3) a (1.1.5).

A reta estimada por mínimos quadrados passa pelas médias amostrais de X e Y, conforme indicado em (1.1.3.)

Os estimadores b_0 e b_1 são **funções lineares** dos Y's, e portanto suas propriedades estatísticas podem ser determinadas a partir das características dos Y's.

Teorema 1.1.1.
Os estimadores b_0 e b_1 são não tendenciosos para os parâmetros desconhecidos β_0 e β_1 .

Demonstração
Da equação do modelo (1.1.1.) segue que : $E(Y) = \beta_0 + \beta_1.x$ e $VAR(Y) = VAR(\varepsilon) = \sigma^2$, uma constante. Daí concluímos que:

$$E\left(Y_i\right) = \beta_0 + \beta_1 . x_i$$

$$VAR\left(Y_i\right) = \sigma^2$$

$$\Rightarrow E\left(\overline{Y}\right) = E\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) = E\left(\frac{1}{n}\sum_{i=1}^{n}\left(\beta_0 + \beta_1 . x_i\right)\right) = \frac{1}{n}\left(n\beta_0 + n\beta_1\overline{X}\right) = \beta_0 + \beta_1\overline{X}$$

$$\Rightarrow VAR\left(\overline{Y}\right) = VAR\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) = \frac{\sigma^2}{n}$$

Então:

$$E\left(b_1\right) = E\left(\frac{SXY}{SXX}\right) = \frac{1}{SXX}E\left(SXY\right) = \frac{1}{SXX}E\left(\sum_{i=1}^{n}\left(x_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)\right) = \frac{1}{SXX}\sum_{i=1}^{n} E\left(\left(x_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)\right) =$$

$$= \frac{1}{SXX}\sum_{i=1}^{n} E\left(x_i Y_i - \overline{X}Y_i - x_i\overline{Y} + \overline{XY}\right) = \frac{1}{SXX}\sum_{i=1}^{n}\left\{E\left(x_i Y_i\right) - \overline{X}E\left(Y_i\right) - x_i E\left(\overline{Y}\right) + \overline{X}E\left(\overline{Y}\right)\right\} =$$

$$= \frac{1}{SXX}\left\{\sum_{i=1}^{n} E\left(x_i Y_i\right) - \overline{X}\sum_{i=1}^{n}\left(\beta_0 + \beta_1 x_i\right) - \sum_{i=1}^{n} x_i\left(\beta_0 + \beta_1\overline{X}\right) + n\overline{X}\left(\beta_0 + \beta_1\overline{X}\right)\right\}$$

$$E(b_1)=\frac{1}{SXX}\left\{\sum_{i=1}^{n}\left(x_iE(Y_i)\right)-\overline{X}\sum_{i=1}^{n}\left(\beta_0+\beta_1x_i\right)-\sum_{i=1}^{n}x_i\left(\beta_0+\beta_1\overline{X}\right)+n\overline{X}\left(\beta_0+\beta_1\overline{X}\right)\right\}=$$

$$=\frac{1}{SXX}\left\{\sum_{i=1}^{n}\left(x_i\left(\beta_0+\beta_1x_i\right)\right)-\overline{X}\sum_{i=1}^{n}\left(\beta_0+\beta_1x_i\right)-\sum_{i=1}^{n}x_i\left(\beta_0+\beta_1\overline{X}\right)+n\overline{X}\left(\beta_0+\beta_1\overline{X}\right)\right\}=$$

$$=\frac{1}{SXX}\left\{\beta_0\left[\sum_{i=1}^{n}x_i-n\overline{X}-\sum_{i=1}^{n}x_i+n\overline{X}\right]+\beta_1\left[\sum_{i=1}^{n}x_i^2-\overline{X}\sum_{i=1}^{n}x_i-\overline{X}\sum_{i=1}^{n}x_i+n\overline{X}^2\right]\right\}=$$

$$=\frac{1}{SXX}\left\{\beta_1\left[\sum_{i=1}^{n}x_i^2-\overline{X}\left(n\overline{X}\right)-\overline{X}\left(n\overline{X}\right)+n\overline{X}^2\right]\right\}=\frac{1}{SXX}\left\{\beta_1\left[\sum_{i=1}^{n}x_i^2-n\overline{X}^2\right]\right\}=\frac{1}{SXX}\left\{\beta_1\left[\sum_{i=1}^{n}\left(x_i-\overline{X}\right)^2\right]\right\}=$$

$$=\frac{1}{SXX}\beta_1\left(SXX\right)=\beta_1$$

Analogamente:

$$E\left(b_0\right)=E\left(\overline{Y}-b_1\overline{X}\right)=E\left(\overline{Y}\right)-\overline{X}E\left(b_1\right)=\beta_0+\beta_1\overline{X}-\overline{X}\beta_1=\beta_0$$

O próximo teorema estabelece as variâncias dos estimadores por mínimos quadrados.

Teorema 1.1.2.
As variâncias dos estimadores b_0 e b_1 são dadas por:

$$VAR\left(b_0\right)=\frac{\sigma^2\sum_{i=1}^{n}x_i^2}{n\left(SXX\right)}=\frac{\sigma^2\sum_{i=1}^{n}x_i^2}{n\left\{\sum_{i=1}^{n}\left(x_i-\overline{X}\right)^2\right\}}$$

e

$$VAR\left(b_1\right)=\frac{n\sigma^2}{n\left(SXX\right)}=\frac{\sigma^2}{SXX}=\frac{\sigma^2}{\sum_{i=1}^{n}\left(x_i-\overline{X}\right)^2}$$

Além disso , a covariância entre estes estimadores é:

$$COV\left(b_0,b_1\right)=\frac{-\overline{X}}{SXX}\sigma^2$$

Demonstração
Note que o estimador b_1 pode ser escrito como:

$$b_1=\left(\frac{SXY}{SXX}\right)=\frac{1}{SXX}\left(\sum_{i=1}^{n}\left(x_i-\overline{X}\right)\left(Y_i-\overline{Y}\right)\right)=\frac{1}{SXX}\left(\sum_{i=1}^{n}\left(x_i-\overline{X}\right)Y_i\right)$$

Assim:

$$VAR(b_1) = \frac{1}{(SXX)^2} VAR\left(\sum_{i=1}^{n}(x_i - \overline{X})Y_i\right) = \frac{1}{(SXX)^2}\left(\sum_{i=1}^{n}(x_i - \overline{X})^2 VAR(Y_i)\right) = \frac{1}{(SXX)^2}\sum_{i=1}^{n}(x_i - \overline{X})^2\sigma^2 =$$

$$= \frac{\sigma^2(SXX)}{(SXX)^2} = \frac{\sigma^2}{SXX}$$

Sejam:

$$c_i = \frac{x_i - \overline{X}}{SXX} \Rightarrow \sum_{i=1}^{n} c_i = 0 \text{ e } \sum_{i=1}^{n} c_i^2 = \frac{1}{SXX} \text{ e } \sum_{i=1}^{n} c_i x_i = 1$$

$$b_1 = \sum_{i=1}^{n} c_i Y_i$$

Mas, pode-se provar que:

$$COV(\overline{Y}, b_1) = 0$$

Pois:

$$COV(\overline{Y}, b_1) = COV\left(\overline{Y}, \sum_{i=1}^{n} c_i Y_i\right) = COV\left(\frac{1}{n}\sum_{i=1}^{n} Y_i, \sum_{i=1}^{n} c_i Y_i\right) =$$

$$= \frac{1}{n} COV(Y_1 + Y_2 + ... + Y_n, c_1 Y_1 + c_2 Y_2 + ... + c_n Y_n) = \frac{1}{n}\sum_{i=1}^{n} COV(Y_i, c_i Y_i) = \frac{1}{n}\sum_{i=1}^{n} c_i VAR(Y_i) =$$

$$= \frac{1}{n}\sum_{i=1}^{n} c_i \sigma^2 = \frac{\sigma^2}{n}\sum_{i=1}^{n} c_i = 0$$

Já que Y_i e Y_j são descorrelatados. Daí:

$$VAR(b_0) = VAR(\overline{Y} - b_1\overline{X}) = VAR(\overline{Y}) + \overline{X}^2 . VAR(b_1) - 2\overline{X} . COV(\overline{Y}, b_1) =$$

$$= VAR(\overline{Y}) + \overline{X}^2 . VAR(b_1) = \frac{\sigma^2}{n} + \overline{X}^2\left(\frac{\sigma^2}{SXX}\right) = \sigma^2\left(\frac{1}{n} + \frac{\overline{X}^2}{SXX}\right) = \sigma^2\left(\frac{SXX + n\overline{X}^2}{n.SXX}\right) =$$

$$= \sigma^2\left(\frac{\sum_{i=1}^{n} x_i^2 - n\overline{X}^2 + n\overline{X}^2}{n.SXX}\right) = \sigma^2\left(\frac{\sum_{i=1}^{n} x_i^2}{n.SXX}\right)$$

Finalmente:

$$COV(b_0, b_1) = COV(\overline{Y} - b_1\overline{X}, b_1) = COV(\overline{Y}, b_1) - COV(b_1\overline{X}, b_1) = 0 - \overline{X} . VAR(b_1) = -\overline{X}\frac{\sigma^2}{SXX}$$

Do teorema 1.1.2. percebemos que as variâncias dos estimadores b_0 e b_1 dependem dos x_i's (que são conhecidos) e de σ^2 , que também é desconhecido. Logo, o próximo passo é encontrar um estimador para σ^2 , que será dado por um múltiplo da Soma do Quadrado dos Resíduos (RSS), definida a seguir.

Definição 1.1.3. (Resíduos)
Os resíduos e_i de uma regressão linear são a diferença entre cada Y_i e o correspondente valor ajustado pela regressão, isto é:

$$e_i = Y_i - \widetilde{Y}_i = Y_i - b_0 - b_1 x_i$$

Note que:

$$\sum_{i=1}^{n} e_i = 0$$

A soma dos resíduos é zero, mas não a soma dos seus quadrados, como definida a seguir!

Definição 1.1.4. (Soma do Quadrado dos Resíduos)
A soma do quadrado dos resíduos, denotada por RSS, é dada por :

$$RSS = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(Y_i - \widetilde{Y}_i\right)^2 = \sum_{i=1}^{n} \left(Y_i - b_0 - b_1 x_i\right)^2$$

A soma do quadrado dos resíduos pode ser escrita em termos do estimador b_1 e das somas de quadrados SXX e SYY como (prove !!!!):

$$RSS = SYY - \frac{(SXY)^2}{SXX} = SYY - b_1^2 (SXX)$$

A soma do quadrado dos resíduos é utilizada na construção de um estimador da variância dos erros, σ^2 . Este estimador, por sua vez, é usado para estimar as variâncias de b_0 e b_1 (estimadores dos parâmetros desconhecidos β_0 e β_1). O estimador da variância σ^2 é indicado a seguir.

Seja RSS a soma do quadrado dos resíduos, já definida. A variância dos erros ε_i (desconhecida) é σ^2 , e será estimada por :

$$s^2 = \frac{RSS}{n-2}$$

Teorema 1.1.5.

O estimador s^2 definido acima é um estimador não tendencioso de σ^2 .

Demonstração
Inicialmente considere a expressão para o i-ésimo resíduo e calcule a sua variância. Note que a média do i-ésimo resíduo é zero (por que?) e assim sua variância é apenas $E(e_i^2)$.

Mas:

$$e_i = Y_i - \widetilde{Y}_i = Y_i - (b_0 + b_1 x_i) = \beta_0 + \beta_1 x_i + \varepsilon_i - (b_0 + b_1 x_i) = (\beta_0 - b_0) + x_i(\beta_1 - b_1) + \varepsilon_i$$

$$e_i^{\,2} = \left((\beta_0 - b_0) + x_i(\beta_1 - b_1) + \varepsilon_i\right)^2 = (\beta_0 - b_0)^2 + x_i^2(\beta_1 - b_1)^2 + 2x_i(\beta_0 - b_0)(\beta_1 - b_1) + \varepsilon_i^2$$

$$+2\varepsilon_i\left((\beta_0 - b_0) + x_i(\beta_1 - b_1)\right)$$

Tomando o valor esperado:

$$E\left(e_i^{\,2}\right) = E\left((\beta_0 - b_0)^2\right) + x_i^2 E\left((\beta_1 - b_1)^2\right) + 2x_i E\left((\beta_0 - b_0)(\beta_1 - b_1)\right) + E\left(\varepsilon_i^2\right)$$

$$+2E\left(\varepsilon_i\left((\beta_0 - b_0) + x_i(\beta_1 - b_1)\right)\right)$$

$$E\left(e_i^{\,2}\right) = E\left((\beta_0 - b_0)^2\right) + x_i^2 E\left((\beta_1 - b_1)^2\right) + 2x_i E\left((\beta_0 - b_0)(\beta_1 - b_1)\right) + E\left(\varepsilon_i^2\right)$$

$$+2E\left(\varepsilon_i(\beta_0 - b_0)\right) + x_i\varepsilon_i\left((\beta_1 - b_1)\right)$$

Daí:

$$E\left(e_i^{\,2}\right) = VAR(b_0) + x_i^2 VAR(b_1) + 2x_i COV(b_0, b_1) + VAR\left(\varepsilon_i^2\right)$$

$$+2.COV(\varepsilon_i, -b_0) + 2x_i COV(\varepsilon_i, -b_1)$$

pois $E(\varepsilon_i) = 0$ e $E\left(\varepsilon_i^{\,2}\right) = VAR(\varepsilon_i) = \sigma^2$

Mas:

$$VAR(b_0) = \sigma^2\left(\frac{\sum_{i=1}^{n} x_i^2}{n.SXX}\right) \quad , \quad VAR(b_1) = \frac{\sigma^2}{SXX} \quad \text{e} \quad COV(b_0, b_1) = -\overline{X}\frac{\sigma^2}{SXX}$$

Logo, ao substituirmos estes resultados na expressão para $E(e_i^2)$ encontramos:

$$E\left(e_i^{\,2}\right) = \sigma^2\left(\frac{\sum_{i=1}^{n} x_i^2}{n.SXX}\right) + x_i^2.\left(\frac{\sigma^2}{SXX}\right) + 2x_i\left(\frac{-\overline{X}\sigma^2}{SXX}\right) + \sigma^2 - 2.COV(\varepsilon_i, b_0) - 2x_i COV(\varepsilon_i, b_1)$$

Mas: $SXX = \sum_{i=1}^{n} x_i^2 - n.\overline{X}^2 \Rightarrow \sum_{i=1}^{n} x_i^2 = SXX + n.\overline{X}^2$

Daí:

$$E\left(e_i^{\,2}\right) = \sigma^2\left(\frac{SXX + n\overline{X}^2}{n.SXX}\right) + x_i^2.\left(\frac{\sigma^2}{SXX}\right) + 2x_i\left(\frac{-\overline{X}\sigma^2}{SXX}\right) + \sigma^2 - 2.COV(\varepsilon_i, b_0) - 2x_i COV(\varepsilon_i, b_1) =$$

$$= \frac{\sigma^2}{SXX}\left(\frac{SXX}{n} + \overline{X}^2 + x_i^2 - 2x_i\overline{X} + SXX\right) - 2.COV(\varepsilon_i, b_0) - 2x_i COV(\varepsilon_i, b_1) =$$

$$= \frac{\sigma^2}{SXX}\left(\frac{(n+1)SXX}{n} + \left(x_i - \overline{X}\right)^2\right) - 2.COV(\varepsilon_i, b_0) - 2x_i COV(\varepsilon_i, b_1)$$

Finalmente:

$$COV(\varepsilon_i, b_1) = COV(y_i - \beta_0 - \beta_1, b_1) = COV(y_i, b_1) = COV\left(y_i, \sum_{i=1}^{n} c_i y_i\right) = c_i VAR(y_i) =$$

$$= c_i \sigma^2 = \sigma^2\left(\frac{x_i - \overline{X}}{SXX}\right) = \frac{\sigma^2}{SXX}(x_i - \overline{X})$$

E também:

$$COV(\varepsilon_i, b_0) = COV(y_i - \beta_0 - \beta_1, b_0) = COV(y_i, b_0) = COV(y_i, \overline{Y} - b_1\overline{X}) =$$

$$= COV(y_i, \overline{Y}) - \overline{X}.COV(y_i, b_1) = COV\left(y_i, \frac{1}{n}\sum_{i=1}^{n} y_i\right) - \overline{X}.\sigma^2\left(\frac{x_i - \overline{X}}{SXX}\right) =$$

$$= \frac{1}{n}VAR(y_i) - \overline{X}.\sigma^2\left(\frac{x_i - \overline{X}}{SXX}\right) = \frac{\sigma^2}{n} - \sigma^2\left(\frac{\overline{X}(x_i - \overline{X})}{SXX}\right) = \frac{\sigma^2}{SXX}\left(\frac{SXX}{n} - \overline{X}(x_i - \overline{X})\right)$$

Substituindo estas expressões na expressão para $E(e_i^2)$ encontramos:

$$E(e_i^2) = \frac{\sigma^2}{SXX}\left(\frac{(n+1)SXX}{n} + (x_i - \overline{X})^2\right) - 2.\frac{\sigma^2}{SXX}\left(\frac{SXX}{n} - \overline{X}(x_i - \overline{X})\right) - 2x_i\sigma^2\left(\frac{x_i - \overline{X}}{SXX}\right) =$$

$$= \frac{\sigma^2}{SXX}\left(\frac{(n+1)SXX}{n} + (x_i - \overline{X})^2 - 2\left(\frac{SXX}{n}\right) - 2\left(-\overline{X}(x_i - \overline{X})\right) - 2\left(x_i(x_i - \overline{X})\right)\right) =$$

$$= \frac{\sigma^2}{SXX}\left(\frac{(n+1-2)SXX}{n} + (x_i - \overline{X})^2 - 2(x_i - \overline{X})(x_i - \overline{X})\right) =$$

$$= \frac{\sigma^2}{SXX}\left(\frac{(n-1)SXX}{n} + (x_i - \overline{X})^2 - 2(x_i - \overline{X})^2\right) = \frac{\sigma^2}{SXX}\left(\frac{(n-1)SXX}{n} - (x_i - \overline{X})^2\right)$$

Mas, o valor esperado da RSS é apenas:

$$E(RSS) = E\left(\sum_{i=1}^{n} e_i^2\right) = \sum_{i=1}^{n} E(e_i^2) = \sum_{i=1}^{n} \frac{\sigma^2}{SXX}\left(\frac{(n-1)SXX}{n} - (x_i - \overline{X})^2\right) =$$

$$= \frac{\sigma^2}{SXX}\left((n-1)SXX - \sum_{i=1}^{n}(x_i - \overline{X})^2\right) =$$

$$= \frac{\sigma^2}{SXX}\left((n-1)SXX - SXX\right) = \frac{\sigma^2}{SXX}(n-2)SXX = (n-2)\sigma^2$$

Logo :

$$E(s^2) = E\left(\frac{RSS}{n-2}\right) = \frac{(n-2)\sigma^2}{n-2} = \sigma^2$$

E assim s^2 é não tendencioso para σ^2 .

O teorema a seguir é demonstrado em livros mais avançados, e o teorema 1.1.5. decorre trivialmente dele.

Teorema 1.1.6.
Se os erros ε_i na equação (1.1.1) são iid N(0, σ^2) então:

$$\frac{RSS}{\sigma^2} = \frac{(n-2)s^2}{\sigma^2} \approx \chi^2_{n-2}$$

Uma conseqüência óbvia deste teorema é o teorema 1.1.5.

O teorema 1.1.6. é importante pois nos permite criar estatísticas t a partir dos estimadores de mínimos quadrados de β_0 e β_1 .

Note que, se os erros ε_i são iid N(0, σ^2), os Y_i 's são também independentes e Normais, com médias $\beta_0 + \beta_1 x_i$ e variância (comum) σ^2 .

Especificamente:

b_0 é Normal com média β_0 e variância $VAR(b_0) = \sigma^2 \left(\frac{\sum_{i=1}^{n} x_i^2}{n.SXX} \right)$

b_1 é Normal com média β_1 e variância $VAR(b_1) = \frac{\sigma^2}{SXX}$

A Normalidade de b_0 e b_1 decorre do fato de ambos serem funções lineares dos Y's, que são Normais e independentes.

Logo, se b_0 e b_1 forem independentes de s^2, pode-se construir uma estatística t, a ser usada no cálculo de intervalos de confiança e testes de hipóteses sobre os parâmetros β_0 e β_1.

Sejam:

$S^2_{b_o} = s^2 \left(\frac{\sum_{i=1}^{n} x_i^2}{n.SXX} \right)$ e $S^2_{b_1} = \frac{s^2}{SXX}$ os estimadores das variâncias de b_0 e b_1 respectivamente.

Pode-se provar (e nós o faremos num contexto mais geral) que, sob a hipótese de Normalidade dos erros:

$\frac{b_0 - \beta_0}{\sqrt{S^2_{b_0}}} \approx t_{n-2}$ e $\frac{b_1 - \beta_1}{\sqrt{S^2_{b_1}}} \approx t_{n-2}$

Logo, a distribuição t com n-2 graus de liberdade pode ser usada no teste de hipóteses relacionadas aos parâmetros desconhecidos β_0 e β_1.

Um intervalo de confiança (1 - α)% para o parâmetro β_i (i = 0 ou 1) é:

$$\left[b_i - t_{n-2,1-\frac{\alpha}{2}} \sqrt{S_{b_i}^2}, b_i + t_{n-2,1-\frac{\alpha}{2}} \sqrt{S_{b_i}^2} \right]$$

O teste da hipótese $H_0 : \beta_0 = c$ = constante versus $H_1 : \beta_0 \neq c$ é obtido através da estatística t dada por : $T = \frac{b_0 - c}{\sqrt{S_{b_0}^2}}$ e rejeita-se a hipótese nula com nível α se | T | > $t_{n-2,1-\alpha/2}$. Podemos encontrar, de maneira análoga, um teste para $H_0 : \beta_1 = d$ = constante versus $H_1 : \beta_1 \neq d$. A estatística de teste é $T = \frac{b_1 - d}{\sqrt{S_{b_1}^2}}$ e rejeita-se a hipótese nula com nível α se | T | > $t_{n-2,1-\alpha/2}$.

O caso particular de maior interesse é o teste de $\beta_1 = 0$ versus $\beta_1 \neq 0$, pois corresponde a testar um modelo linear versus um modelo constante. Em outras palavras, sob a hipótese nula $\beta_1 = 0$ supõe-se que X não afeta Y, e este é modelado por uma constante somada a um erro aleatório, enquanto que , sob a hipótese alternativa, existe uma relação linear entre X e Y. A estatística de teste é , neste caso : $T = \frac{b_1}{\sqrt{S_{b_1}^2}}$ e rejeita-se a hipótese nula se | T | > $t_{n-2,1-\alpha/2}$. Note que este teste pode ser escrito em termos de uma variável com distribuição F, pois $T^2 \sim F(1, n-2)$ sob a hipótese nula. Este resultado será empregado a seguir, no desenvolvimento da tabela de análise de variância para a regressão.

A seguir desenvolvemos um teste equivalente para o parâmetro β_1, a partir da idéia de comparar o modelo linear (onde $\beta_1 \neq 0$) e o modelo constante, no qual $\beta_1 = 0$. O teste resultante baseia-se na distribuição F, e introduziremos uma tabela de Análise de Variância (ANOVA).

Suponha que desejamos testar $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0$. A hipótese nula é equivalente à suposição de que o modelo para as observações é:
$Y_i = \beta_0 + \varepsilon_i$ onde os ε_i são iid com média zero e variância constante σ^2 para i = 1,2, ..., n . O método de mínimos quadrados produz o estimador b_0 que minimiza:

$$S(b_0) = \sum_{i=1}^{n} (Y_i - b_0)^2 \Rightarrow b_0 = \overline{Y}$$

A soma de quadrado dos resíduos torna-se, neste modelo simplificado:

RSS = $\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2 = SYY$

Esta soma de quadrado dos resíduos tem n - 1 graus de liberdade. Em seguida considere o modelo linear dado por : $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$. Note que os estimadores de β_0 são diferentes sob os 2 modelos.

Já vimos antes que a soma de quadrado dos resíduos do modelo linear pode ser escrita como :

$$RSS = SYY - \frac{(SXY)^2}{SXX} = SYY - b_1^2(SXX)$$

Pode-se encarar esta expressão de uma maneira que torna explícita a relação entre o modelo linear e o modelo constante, pois a RSS é igual a SYY (soma do quadrado dos resíduos sob o modelo constante) menos um múltiplo de b_1^2 . Isto é, ao considerarmos uma estrutura um pouco mais elaborada (linear ao invés de constante), provocamos uma redução na soma do quadrado dos resíduos. Esta redução na soma do quadrado dos resíduos devido à introdução do parâmetro β_1 (que torna o modelo linear) é chamada de soma dos quadrados da regressão, e denotada, SSReg. Logo, podemos escrever:

$$SS\,\mathrm{Re}\,g = SYY - RSS = SYY - \left(SYY - \frac{(SXY)^2}{SXX}\right) = \frac{(SXY)^2}{SXX} = b_1^2(SXX)$$

O número de graus de liberdade associado a SSReg é a diferença entre os graus de liberdade das somas dos quadrados dos resíduos sob os modelos constante e linear, ou seja (n-1) - (n-2) = 1.

Se SSReg é “grande”, o modelo linear representa uma melhora substancial em relação ao modelo constante, em termos da redução da soma de quadrado dos resíduos. Isto é equivalente a dizer que o parâmetro adicional β_1 existente no modelo linear é significante (isto é, estatisticamente diferente de zero). Para formalizar esta afirmação em termos estatísticos, devemos comparar as somas de quadrados (escalonadas por seus graus de liberdade) sob os 2 modelos (constante e linear), e a estatística F da tabela de análise de variância faz exatamente isto !

Estes resultados são normalmente condensados numa tabela de análise de variância (ANOVA), como indicado a seguir.

Tabela ANOVA para regressão simples

Fonte	graus de liberdade (df)	soma de quadrados (SS)	mean square = SS/df	estatística F
regressão em X	1	SSReg	MSReg = Ssreg/1	MSReg/MSE
resíduo do modelo linear	n-2	RSS	s^2 = RSS/(n-2)	
Total	n-1	SYY		

A estatística F que aparece na tabela ANOVA é:

$$F = \frac{MS\,\mathrm{Re}\,g}{s^2} = \frac{(SXY)^2/(SXX)}{RSS/(n-2)} = \frac{b_1^2(SXX)}{s^2} = \frac{(b_1-0)^2}{\left(\frac{s^2}{SXX}\right)} = \left(\frac{b_1-0}{\sqrt{S_{b_1}^2}}\right)^2 = T^2 \approx F(1,n-2)$$

Note que o uso da distribuição F é justificado pois a estatística considerada é apenas o quadrado da estatística t desenvolvida para o teste de $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0$. Note que esta estatística tem distribuição F apenas quando os erros ε_i são Normais.

A tabela ANOVA simplesmente compara as somas de quadrados decorrentes da especificação de modelos "aninhados", e esta idéia pode ser generalizada para modelos com mais parâmetros.

Definição 1.1.7. (Coeficiente de Determinação)
O coeficiente de determinação da regressão, R^2, é definido como :

$$R^2 = 1 - \frac{RSS}{SYY} = \frac{SS\,\mathrm{Re}\,g}{SYY}$$

R^2 é um número entre 0 e 1, e quanto maior, mais forte é a relação linear entre as variáveis X e Y.

O coeficiente de determinação está relacionado ao coeficiente de correlação amostral entre as variáveis. Especificamente :

$R^2 = \frac{SS\,\mathrm{Re}\,g}{SYY} = \frac{(SXY)^2}{(SXX)(SYY)} = r_{xy}^2$, isto é, R^2 representa o quadrado do coeficiente de correlação amostral entre X e Y.

Algumas observações sobre os resíduos de uma regressão

Os resíduos de uma regressão trazem informações fundamentais sobre a qualidade do ajuste do modelo e possíveis violações das hipóteses do modelo de regressão, tais como : não normalidade dos erros (e consequentemente dos Y's) e variâncias não

constantes (heteroscedasticidade). Aqui faremos alguns comentários muito básicos sobre diagnósticos obtidos a partir dos resíduos.

O gráfico mais importante ao diagnosticar um modelo de regressão talvez seja o dos resíduos versus valores ajustados. A princípio, não deve existir qualquer correlação entre eles, e o gráfico deve ter um aspecto puramente aleatório se a hipótese de variância constante dos erros for válida. Qualquer padrão sistemático neste gráfico nos leva a questionar esta hipótese. Por exemplo, um padrão bastante comum é o de um cone que se abre para o lado direito, isto é, existe maior dispersão nos resíduos para valores ajustados altos, enquanto a dispersão nos resíduos é pequena para valores ajustados baixos. Isto indica heteroscedasticidade na variância dos resíduos.

A existência de alguns poucos resíduos grandes pode indicar a presença de “outliers” , o que é problemático no contexto de parâmetros ajustados por mínimos quadrados, pois a estimação por mínimos quadrados é muito sensível à presença de “outliers” nos dados. “Outliers” podem ter um efeito substancial sobre os parâmetros estimados do modelo, especialmente se os valores de x correspondentes estiverem nos extremos do conjunto de dados. Mas, também é possível que um valor extremo de x “puxe” a reta de regressão na sua direção, provocando um resíduo pequeno.
QQPLOTS dos resíduos devem também ser produzidos para tentar identificar discrepâncias em relação à hipótese de Normalidade, e para verificar a existência de “outliers”.

Por exemplo, no caso de regressões envolvendo dados de contagens, observa-se que, freqüentemente, a variância cresce à medida em que o nível aumenta. Neste caso procura-se aplicar aos dados uma transformação para a estabilização da variância, e a transformação mais comum para este propósito é a raiz quadrada (aplicada a Y e X simultaneamente).

A hipótese de variância constante dos erros é fundamental nos modelos de regressão. Se esta hipótese for violada, os desvios padrões estimados dos parâmetros e os intervalos de confiança obtidos para os parâmetros serão altamente questionáveis.
No caso de regressão de séries temporais, é fundamental olhar para o gráfico da evolução dos resíduos ao longo do tempo, e também para as suas autocorrelações. A existência de padrões sistemáticos nestes gráficos revela que o resíduo resultante do modelo ainda tem alguma estrutura de dependência serial, e esta dependência deve ser explicitamente modelada.

Em resumo, ao analisar a relação entre duas variáveis através de um modelo de regressão não se deve incorrer no erro de só olhar o R^2. Este é, sem dúvida um instrumento importante, mas também é preciso verificar se as hipóteses básicas do modelo estão sendo satisfeitas.

A sintaxe do comando de regressão no MINITAB é mostrada a seguir, e concluímos esta seção com um exemplo.

MTB > help regress
REGRESS C on K predictors C,...,C
REGRESS C on K pred. C,...,C [st. res. in C [fits in C]]

Subcommands: NOCONSTANT XPXINV COOKD VIF WEIGHTS RMATRIX DFITS DWMSE RESIDS HI PURE COEF TRESIDS PREDICT XLOF TOLERANCE

Fits a regression equation to data. To fit an equation without a constant (intercept) use the subcommand NOCONSTANT. To do a weighted fit, use the subcommand WEIGHTS.

If you give an additional column, the standardized residuals will be stored. If you give a second column the fits will be stored in it. See HELP REGRESS RESIDS for a definition of standardized residual.

To control the amount of printed output, use the (main) command BRIEF.

Missing Data in REGRESS.

All observations which contain one or more missing values (either in the dependent or one or more of the independent variables) are not used in any of the regression calculations, with two exceptions: If Yi is missing but all predictors are present (or if case i hasWEIGHT = 0 and all predictors are present), then Yhati is calculated, and hi is calculated as xi*(INV(X'X))*xi', where xi is the row vector of predictors for the i-th observation and X is the design matrix with the i-th observation deleted.

See the Minitab Reference Manual for a discussion on the handling of ill-conditioned data. Ill-conditioned data refers to cases where some predictors are highly correlated with other predictors, or when a predictor variable has a small coefficient of variation.

Exemplo 1.1.8.
Neste exemplo examinamos a relação entre consumo de gasolina e outras variáveis relevantes para 48 estados americanos.

As variáveis sob estudo, e as colunas onde estão localizadas na planilha MINITAB, estão descritas a seguir.

variável	coluna	número de observações	descrição
consumo	C1	48	consumo em milhões de galões em 1972
pop	C2	48	população em milhares (1971)
renda	C3	48	renda per capita em dólares (1972)

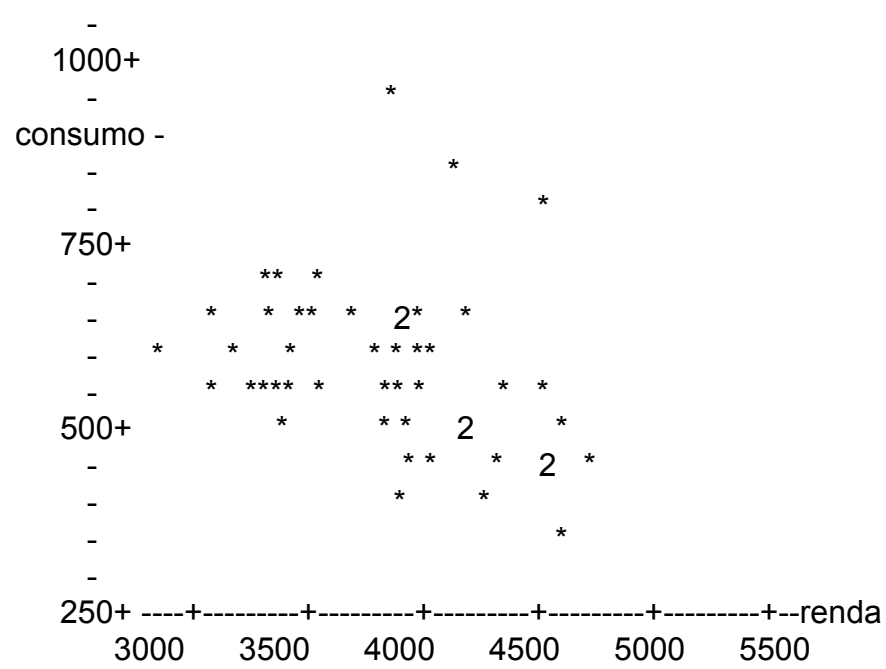
estrada	C4	48	número de milhas de estradas federais em 1971
taxa	C5	48	imposto sobre combustíveis em centavos por galão de gasolina (1972)

Algumas estatísticas descritivas das variáveis estão abaixo.

variável	Média	mediana	mínimo	máximo	desvio padrão
consumo	576.8	568.5	344	968	111.9
pop	0.57033	0.56450	0.451	0.724	0.05547
renda	4241.8	4298.0	3063	5342	573.6
estrada	5565	4736	431	17782	3492
taxa	7.668	7.500	5	10	0.951

O próximo gráfico exibe a relação entre consumo e renda. É surpreendente notar que o consumo é alto em estados com renda per capita mais baixa. Isto se deve principalmente à forma como as variáveis foram padronizadas. “Renda” é expressa “per capita”, ou seja, numa base individual. Por outro lado, o consumo representa o estado inteiro, e será relativamente grande em estados populosos e pequeno em estados pouco habitados. Logo, a relação entre estas variáveis pode estar sendo deturpada por problemas na padronização. Faz sentido transformar a variável consumo num consumo individual, dividindo-a pela população do estado, antes de tentar explicar a razão desta relação “estranha” entre as variáveis.

MTB > plot c1 c3



O próximo gráfico exibe a relação entre consumo e taxa. Aqui o comportamento observado é intuitivamente razoável: o consumo é alto quando o imposto cobrado é baixo e vice versa.

```
 MTB > plot c1 c5

          -
       1000+
          -                    *
    consumo -
          -                    *
          -            *
          -
        750+
          -                    3
          -     *          *   3    2        2
          -                    4        3
          -                    4        2        5
        500+                      3        *        2
          -                        *    2    *    *        *
          -                        *    *
          -                             *
          -
          -
        250+
            ----+---------+---------+---------+---------+---------+--taxa
              5.0       6.0       7.0       8.0       9.0      10.0
```

No quadro abaixo apresentamos o resultado da regressão no MINITAB. Note que as colunas c10 e c11 contém, respectivamente, os resíduos e valores ajustados da regressão.

```
 MTB > regr c1 1 c5 res c10 fv c11

The regression equation is
 consumo = 984 - 53.1 taxa

 Predictor       Coef      Stdev    t-ratio       p
 Constant       984.0      119.6       8.23   0.000
 taxa         -53.11      15.48      -3.43   0.001
s = 100.9       R-sq = 20.4%     R-sq(adj) = 18.6%
```

O valor de R^2 é baixo, mas mesmo assim os coeficientes estimados são significantes. Isso quer dizer que a taxa “explica” em parte, o consumo de gasolina, mas outras variáveis devem também ser levadas em consideração.

A tabela ANOVA para esta regressão está a seguir.

Analysis of Variance

SOURCE	DF	SS	MS	F	p
Regression	1	119823	119823	11.76	0.001
Error	46	468543	10186		
Total	47	588366			

O MINITAB também produz diagnósticos indicando observações influentes e “outliers” do modelo.

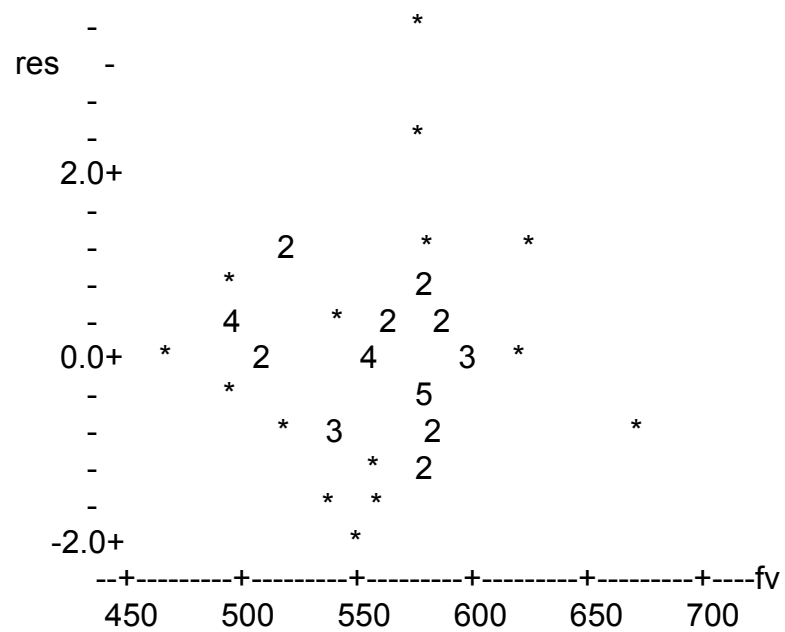
Unusual Observations

Obs.	taxa	consumo	Fit	Stdev.Fit	Residual	St.Resid
6	10.0	457.0	452.9	38.9	4.1	0.04 X
7	8.0	344.0	559.2	15.4	-215.2	-2.16R
19	7.0	865.0	612.3	17.9	252.7	2.54R
37	5.0	640.0	718.5	43.8	-78.5	-0.86 X
40	7.0	968.0	612.3	17.9	355.7	3.58R

R denotes an obs. with a large st. resid.
X denotes an obs. whose X value gives it large influence.

A seguir apresentamos o gráfico dos resíduos versus os valores ajustados. Nota-se claramente a existência de um padrão sistemático neste gráfico, pois existem resíduos altos à medida que crescem os valores ajustados pela regressão, indicando a heteroscedasticidade dos erros. Este problema possivelmente será contornado ao incluirmos outras variáveis explicativas no modelo.

```
MTB > name c10 'res' c11 'fv'
MTB > plot c10 c11
```



1.2. Modelos de Regressão Linear Múltipla

Nesta seção estendemos os resultados da seção 1.1. para o caso geral, em que existe mais de uma variável explicativa. Diversos resultados gerais sobre médias e matrizes de covariância de vetores aleatórios são apresentados, mas as demonstrações serão omitidas na maioria das vezes.

Suponha que o modelo a ser ajustado aos dados tem a forma:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_{p-1} x_{p-1} + \varepsilon \qquad (1.2.1.)$$

Onde x_1, x_2, ..., x_{p-1} são covariáveis (variáveis explicativas) supostas fixas e ε é um erro aleatório com média zero e variância constante σ^2 .

Suponha também que foram coletadas n observações, onde cada observação consiste na p-upla (y, x_1,x_2, ..., x_{p-1}). O objetivo é ajustar um hiperplano (em p dimensões) a estas n observações, utilizando o critério de mínimos quadrados, a exemplo do que fizemos na seção 1.1.

O desenvolvimento aqui será apresentado em forma matricial. Para tanto é preciso definir algumas matrizes de interesse.

A matriz de "design" da regressão é a matriz n por p onde cada linha corresponde aos elementos que multiplicam os β's na equação (1.2.1). Especificamente, a i-ésima linha da matriz de design é dada por : (1, x_{i1} , x_{i2}, ..., $x_{i,p-1}$).

A matriz de design é então a matriz com n linhas e p colunas dada por :

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & ... & x_{1,p-1} \\ 1 & x_{21} & x_{22} & ... & x_{2,p-1} \\ & & ..... & & \\ 1 & x_{n1} & x_{n2} & ... & x_{n,p-1} \end{bmatrix}$$

Seja $\beta = (\beta_0, \beta_1, ..., \beta_{p-1})^t$ o vetor p x 1 de parâmetros desconhecidos a serem estimados. A equação (1.2.1.) pode ser escrita, em forma matricial, como:

$$Y = X.\beta + \varepsilon \qquad (1.2.2.)$$

Seja $b = (b_0, b_1, ..., b_{p-1})^t$ o vetor de estimadores dos coeficientes obtido por mínimos quadrados. Então b é obtido de forma a minimizar a soma de quadrado dos resíduos dada por:

$$RSS = \sum_{i=1}^{n} \left(Y_i - b_0 - b_1 x_{i1} - ... - b_{p-1} x_{i,p-1} \right)^2 = \sum_{i=1}^{n} \left(Y_i - \widetilde{Y}_i \right)^2 = \| Y - Xb \|^2$$

Para encontrar os estimadores (b_0, b_1, ..., b_{p-1}) diferenciamos RSS em relação a cada um dos b's e igualamos as p derivadas a zero. A solução deste sistema de equações produz os estimadores por mínimos quadrados do vetor b = $(b_0, b_1, ..., b_{p-1})^t$. O estimador por mínimos quadrados do vetor β é dado por :

$$b = (X^tX)^{-1}X^tY \qquad (1.2.3)$$

A equação (1.2.3) surge como solução das **equações normais**, dadas por:

$$(X^tX)b = X^t.Y$$

Claramente, o requisito necessário para que (1.2.3) seja a solução por mínimos quadrados das equações normais é que a matriz X^tX seja inversível. Note também (prove !!!) que X^tX é uma matriz simétrica, ou seja : $(X^tX)^t = (X^tX)$.

O lema a seguir nos dá as condições para a existência da solução por mínimos quadrados.

Lema 1.2.1.
A solução (1.2.3) existe se e somente se, X^tX é não singular, e isto ocorre se X é uma matriz de posto p.

Em termos matriciais, a equação para os valores ajustados de Y é análoga à equação (1.2.2.), a saber:

$$Y = X.b$$

A soma do quadrado dos resíduos pode ser escrita em forma matricial como :

$$RSS = \sum_{i=1}^{n}\left(Y_i - \tilde{Y}_i\right)^2 = \left(Y - \tilde{Y}\right)^t\left(Y - \tilde{Y}\right) = (Y - Xb)^t(Y - Xb) = \left(Y^t - b^tX^t\right)(Y - Xb) =$$
$$= Y^tY - b^tX^tY - Y^tXb + b^tX^tXb$$

Mas, note que todos os termos no lado direito da equação acima são escalares (por exemplo, b^tX^tY é (1 x p)(p x n)(n x 1)) e então podemos escrever : $Y^t.X.b = (Y^t.X.b)^t = b^t.X^t.Y$ e então:

$$RSS = Y^tY - b^tX^tY - Y^tXb + b^tX^tXb = Y^tY - b^tX^tY - b^tX^tXb + b^tX^tXb =$$
$$= Y^tY - b^tX^tY$$

Substituindo b^t nesta última equação leva a:

$$RSS = Y^tY - b^tX^tY = Y^tY - \left[\left(X^tX\right)^{-1}X^tY\right]^tX^tY = Y^tY - \left(X^tY\right)^t\left[\left(X^tX\right)^{-1}\right]^tX^tY =$$
$$= Y^tY - Y^tX\left(X^tX\right)^{-1}X^tY = Y^t\left[I - X\left(X^tX\right)^{-1}X^t\right]Y$$

onde I é a matriz identidade de dimensão n.

Seja P = X.($X^tX)^{-1}.X^t$. Note que P é uma matriz quadrada de dimensão n. Então a soma dos quadrados dos resíduos pode ser escrita como :

$$RSS = Y^t\left[I - X\left(X^t X\right)^{-1} X^t \right]Y = Y^t\left[I - P\right]Y \qquad (1.2.4)$$

A matriz P é freqüentemente chamada de matriz "chapéu". Por que? Porque transforma Y em "Y chapéu", ou seja, transforma o vetor de valores observados no vetor de valores ajustados. De fato:

$$PY = X\left(X^t X\right)^{-1} X^t Y = Xb = \widetilde{Y}$$

Daí segue que o vetor de resíduos é apenas:

$$e = Y - \widetilde{Y} = Y - PY = \left(I - P\right)Y$$

e então a soma dos quadrados dos resíduos toma a forma indicada em (1.2.4).

Exemplo 1.2.2. (regressão linear simples)

Considere o modelo de regressão linear simples estudado na seção 1.1., isto é:

$Y_i = \beta_0 + \beta_1 . x_i + \varepsilon_i$ onde os ε_i são supostos independentes com média nula e variância constante σ^2 . A seguir especificamos este modelo na forma matricial dada por (1.2.2).

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \dots \\ Y_n \end{bmatrix} , \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} , X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \dots & \dots \\ 1 & x_n \end{bmatrix} , Y - X\beta = \begin{bmatrix} Y_1 - \beta_0 - \beta_1 x_1 \\ Y_2 - \beta_0 - \beta_1 x_2 \\ \dots \\ Y_n - \beta_0 - \beta_1 x_n \end{bmatrix}$$

$$X^t X = \begin{bmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_n \end{bmatrix} . \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \dots & \dots \\ 1 & x_n \end{bmatrix} = \begin{bmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{bmatrix}$$

$$\left(X^t X\right)^{-1} = \frac{1}{nSXX} \begin{bmatrix} \sum_{i=1}^{n} x_i^2 & -\sum_{i=1}^{n} x_i \\ -\sum_{i=1}^{n} x_i & n \end{bmatrix} , X^t Y = \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i y_i \end{bmatrix}$$

E então:

$$b = \left(X^t X\right)^{-1} X^t Y = \frac{1}{nSXX}\begin{bmatrix} \sum_{i=1}^{n} x_i^2 & -\sum_{i=1}^{n} x_i \\ -\sum_{i=1}^{n} x_i & n \end{bmatrix} . \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i y_i \end{bmatrix} =$$

$$= \frac{1}{nSXX}\begin{bmatrix} n\overline{Y}\sum_{i=1}^{n} x_i^2 - n\overline{X}\sum_{i=1}^{n} x_i y_i \\ n\sum_{i=1}^{n} x_i y_i - \left(n\overline{X}\right)\left(n\overline{Y}\right) \end{bmatrix} = \frac{1}{nSXX}\begin{bmatrix} n\overline{Y}\left(SXX + n\overline{X}^2\right) - n\overline{X}\left(SXY + n\overline{XY}\right) \\ n\left(SXY\right) \end{bmatrix} =$$

$$= \frac{1}{SXX}\begin{bmatrix} \overline{Y}\left(SXX + n\overline{X}^2\right) - \overline{X}\left(SXY + n\overline{XY}\right) \\ SXY \end{bmatrix} = \begin{bmatrix} \overline{Y} + \frac{n\overline{X}^2}{SXX}\overline{Y} - \frac{\overline{X}}{SXX}SXY - \frac{n\overline{X}^2}{SXX}\overline{Y} \\ \frac{SXY}{SXX} \end{bmatrix} =$$

$$= \begin{bmatrix} \overline{Y} - \frac{\overline{X}}{SXX}SXY \\ \frac{SXY}{SXX} \end{bmatrix} = \begin{bmatrix} \overline{Y} - \frac{SXY}{SXX}\overline{X} \\ \frac{SXY}{SXX} \end{bmatrix}$$

Como já demonstrado.

A seguir examinamos algumas propriedades estatísticas de vetores aleatórios e aplicamos estas propriedades para obter a média e a matriz de covariância dos estimadores por mínimos quadrados.

Resultado 1.2.3.
Seja Y um vetor coluna aleatório de dimensão n, com vetor de médias μ e matriz de covariância Σ (não necessariamente diagonal).

Seja Z um vetor de dimensão coluna de dimensão k definido por :
Z = c + A.Y onde c é um vetor (k x 1) de constantes e A é uma matriz fixa de dimensão k x n. Então o vetor de médias e a matriz de covariância de Z estão relacionadas aos respectivos momentos de Y da seguinte forma :

$E(Z) = a^t.E(Y) = a^t.\mu$ e
$COV(Z) = A.COV(Y).A^t = A.\Sigma.A^t$

Exemplo 1.2.4.
Considere um caso particular do resultado 1.2.3. e suponha que Z é um escalar dado por $Z = a^t.Y = \sum_{i=1}^{n} a_i Y_i$. Suponha que Y tem vetor de médias μ e matriz de covariância Σ =

COV(Y) = σ^2.I (diagonal). Então, o vetor de médias de Z é apenas a^t .μ = $\sum_{i=1}^{n} a_i \mu_i$ e a matriz de covariância de Z reduz-se a um escalar dado por : Var(Z) = a^t.COV(Y).a = a^t.(σ^2 I).a =

= σ^2 (a^t.a)= $\sigma^2 \sum_{i=1}^{n} a_i^2$

Definição 1.2.5. (Forma quadrática)
Seja A uma matriz simétrica n x n e seja x um vetor coluna de dimensão n. A expressão:

$x^t Ax = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij} x_i x_j$ é chamada de forma quadrática.

Note que uma forma quadrática é um escalar, mas suas propriedades estatísticas são geralmente obtidas a partir daquelas do vetor aleatório x. Formas quadráticas surgem naturalmente em modelos de regressão. Um exemplo é SXX = $\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$

O próximo teorema nos dá a média de uma forma quadrática.

Teorema 1.2.6.
Seja Y um vetor coluna de dimensão n com vetor de médias μ e matriz de covariância Σ. Seja A uma matriz constante de dimensão n x n, e considere a forma quadrática U = Y^t.A.Y . Então, a média de U é dada por :
E(U) = E(Y^t.A.Y) = tr(A.Σ) + μ^t.A.μ , onde tr(.) indica o traço da matriz.

Exemplo 1.2.7.
Sejam Y_i (i = 1, 2 ,..., n) variáveis aleatórias descorrelatadas com média μ e variância σ^2 . Então o vetor aleatório Y tem vetor de médias constante e matriz de covariância Σ = σ^2.I, um múltiplo da matriz identidade.

Considere a forma quadrática SYY = $\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2$. Desejamos inicialmente escrever SYY na forma Y^t.A.Y (o que equivale a encontrar a matriz A) e, em seguida, utilizar o teorema 1.2.6. para encontrar a média de SYY.
Note que a média amostral pode ser escrita como:

$\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i = \frac{1}{n}\left(1^t Y\right)$ onde 1 indica um vetor coluna de dimensão n, onde todos os elementos são fixos e iguais a um. Note também que 1^t.1 = n. Logo, o vetor n x 1 cujos elementos são todos iguais à média amostral pode ser escrito como : $\frac{1}{n}\left(1.1^t Y\right)$. Assim,

$Y_i - \overline{Y}$ é apenas o i-ésimo elemento do vetor $Y - \frac{1}{n}\left(1.1^t Y\right) = \left(I - \frac{1}{n}\left(1.1^t\right)\right)Y$ onde I novamente indica a matriz identidade.

Seja A a matriz : $\left(I - \frac{1}{n}\left(1.1^t\right)\right)$. Note que A é uma matriz quadrada de dimensão n e também A é simétrica. Pode-se verificar também que $A^2 = A.A = A^t.A = A.A^t = A$ (isto é, a matriz A é *idempotente*).

Então a forma quadrática SYY pode ser escrita como:

$$SYY = \left(\left(I - \frac{1}{n}\left(1.1^t\right)\right)Y\right)^t \left(\left(I - \frac{1}{n}\left(1.1^t\right)\right)Y\right) = \left(AY\right)^t \left(AY\right) = Y^t A^t AY = Y^t AY$$

Usando o teorema 1.2.6. leva a:

$$E\left(SYY\right) = E\left(Y^t AY\right) = \sigma^2 tr\left(A\right) + \mu^t A\mu$$

Como os elementos do vetor de médias são todos iguais pode-se escrever: $\mu = \mu.1$ e segue que $\mu^t A\mu = 0$ (verifique !!!). Também : tr(A) = n-1 e assim concluímos que $E(SYY) = (n-1).\sigma^2$.

O próximo resultado fornece a matriz de covariâncias entre dois vetores que são funções lineares do mesmo vetor aleatório.

Teorema 1.2.8.
Seja X um vetor aleatório de dimensão n x 1 com matriz de covariância Σ, e sejam A e B matrizes constantes com dimensões, respectivamente,
k x n e l x n. Defina os vetores aleatórios Y e Z como:
Y = A.X e Z = B.X .

Então:

A matriz de covariâncias cruzadas entre os elementos de Y e Z é uma matriz de dimensão k por l dada por : $\Sigma_{YZ} = A.\Sigma.B^t$.

A matriz de covariância de Y é uma matriz quadrada de dimensão k dada por: $\Sigma_{YY} = A.\Sigma.A^t$ (analogamente para a matriz de covariância de Z).

Nos casos particulares $Y = a^t.X$ e $Z = b^t.X$ onde a e b são, respectivamente, vetores coluna de dimensão p, e daí Y e Z são escalares, os resultados anteriores reduzem-se a:

- A matriz de covariâncias cruzadas entre Y e Z reduz-se a um escalar, a covariância entre Y e Z, dada por : $Cov(Y,Z) = a^t.\Sigma.b$.

- A variância de Y é apenas $a^t.\Sigma.a$, enquanto a variância de Z é $b^t.\Sigma.b$.

A seguir aplicamos o teorema 1.2.8. a um problema extremamente importante em regressão, a saber: determinar a covariância entre a média amostral e $X_i - \overline{X}$.

Exemplo 1.2.9.
Seja X um vetor aleatório de dimensão n com vetor de médias constante e matriz de covariância $\Sigma_{XX} = \sigma^2.I$. Sejam Y = $\overline{X}$ e Z o vetor cujo i-ésimo elemento é $X_i - \overline{X}$. A matriz de covariâncias cruzadas entre Z e Y é uma matriz n por 1. Note que ambos Y e Z são funções lineares do vetor aleatório X, e podem ser escritos como:

$$Z = \left(I - \frac{1}{n} 1.1^t \right) X \quad \text{e} \quad Y = \left(\frac{1}{n} \right) .1^t \, X$$

onde 1 representa um vetor coluna de dimensão n no qual todos os elementos são iguais a um.

Pelo teorema anterior, a matriz de covariâncias cruzadas entre Z e Y é:

$$\Sigma_{ZY} = \left(I - \frac{1}{n} 1.1^t \right) \left(\sigma^2 I \right) \left(\frac{1}{n} \right) .1 = \frac{\sigma^2}{n} \left(I - \frac{1}{n} 1.1^t \right) .1 = \frac{\sigma^2}{n} \left[\begin{pmatrix} 1 & 0 & 0 & \ldots & 0 \\ 0 & 1 & 0 & \ldots & 0 \\ & & \ldots\ldots\ldots & & \\ 0 & 0 & 1 & \ldots. & 1 \end{pmatrix} - \frac{1}{n} \begin{pmatrix} 1 & 1 & 1 & \ldots & 1 \\ 1 & 1 & 1 & \ldots & 1 \\ & & \ldots\ldots\ldots & & \\ 1 & 1 & 1 & \ldots. & 1 \end{pmatrix} \right] . \begin{bmatrix} 1 \\ 1 \\ \ldots \\ 1 \end{bmatrix} =$$

$$= \frac{\sigma^2}{n} \left[\begin{pmatrix} (n-1)/n & -1/n & -1/n & \ldots & -1/n \\ -1/n & (n-1)/n & -1/n & \ldots & -1/n \\ & & \ldots\ldots\ldots\ldots & & \\ -1/n & -1/n & -1/n & & (n-1)/n \end{pmatrix} \right] . \begin{bmatrix} 1 \\ 1 \\ \ldots \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \ldots \\ 0 \end{bmatrix}$$

Logo, nas condições enunciadas acima, $\overline{X}$ e cada um dos $X_i - \overline{X}$ são descorrelatados.

Considere agora o modelo linear descrito por (1.2.2.) e suponha que o estimador por mínimos quadrados do vetor β seja b definido em (1.2.3.). A seguir empregamos os teoremas enunciados anteriormente para derivar as propriedades estatísticas do estimador de mínimos quadrados.

Suponha que o vetor de erros, ε, é um vetor aleatório de dimensão n x 1 com vetor de médias zero e matriz de covariância $\sigma^2.I$.

Teorema 1.2.10.
Sob a hipótese de erros com média nula, o estimador por mínimos quadrados é não tendencioso para o vetor de parâmetros β.

Demonstração
O estimador por mínimos quadrados de β é:
$b = (X^tX)^{-1}X^tY$

Aplicando-se o valor esperado encontramos:

$E(b) = E\{(X^tX)^{-1}X^tY\} = (X^tX)^{-1}X^t.E(Y) = (X^tX)^{-1}X^t.E(X.\beta + \varepsilon) =$

$= (X^tX)^{-1}X^t.(X.\beta + E(\varepsilon)) = (X^tX)^{-1}X^t.(X.\beta) = (X^tX)^{-1}X^t.X.\beta = \beta$

Note que a única hipótese utilizada nesta demonstração é a de que os erros têm média nula ! Em particular, mesmo que os erros sejam correlatados ou tenham variância não constante, os estimadores por mínimos quadrados são não tendenciosos.

A matriz de covariância dos elementos do vetor b também pode ser calculada, mas a demonstração do próximo teorema depende da forma especificada para a matriz de covariância dos erros.

Teorema 1.2.11.
Sob a hipótese de que os erros têm média zero, variâncias constantes e são descorrelatados, a matriz de covariância dos estimadores por mínimos quadrados de β é :

$$\Sigma_{bb} = \sigma^2.\left(X^tX\right)^{-1}$$

Demonstração
Do teorema 1.2.8. segue que:

$$\Sigma_{bb} = COV(b) = COV\left(\left(X^tX\right)^{-1}X^tY\right) = \left(\left(X^tX\right)^{-1}X^t\right)COV(Y)\left(\left(X^tX\right)^{-1}X^t\right)^t =$$

$$= \left(\left(X^tX\right)^{-1}X^t\right)\sigma^2.I.\left(\left(X^tX\right)^{-1}X^t\right)^t = \sigma^2\left(\left(X^tX\right)^{-1}X^t\right)\left(X.\left(X^tX\right)^{-1}\right) = \sigma^2.\left(X^tX\right)^{-1}$$

Exemplo 1.2.12.
Considere o modelo de regressão linear simples de seção 1.1. No exemplo 1.2.2. deduzimos que :

$$\left(X^tX\right)^{-1} = \frac{1}{nSXX}\begin{bmatrix} \sum_{i=1}^{n} x_i^2 & -\sum_{i=1}^{n} x_i \\ -\sum_{i=1}^{n} x_i & n \end{bmatrix}$$

Logo, a matriz de covariância de $b = (b_0 , b_1)^t$ é dada por:

$$COV(b) = \sigma^2\left(X^t X\right)^{-1} = \frac{\sigma^2}{nSXX}\begin{bmatrix} \sum_{i=1}^{n} x_i^2 & -\sum_{i=1}^{n} x_i \\ -\sum_{i=1}^{n} x_i & n \end{bmatrix} = \frac{\sigma^2}{nSXX}\begin{bmatrix} \sum_{i=1}^{n} x_i^2 & -n.\overline{X} \\ -n.\overline{X} & n \end{bmatrix} =$$

$$= \begin{bmatrix} \frac{\sigma^2}{nSXX}.\sum_{i=1}^{n} x_i^2 & -.\frac{\overline{X}\sigma^2}{SXX} \\ -.\frac{\overline{X}\sigma^2}{SXX} & \frac{\sigma^2}{SXX} \end{bmatrix}$$

O que concorda com o resultado do teorema 1.1.2.

Estimação de σ^2

A exemplo do que fizemos no caso de regressão simples, aqui também precisamos encontrar um bom estimador para σ^2 para podermos construir intervalos de confiança para os elementos do vetor β. A idéia é, novamente, empregar a soma de quadrado dos resíduos na estimação de σ^2, já que os resíduos são, por sua vez, estimadores dos erros (não observáveis) do modelo, e portanto o quadrado dos resíduos "deve servir" como estimador do valor esperado do quadrado do erro, que é a variância do erro.

A seguir exibimos algumas propriedades da matriz "chapéu" P. Lembre-se que o efeito desta matriz, quando aplicada sobre Y, é transformá-lo no vetor de valores ajustados, e assim o vetor de resíduos pode ser escrito como :

$$e = Y - \widetilde{Y} = Y - Xb = Y - X\left(X^t X\right)^{-1} X^t Y = Y - PY = (I - P)Y \quad \text{onde} \quad P = X\left(X^t X\right)^{-1} X^t$$

Lema 1.2.13.

Seja P a matriz "chapéu' definida anteriormente. Então:

$P = P^t = P^2$ e

$(I - P) = (I - P)^t = (I - P)^2$

isto é, P e (I-P) são simétricas e idempotentes.

A soma do quadrado dos resíduos pode ser escrita como :

$$RSS = Y^t\left[I - X\left(X^t X\right)^{-1} X^t\right]Y = Y^t[I - P]Y$$

Isto é uma forma quadrática, e o teorema 1.2.6. fornece diretamente o valor esperado desta forma quadrática. Daí :

$$E(RSS) = tr\left\{\left[I - X\left(X^t X\right)^{-1} X^t\right].COV(Y)\right\} + (X\beta)^t\left[I - X\left(X^t X\right)^{-1} X^t\right](X\beta) =$$

$$= tr\left\{[I - P].\sigma^2 I\right\} + (X\beta)^t\left[I - X\left(X^t X\right)^{-1} X^t\right](X\beta) =$$

$$= \sigma^2.tr\{I - P\} + \beta^t X^t\left[I - X\left(X^t X\right)^{-1} X^t\right](X\beta)$$

Mas,
tr(I - P) = tr(I) - tr(P) = n - tr(P)

Também , em geral para matrizes A e B , tr(AB) = tr(BA), e aplicando este resultado no cálculo do traço da matriz "chapéu" leva a :

$$tr(P) = tr\left\{X\left(X^t X\right)^{-1} X^t\right\} = tr\left\{\left(X^t X\right).\left(X^t X\right)^{-1}\right\} = tr\left(I_{pxp}\right) = p$$

Finalmente :

$$E(RSS) = \sigma^2.tr\{I - P\} + \beta^t X^t\left[I - X\left(X^t X\right)^{-1} X^t\right](X\beta) =$$

$$= \sigma^2.(n - p) + \beta^t X^t X\beta - \beta^t X^t X\left(X^t X\right)^{-1} X^t X\beta =$$

$$= \sigma^2.(n - p) + \beta^t X^t X\beta - \beta^t X^t X\beta =$$

$$= \sigma^2.(n - p)$$

Esta expressão produz um estimador não tendencioso "natural" para σ^2 , a saber :

$$s^2 = \frac{RSS}{n - p}$$

Sob a hipótese de erros descorrelatados, com variância constante σ^2 e com média nula, o estimador s^2 é não tendencioso para σ^2 . Note que este estimador é o encontrado na seção 1.1. fazendo p = 2 .

Alguns comentários adicionais sobre os resíduos

Já vimos que o vetor de resíduos pode ser escrito como:

$$e = Y - \widetilde{Y} = (I - P)Y$$

Isso nos permite calcular facilmente a matriz de covariância dos resíduos, que é:

$$COV(e) = COV\left((I - P)Y\right) = (I - P)^t COV(Y)(I - P) = (I - P)^t \sigma^2 I(I - P) = \sigma^2(I - P)$$

Note que (I - P) não é uma matriz diagonal, e portanto os resíduos de diferentes observações são correlatados ! Também, as variâncias dos resíduos são diferentes. Logo, para tornar os resíduos de diferentes observações comparáveis é necessário padronizá-los, de tal forma que os resíduos resultantes tenham média zero e variância um. Esta padronização, entretanto, depende de σ^2 , e a solução óbvia é substituí-lo por seu estimador não tendencioso s^2 no processo de padronização dos resíduos.

O i-ésimo resíduo padronizado é definido como:

$\frac{Y_i - \widetilde{Y}_i}{s\sqrt{1-p_{ii}}}$ onde p_{ii} é o i-ésimo elemento da diagonal de P.

Uma propriedade importante dos resíduos é dada no próximo teorema.

Teorema 1.2.14.
Se os erros são descorrelatados e com variância constante σ^2 então os resíduos e valores ajustados são descorrelatados.

Demonstração
Os resíduos são $e = Y - \widetilde{Y} = (I - P)Y$ e os valores ajustados são dados por $\widetilde{Y} = PY$. Do teorema 1.2.8. segue que a matriz de covariâncias cruzadas entre os resíduos e valores ajustados é apenas : $(I - P)\sigma^2 I.P^t = \sigma^2(I - P)P^t = \sigma^2\left(P^t - PP^t\right) = \sigma^2(P - P) = 0$ pois P é simétrica e idempotente.

Por que este teorema é importante ???
Se as hipóteses do teorema são válidas, os resíduos e valores ajustados são descorrelatados, e portanto o gráfico destas duas quantidades deve ter um padrão puramente aleatório. A existência de quaisquer padrões neste gráfico indica que existe correlação, e portanto, que alguma hipótese básica do modelo de regressão (como homoscedasticidade dos erros !) não é satisfeita.

Inferência a respeito dos β's

Note que os estimadores por mínimos quadrados dos parâmetros desconhecidos (β_0, β_1, ..., β_{p-1}) são não tendenciosos e a matriz de covariância do vetor de estimadores é dada por : $\Sigma_{bb} = \sigma^2.\left(X^t X\right)^{-1}$.

Suponha agora que os erros no modelo (1.2.1) são Normais e independentes com média zero e variância constante σ^2 . Note que isso implica na Normalidade dos Y_i 's e, consequentemente, os estimadores b_j também são Normais, pois são funções lineares dos Y's.

Em particular segue que cada estimador b_j (j = 0, 1, 2, ..., p-1) é Normal com média β_j e variância $\sigma^2(c_{jj})$ onde c_{jj} é o j-ésimo elemento da diagonal da matriz $(X^t.X)^{-1}$. Assim, o erro padrão de b_j pode ser estimado por $S_{b_j} = s\sqrt{c_{jj}}$ onde s é a raiz do estimador da variância σ^2 .

O próximo teorema indica qual a distribuição de probabilidade do estimador b_j.

Teorema 1.2.15.
Sob a hipótese de erros Normais e com variância constante pode-se mostrar que:

$$\frac{b_j - \beta_j}{S_{b_j}} \approx t_{n-p}$$

Este teorema nos permite encontrar um intervalo de confiança 100(1-α)% para o parâmetro desconhecido β_j (onde j = 0,1, ..., p-1). Este intervalo é dado por: $\left(b_j - t_{n-p,1-\alpha/2}.S_{b_j}, b_j + t_{n-p,1-\alpha/2}.S_{b_j}\right)$ onde $t_{n-p,1-\alpha/2}$ indica o percentil (1 - α/2) da distribuição t com n-p graus de liberdade.

Podemos estender a idéia de intervalos de confiança para testar hipóteses a respeito dos parâmetros desconhecidos.

Suponha que a hipótese nula é:
$H_0 : \beta_j = 0$
Sob esta hipótese nula, a estatística $\frac{b_j}{S_{b_j}}$ tem distribuição t com n-p graus de liberdade, e portanto rejeitamos H_0 se esta estatística é "grande" em módulo. Especificamente, para um teste de nível α, a hipótese nula $\beta_j = 0$ é rejeitada se : $|T| = \left|\frac{b_j}{S_{b_j}}\right| \geq t_{n-p,1-\alpha/2}$, ou seja, se o intervalo de confiança correspondente não inclui zero.

Tabela de análise de variância

Analogamente ao que fizemos no caso de regressão simples, aqui procuramos comparar o modelo geral (com p-1 variáveis explicativas) com o modelo constante $Y = \beta_0 + \varepsilon$. Para o modelo constante, o estimador de β_0 é a média amostral dos Y's e a soma dos quadrados dos resíduos se reduz a SYY. Claramente, RSS, a soma dos quadrados dos resíduos do modelo com variáveis explicativas, é menor que SYY, e a diferença entre as duas é a soma de quadrados devido à regressão, isto é:

SSReg = SYY - RSS

A tabela a seguir indica os graus de liberdade associados a cada uma destas somas de quadrados, e permite a construção de um teste F para verificar a validade do modelo de regressão. Note que o modelo suposto para os dados envolve p -1 variáveis explicativas além de um termo constante, isto é, tem a forma : $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_{p-1} x_{p-1} + \varepsilon$

Tabela ANOVA
(Regressão múltipla baseada em p-1 variáveis X_1 , X_2 , ... , X_{p-1})

Fonte	graus de liberdade (df)	soma de quadrados (SS)	mean square = SS/df	estatística F
regressão em X	p-1	SSReg	MSReg = Ssreg/(p-1)	MSReg/ s^2
resíduo do modelo linear	n-p	RSS	s^2 = RSS/(n-p)	
Total	n-1	SYY		

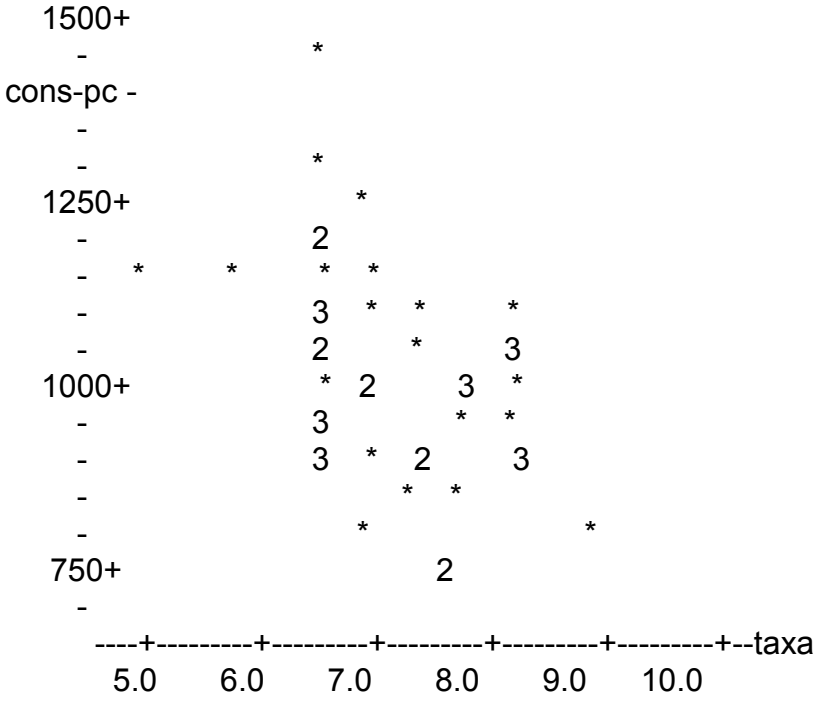
Podemos julgar se as variáveis explicativas são importantes na previsão de Y através da verificação do quanto SSReg é "grande" em relação a SYY. A idéia é comparar MSReg/s^2 com pontos percentuais da distribuição F com p-1 graus de liberdade no numerador e n - p graus no denominador. Se este valor excede um percentil apropriado da distribuição F dizemos que a regressão é significante. Do contrário, as variáveis explicativas escolhidas não têm (em conjunto) a capacidade de "explicar" a variável dependente Y.

O coeficiente de determinação R^2 é definido de maneira análoga ao caso de regressão simples, isto é : $R^2 = 1 - \frac{RSS}{SYY} = \frac{SS\,\mathrm{Re}\,g}{SYY}$, e fornece a porção da variabilidade dos Y's explicada pela regressão nos X's. Também pode-se mostrar que R^2 representa o quadrado da maior correlação entre Y e qualquer função linear de X_1 , X_2 , ... , X_{p-1}.

Exemplo 1.2.16.
(Vide exemplo 1.1.8)
Aqui consideramos novamente a relação entre consumo de gasolina e algumas variáveis explicativas para 48 estados americanos.
Y = consumo de galões de gasolina *per capita* (1 galão = 3.8 litros)
X_1 = taxa, em centavos de dólar, por galão
X_2 = proporção da população com carteira de motorista (em 1971)
X_3= renda per capita em US$
X_4= milhas de estrada principais administradas pelo governo federal (em 1971)
O objetivo desse trabalho é explicar a variação de consumo de gasolina através de variáveis relevantes.

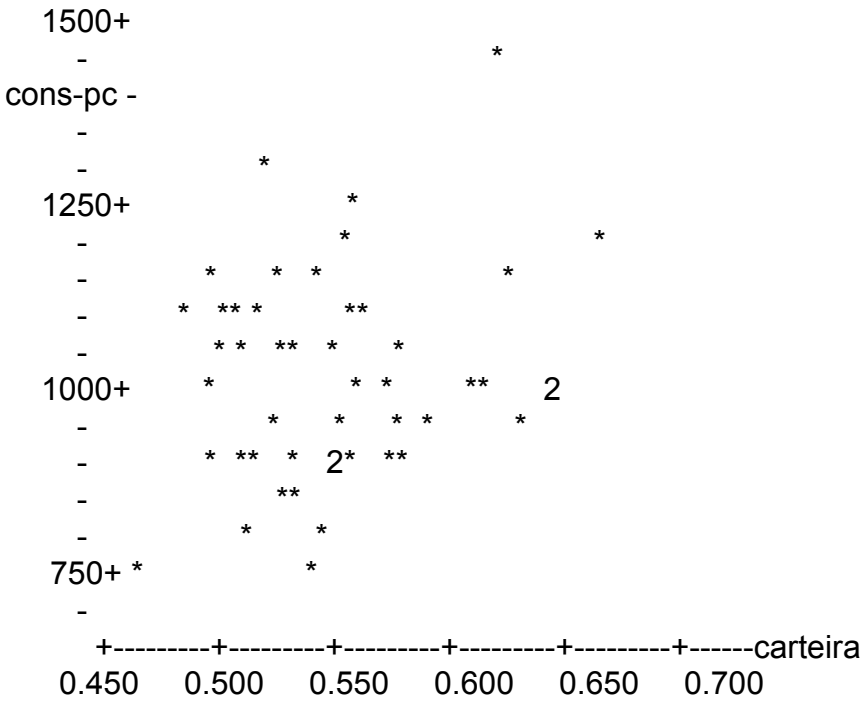
a) Obtenha o diagrama de dispersão entre a variável dependente e os demais regressores. Comente os padrões observados.

MTB > plot c1 c5



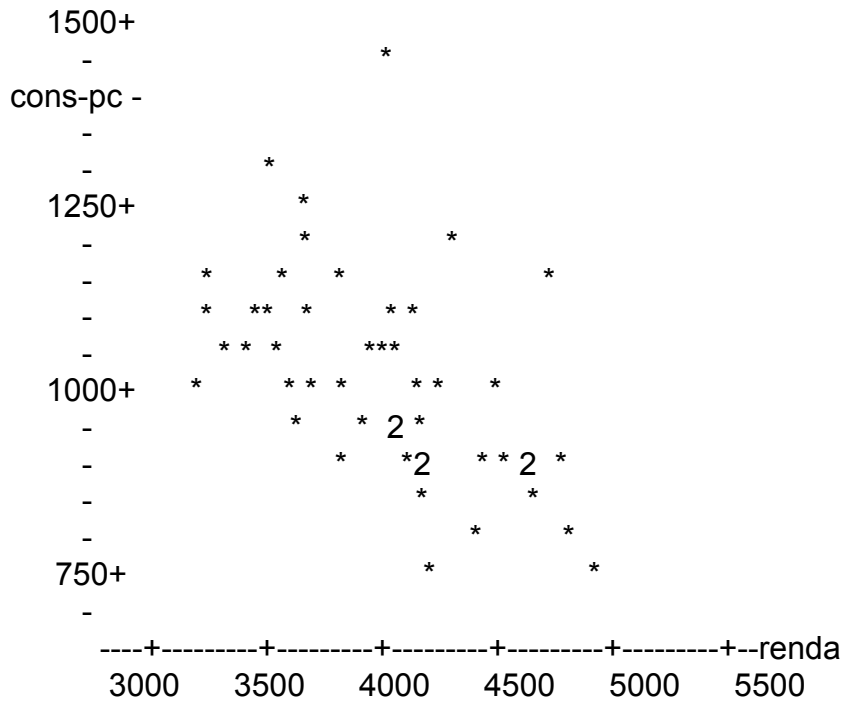
O gráfico do consumo per capita versus a taxa tem o padrão esperado. A relação entre as 2 variáveis é inversa.

MTB > plot c1 c7

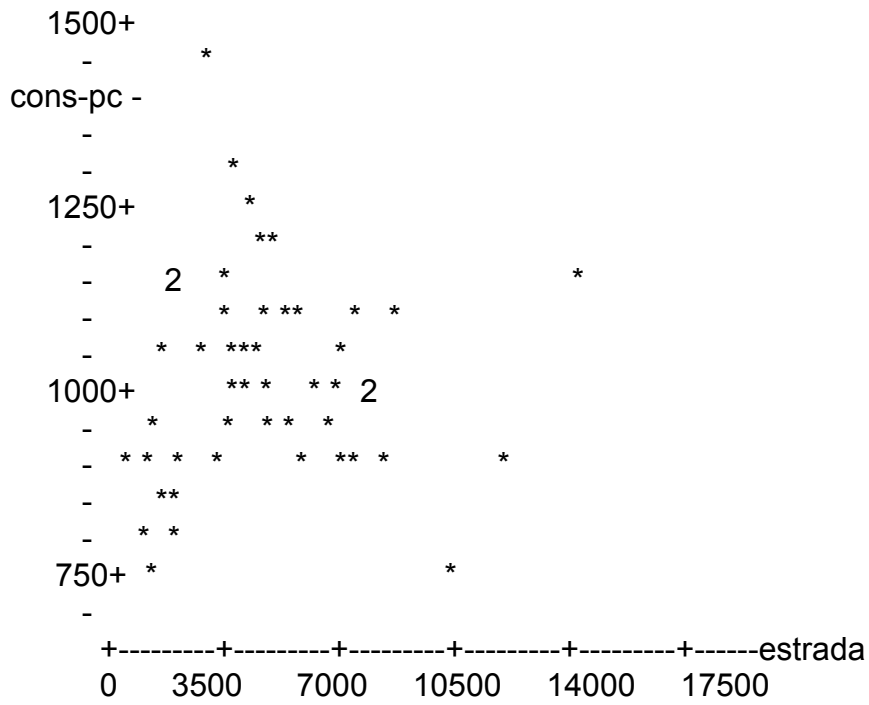


O consumo per capita de gasolina aumenta à medida em que cresce a proporção de motoristas com carteira na população do estado.

MTB > plot c1 c3



MTB > plot c1 c4



A relação entre consumo per capita e renda per capita, entretanto, não é a esperada. A relação observada entre as variáveis é o oposto do que seria intuitivamente razoável, e o consumo tende a ser grande em estados com baixa renda per capita. Note que aqui o consumo já está padronizado (per capita), ao contrário do exemplo 1.1.8.. Entretanto, ao contrário do que comentamos no exemplo 1.1.8., a mudança de escala não foi suficiente para transformar a relação entre estas 2 variáveis em numa relação direta.

O consumo tende a ser alto em estados com pequeno número de estradas federais, o que também não é facilmente explicável.

b) Obtenha a regressão de Y em X_1.

MTB > regr c1 1 c5

The regression equation is
cons-pc = 1477 - 61.2 taxa

Predictor	Coef	Stdev	t-ratio	p
Constant	1477.4	155.9	9.48	0.000
taxa	-61.21	20.17	-3.03	0.004

s = 131.5 R-sq = 16.7% R-sq(adj) = 14.9%

Analysis of Variance

SOURCE	DF	SS	MS	F	p
Regression	1	159195	159195	9.21	0.004
Error	46	795475	17293		
Total	47	954670			

Unusual Observations

Obs.	taxa	cons-pc	Fit	Stdev.Fit	Residual	St.Resid
6	10.0	800.4	865.2	50.7	-64.9	-0.53 X
18	7.0	1322.2	1048.9	23.3	273.4	2.11R
37	5.0	1130.7	1171.3	57.1	-40.5	-0.34 X
40	7.0	1440.5	1048.9	23.3	391.6	3.03R

R denotes an obs. with a large st. resid.
X denotes an obs. whose X value gives it large influence.

A regressão é claramente significante, mas existem diversas observações que não se ajustam bem a ela.

c) Obtenha a regressão de Y em X_2.

MTB > regr c1 1 c7

The regression equation is
cons-pc = 648 + 631 carteira

Predictor	Coef	Stdev	t-ratio	p
Constant	647.8	210.4	3.08	0.003
carteira	631.4	367.2	1.72	0.092

s = 139.6 R-sq = 6.0% R-sq(adj) = 4.0%

Analysis of Variance

SOURCE	DF	SS	MS	F	p
Regression	1	57652	57652	2.96	0.092
Error	46	897018	19500		
Total	47	954670			

Unusual Observations

Obs.	carteira	cons-pc	Fit	Stdev.Fit	Residual	St.Resid
18	0.540	1322.2	988.8	23.0	333.4	2.42R
19	0.724	1194.8	1105.0	59.9	89.8	0.71 X
40	0.672	1440.5	1072.1	42.4	368.3	2.77R

R denotes an obs. with a large st. resid.
X denotes an obs. whose X value gives it large influence.

Os mesmos comentários feitos em relação à regressão anterior se aplicam aqui. A regressão é significante, mas a qualidade do ajuste deixa a desejar.

d) Obtenha a regressão de Y em X_1 e X_2 simultaneamente. Verifique que o R^2 da regressão múltipla não pode ser obtido pela adição dos R^2 ´s das regressões simples. Explique o porque deste fato.

MTB > regr c1 2 c5 c7

The regression equation is
cons-pc = 1226 - 55.2 taxa + 359 carteira

Predictor	Coef	Stdev	t-ratio	p
Constant	1226.4	296.8	4.13	0.000
taxa	-55.18	21.07	-2.62	0.012
carteira	359.0	361.2	0.99	0.326

s = 131.5 R-sq = 18.5% R-sq(adj) = 14.8%

Analysis of Variance

SOURCE	DF	SS	MS	F	p
Regression	2	176284	88142	5.10	0.010
Error	45	778386	17297		
Total	47	954670			

SOURCE	DF	SEQ SS
taxa	1	159195
carteira	1	17089

Unusual Observations

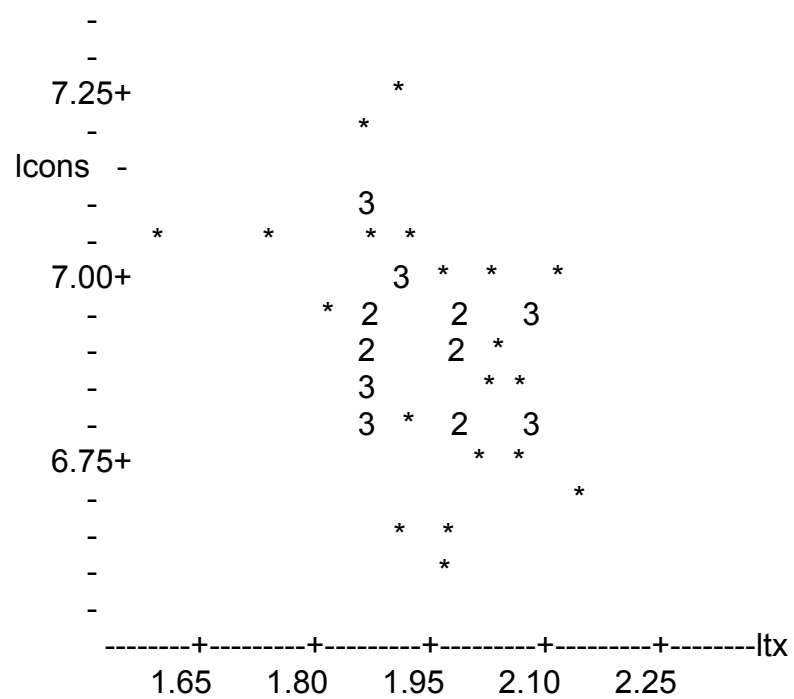
Obs.	taxa	cons-pc	Fit	Stdev.Fit	Residual	St.Resid
18	7.0	1322.2	1033.9	27.7	288.3	2.24R
37	5.0	1130.7	1153.6	59.8	-22.9	-0.20 X
40	7.0	1440.5	1081.3	40.1	359.2	2.87R

R denotes an obs. with a large st. resid.
X denotes an obs. whose X value gives it large influence.

A regressão é significante, e o R^2 é 18.5 %, diferente da soma dos R^2 's das regressões com cada variável individualmente. Isso se deve à própria definição do coeficiente de determinação.

A seguir analisamos novamente o problema, fazendo uma mudança de escala para todas as variáveis envolvidas. Esta desta mudança de escala (transformação logaritmo) é frequententemente usada para eliminar problemas de heteroscedasticidade.

MTB > plot c11 c12



O gráfico do logaritmo do consumo versus o logaritmo da taxa ainda exibe o mesmo tipo de relação inversa observado para as variáveis na escala original.

Os mesmos comentários se aplicam ao gráfico do logaritmo do consumo versus o logaritmo da proporção de motoristas habilitados na população, exibido a seguir.

```
MTB > plot c11 c13

        -
        -
    7.25+                                   *
        -                  *
 lcons  -
        -                      *   *                *
        -               *      * *            *
        -                *   ** *     * *
    7.00+                 * * **  **    *   *
        -                *           *     *    2
        -                     *     *   * *     *
        -                  * 2    * 3  **
        -                         **
    6.75+                         *
        -
        -    *           *
        -                  *
        -
        -
          ----+---------+---------+---------+---------+---------+--lcart
          -0.80     -0.70     -0.60     -0.50     -0.40     -0.30
```

e) Obtenha a regressão de lnY em ln X_1 e ln X_2. Verifique, através do procedimento estatístico adequado, se neste período a relação entre consumo e taxa é inelástica.

```
MTB > regr c11 2 c12 c13

The regression equation is
lcons = 7.84 - 0.406 ltx + 0.198 lcart

Predictor       Coef      Stdev    t-ratio       p
Constant      7.8430     0.3028      25.90   0.000
ltx          -0.4063     0.1551      -2.62   0.012
lcart         0.1984     0.2052       0.97   0.339

s = 0.1285      R-sq = 18.4%     R-sq(adj) = 14.7%

Analysis of Variance

SOURCE       DF         SS         MS        F       p
Regression    2    0.16700    0.08350     5.06   0.010
Error        45    0.74273    0.01651
Total        47    0.90973
```

```
SOURCE      DF      SEQ SS
ltx          1     0.15157
lcart        1     0.01543


Unusual Observations
Obs.     ltx     lcons       Fit Stdev.Fit  Residual   St.Resid
 18     1.95    7.1871    6.9301    0.0264    0.2570      2.04R
 37     1.61    7.0306    7.0762    0.0679   -0.0455     -0.42 X
 40     1.95    7.2727    6.9735    0.0380    0.2992      2.44R

R denotes an obs. with a large st. resid.
X denotes an obs. whose X value gives it large influence.
```

A resposta é : *não !* A taxa tem um efeito significante sobre o consumo per capita.

f) Se voce tivesse que escolher entre o modelo em d) e o modelo em e) qual modelo voce escolheria ? Justifique a sua resposta, usando argumentos estatísticos e extra-estatísticos.

Eu escolheria o modelo na escala dos logs, até por causa da interpretação mais clara em termos de elasticidade. Entretanto, para que esta escolha seja realmente justificada, é necessário olhar com atenção para os resíduos do modelo na escala original e transformada, e verificar se a transformação representa algum "ganho" em termos de tornar as hipóteses básicas do modelo de regressão mais realistas.

g) Ao melhor modelo encontrado na questão anterior, verifique se a adição conjunta das variáveis X_3 e X_4 (ou ln das variáveis) melhora significativamente o ajuste do modelo.

```
MTB > regr c11 4 c12 c13 c14 c15

The regression equation is
lcons = 12.4 - 0.345 ltx + 0.355 lcart - 0.560 lrenda + 0.0111 lestrad

Predictor       Coef      Stdev    t-ratio       p
Constant      12.390      1.064      11.64   0.000
ltx          -0.3449     0.1456      -2.37   0.022
lcart         0.3548     0.1716       2.07   0.045
lrenda       -0.5605     0.1136      -4.93   0.000
lestrad       0.01109    0.02352     0.47    0.640

s = 0.1044      R-sq = 48.5%     R-sq(adj) = 43.7%
```

Existe um ganho (em termos de R^2) devido à inclusão das novas variáveis. Note que log(estrada) não é significante, e poderia ser omitido da equação de regressão sem

perda substancial da capacidade explicativa do modelo. O coeficiente de log(renda) é negativo, e isso concorda com a relação expressa nos gráficos, embora seja de difícil interpretação.

Analysis of Variance

SOURCE	DF	SS	MS	F	p
Regression	4	0.44125	0.11031	10.13	0.000
Error	43	0.46847	0.01089		
Total	47	0.90973			

SOURCE	DF	SEQ SS
ltx	1	0.15157
lcart	1	0.01543
lrenda	1	0.27183
lestrad	1	0.00242

Unusual Observations

Obs.	ltx	lcons	Fit	Stdev.Fit	Residual	St.Resid
5	2.08	6.6250	6.8224	0.0551	-0.1974	-2.23R
40	1.95	7.2727	6.9748	0.0312	0.2980	2.99R

R denotes an obs. with a large st. resid.

Regressão stepwise

Um problema importante em regressão múltipla é: como escolher dentre muitas possíveis variáveis explicativas. Uma saída possível é usar regressão stepwise, sempre mantendo em mente os comentários a seguir.

O procedimento stepwise permite selecionar variáveis a partir de um conjunto inicial de variáveis explicativas. A escolha de variáveis se baseia num procedimento heurístico que, apesar de intuitivamente razoável, não garante que a regressão encontrada possui o maior R^2 , nem que o modelo encontrado é o melhor, do ponto de vista prático. Entretanto, o procedimento é útil nos estágios iniciais de análise, especialmente quando existe um número muito grande de possíveis variáveis explicativas.

A escolha das variáveis é feita a partir da estatística F de cada variável, onde F é o quadrado do " t-ratio " da varíavel.

Características do procedimento no MINITAB

É necessário especificar 2 valores positivos, FENTER e FREMOVE. Uma variável será incluída no modelo se a estatística F correspondente exceder FENTER, obedecendo ao procedimento mostrado a seguir. O valor FREMOVE é usado para remover variáveis ao

longo do procedimento, e uma variável será excluída se sua estatística F for menor que FREMOVE.

Note que FENTER ≥ FREMOVE sempre.

Passo 1 - ajustar todos os modelos com apenas 1 variável explicativa, e escolher aquela com o maior valor da estatística F, desde que F > FENTER.

Passo 2 - ajustar todos os modelos com 2 variáveis, sendo uma delas a variável incluída no passo 1. Incluir a variável com maior valor de F, desde que F > FENTER.

Passo 3 - verificar se alguma variável (exceto a incluída no passo 2) pode ser retirada do modelo. A variável com menor valor de F é excluída, desde que F < FREMOVE.

Passo 4 - se nenhuma variável é removida, STEPWISE tenta inserir uma nova variável. O procedimento se repete, e variáveis são incluídas se F > FENTER, e removidas se F < FREMOVE. O procedimento continua até que nenhuma variável seja incluída ou removida do modelo.

Sintaxe do comando STEPWISE no MINITAB

STEPWISE y na coluna C, variáveis nas colunas C, C , ... , C

Subcomandos :

FENTER = k (o default é k = 4)

FREMOVE = k (o default é k = 4)

FORCE C, C , ... , C (incluir estas variáveis no primeiro passo e não excluí-las posteriormente).

ENTER C, C , ... , C (incluir estas variáveis no primeiro passo, permitindo uma posterior exclusão).

Casos Especiais

1) Forward Selection

As variáveis são inseridas da mesma maneira que no comando STEPWISE, mas nunca são removidas. O procedimento termina quando nenhuma variável tem F > FENTER. A implementação no MINITAB é :

STEPWISE y na coluna C, variáveis nas colunas C, C , ... , C
FREMOVE = 0 (nenhuma variável será removida)

2) Backward Elimination

A partir de todas as variáveis disponíveis, elimina-se uma variável a cada passo. Não se permite o reingresso de variáveis. O procedimento termina quando nenhuma variável pode mais ser removida, isto é, quando todas as variáveis no modelo apresentam F > FREMOVE.

A implementação no MINITAB é :

STEPWISE y na coluna C, variáveis nas colunas C, C , ... , C
ENTER C, C, ... , C (para forçar a inclusão de todas as variáveis no primeiro passo).
FENTER = 10000 (ou outro número grande, para impedir a inclusão de variáveis no procedimento STEPWISE)

1.3. Introdução aos Modelos Lineares Generalizados

Os Modelos Lineares Generalizados (MLG's) combinam, dentro de uma mesma estrutura, diversos modelos usuais em Estatística, como os modelos de regressão linear, logística , modelos log-lineares, análise de variância e covariância, entre outros. Na verdade, esta capacidade de "unificação" de modelos tão distintos quanto os mencionados acima é uma das características principais dos MLG's.

Antes de prosseguir, é interessante caracterizar as principais componentes dos modelos que serão aqui estudados. Em muitos problemas de interesse em Estatística procuramos estudar a relação entre uma variável de "resposta" (ou variável dependente) Y e um conjunto de variáveis explicativas X_1, X_2, ..., X_{p-1}. Estas variáveis explicativas são freqüentemente chamadas de variáveis "independentes" , preditores ou co-variáveis. Regressão linear múltipla é o exemplo mais simples de um modelo com esta estrutura. Como podemos escrever um modelo de regressão linear múltipla ?

Regressão Linear Múltipla

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_{p-1} x_{p-1} + \varepsilon \qquad (1.3.1)$$

Quais as características principais deste modelo?

1- Y (a variável de resposta) é uma variável aleatória contínua.

2- X_1, X_2, ..., X_{p-1} são encarados como valores fixos (não aleatórios). A relação entre Y e X_1, X_2, ..., X_{p-1} representa a *parte sistemática* do modelo, e esta relação é linear, segundo a equação (1.3.1).

3- ε é um ruído aleatório, e supõe-se que ele tem média zero e variância constante. A presença de ε na equação (1.3.1) representa a parte aleatória do modelo.

4- Muitas vezes supomos também que ε é uma variável Normal, e então Y também é Normal. Isto é conveniente para que possamos produzir intervalos de confiança e testes de hipóteses.

Tomando o valor esperado na expressão (1.3.1) encontramos:

$$E(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_{p-1} x_{p-1} \qquad (1.3.2)$$

A expressão (1.3.2) nos diz que a média de Y, μ, e as variáveis explicativas X_1, X_2, ..., X_{p-1} exibem uma *relação linear*. Num caso mais geral poderíamos supor que uma função de μ, g(μ), estaria "ligada" aos preditores X_1, X_2, ..., X_{p-1}.

Suponha agora que observamos Y para n indivíduos diferentes. Então a equação (1.3.1) deve ser aplicada a cada um dos indivíduos, e a hipótese convencional sobre os ruídos ε_1, ε_2, ..., ε_n é a de média zero, variância constante e correlação nula entre quaisquer ε_i e ε_j , com $i \neq j$.

Em resumo, num modelo de regressão linear múltipla temos a seguinte estrutura:
1- variável dependente contínua,
2- relação entre a variável independente e os preditores é linear,
3- relação entre a média da variável independente e os preditores é linear,
4- a distribuição dos erros é Gaussiana.

Qualquer modificação nestas componentes gera um modelo que não pode mais ser chamado de "modelo de regressão linear". Na verdade, ao considerarmos MLG's estaremos admitindo que estas 4 características possam ser diferentes do que está exposto acima. Por exemplo, poderíamos estipular uma relação não linear entre Y e os preditores, ou a distribuição dos erros não seria Gaussiana, ou ainda Y não seria uma variável contínua.

É importante observar quais são os “tipos” possíveis de variáveis resposta existentes, pois a identificação destes tipos nos leva a considerar classes de modelos diferentes.

Tipos de Variáveis

Variável	Exemplos
Categórica com 2 valores possíveis (variável binária)	homem/mulher, sim/não, presença ou ausência de uma doença, comprou ou não um bem de consumo, vivo ou morto, etc ...
Categórica com mais de 2 grupos não ordenados (variável nominal)	tipo de sangue, causa mortis, religião, partido político, bairro de residência, estado civil, etc...
Categórica com valores ordenados (variável ordinal)	graus de severidade de uma doença, ano na Universidade, pressão sanguínea (categorizada por grupos, por exemplo, < 70, entre 71 e 90, entre 91 e 110 , entre 111 e 130, acima de 130, etc...
Contínua	peso, altura, tempo de duração de uma lâmpada, tempo de sobrevivência após o diagnóstico de uma doença, etc...

Discreta (contagens)	número de filhos numa família, número de acidentes de trânsito num ano, número de chamadas não completadas num tronco telefônico, etc...

O termo variável quantitativa é geralmente empregado para designar variáveis contínuas, e variáveis nominais (e às vezes ordinais) são muitas vezes chamadas de variáveis qualitativas. As variáveis explicativas qualitativas são chamadas fatores, e as suas categorias são chamadas de níveis.

Por exemplo, suponha que Y é uma variável binária, onde Y = 0 se uma pessoa comprou um carro de luxo numa concessionária este mês, e Y = 1 do contrário. É claro que seria interessante "explicar" a compra em termos de outras variáveis, como : idade do cliente, sexo do cliente, profissão, local de residência, número de filhos, estado civil, renda familiar, se possui ou não apartamento próprio, se é cliente habitual da concessionária, etc ... O efeito de cada uma destas variáveis sobre uma venda potencial é de extremo interesse para qualquer pessoa inserida no mercado de automóveis de luxo, pois ajuda a identificar o "target" deste tipo de produto. Em particular, também nos interessa tentar “prever” a probabilidade de um indivíduo comprar o automóvel sabendo qual o seu perfil em termos das variáveis explicativas.

A combinação de diversos tipos de variáveis de resposta e preditores leva a diversos modelos estatísticos. O quadro a seguir exibe algumas das combinações mais comuns, e os modelos estatísticos decorrentes.

Variável de resposta →	Binária	Nominal com mais de 2 categorias	Contínua
Variável explicativa ↓			
Binária	Regressão logística, modelos de resposta a dosagens (por exemplo, análise probit) , comparação de proporções	tabelas de contingência, modelos log-lineares	análise de variância
Nominal com mais de 2 categorias	idem acima	idem acima	análise de variância
Contínua	idem acima	idem acima	regressão múltipla e análise de covariância
Mistura de tipos	idem acima	idem acima	regressão múltipla e análise de covariância

Modelos Lineares

Os modelos lineares incluem como casos particulares os modelos de regressão simples e múltipla e análise de variância.

Notação

$y = (y_1, y_2, y_3, ..., y_n)^t$ representa a variável "dependente" (variável de resposta), onde y_i refere-se à resposta medida no i-ésimo indivíduo.

X é uma matriz de “design”, cujas colunas são os valores das variáveis explicativas. A forma da matriz X indica a estrutura do modelo linear, e representa a parte sistemática do modelo. No caso de modelos de análise de variância, os elementos de X são zero ou um, e nos modelos de regressão linear a primeira coluna de X geralmente é composta de 1's, no caso do modelo incluir o termo constante. Nos modelos de regressão linear, as colunas de X representam variáveis explicativas contínuas.

$\varepsilon = (\varepsilon_1 , \varepsilon_2 , ..., \varepsilon_n)^t$ é o vetor de ruídos aleatórios, e representa a componente aleatória do modelo.

$\beta = (\beta_0 , \beta_1, \beta_2, ..., \beta_{p-1})^t$ é o vetor de parâmetros a serem estimados.

Um modelo linear pode então ser escrito como:

$$y = X\beta + \varepsilon \qquad (1.3.3)$$

O modelo (1.3.3) é dito linear por que nele a parte determinística é uma função linear dos parâmetros desconhecidos e também a parte aleatória do modelo é aditiva.

No caso de variáveis explicativas quantitativas, o modelo (1.3.3) contém termos da forma $\beta_i.x_i$ onde o parâmetro β_i representa a taxa de variação da variável resposta correspondendo a mudanças no preditor x_i .

Se as variáveis explicativas são qualitativas, os parâmetros representam os diversos níveis dos fatores.

Exemplo 1.3.1.

Os dados a seguir representam o peso em gramas de 2 amostras de 10 plantas cultivadas sob 2 condições experimentais diferentes.

Controle	4.17	5.58	5.18	6.11	4.50	4.61	5.17	4.53	5.33	5.14
Tratamento	4.81	4.17	4.41	3.59	5.87	3.83	6.03	4.89	4.32	4.69

O objetivo aqui é descrever o efeito do tratamento aplicado sobre o crescimento das plantas. Um possível modelo neste caso é :

$$Y_{jk} = \mu_j + \varepsilon_{jk} \qquad (1.3.4)$$

onde:
Y_{jk} é o peso da k-ésima planta (k = 1, 2, ..., 10) da j-ésima amostra (j = 1, 2), onde j = 1 indica o tratamento usual (controle) e j = 2 representa o tratamento novo.
μ_j é um parâmetro que indica o peso esperado das plantas em cada uma das amostras.
ε_{jk} é um termo de erro aleatório. Supomos que os ε_{jk} são independentes e identicamente distribuídos com densidade $N(0, \sigma^2)$.

Uma conseqüência direta desta suposição é $E(Y_{jk}) = \mu_j$, isto é, o peso médio das plantas da amostra j é μ_j.

Gostaríamos de saber se o novo tratamento foi eficiente no sentido de aumentar o peso médio das plantas cultivadas, e então a quantidade de interesse aqui é $\mu_2 - \mu_1$. Devemos testar a hipótese desta diferença ser ou não estatisticamente diferente de zero.

O modelo (1.3.4) pode ser especificado de maneira equivalente como:

$$Y_{jk} = \mu + \alpha_j + \varepsilon_{jk} \qquad (1.3.5)$$

onde Y_{jk} e ε_{jk} têm a mesma representação que em (1.3.4) e agora μ indica um fator representando o crescimento comum às duas amostras.

Desta forma $\alpha_j = \mu_j - \mu$ (j=1,2) indica o efeito do j-ésimo tratamento. Se estes efeitos não são significativos, o modelo (1.3.5) reduz-se a :

$$Y_{jk} = \mu + \varepsilon_{jk} \qquad (1.3.6)$$

Logo, o teste da hipótese $\mu_2 - \mu_1 = 0$ é equivalente a testar $\alpha_1 = \alpha_2 = 0$, isto é, testar se os efeitos dos tratamentos são insignificantes.

A verossimilhança sob o modelo (1.3.4) é dada por:

$$\prod_{j=1}^{2}\prod_{k=1}^{10} \frac{1}{\left(2\pi\sigma^2\right)^{1/2}} . \exp\left\{\frac{-1}{2\sigma^2}\left(y_{jk} - \mu_j\right)^2\right\}$$

A log-verossimilhança é:

$$l = -10.\log\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\sum_{j=1}^{2}\sum_{k=1}^{10}\left(y_{jk} - \mu_j\right)^2$$

De maneira análoga, a log-verossimilhança sob o modelo (1.3.6) é:

$$l = -10.\log\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\sum_{j=1}^{2}\sum_{k=1}^{10}\left(y_{jk} - \mu\right)^2$$

O próximo passo é estimar os parâmetros desconhecidos. A estimação destes parâmetros é feita por máxima verossimilhança, e pode-se mostrar que , sob o modelo (1.3.4):

$$\mu_j = \frac{1}{10}\sum_{k=1}^{10} y_{jk} = y_j$$

onde j = 1, 2

O estimador do nível de crescimento comum aos dois tratamentos, μ, é dado por: $\mu = \frac{1}{20}\sum_j\sum_k y_{jk} = y_{..}$

O estimador da variância σ^2 sob o modelo (1.3.4) é um múltiplo de:

$$S_1 = \sum_{j=1}^{2}\sum_{k=1}^{10}\left(y_{jk} - y_j\right)^2$$

Por outro lado, o estimador de σ^2 sob o modelo mais simples (1.3.6) é múltiplo de:

$$S_0 = \sum_{j=1}^{2}\sum_{k=1}^{10}\left(y_{jk} - y_{..}\right)^2$$

Sob o modelo mais simples não há efeitos devidos aos tratamentos, e então intuitivamente, os y_j devem estar próximos entre si e próximos da média computada utilizando-se as 2 amostras, $y_{..}$.

Logo, se o modelo mais simples (1.3.6) é verdadeiro, as estimativas S_0 e S_1 são também semelhantes. Ao contrário, se o modelo verdadeiro é (1.3.4), e existem diferenças entre o grupo de controle e o tratamento, S_0 deve ser bem maior que S_1, pois S_0 calcula discrepâncias em relação ao valor médio sob cada um dos j tratamentos, ao invés de calculá-las em relação à média global das 2 amostras.

O teste das hipóteses:
$H_0 : \mu_1 = \mu_2$ (isto é, o modelo verdadeiro é o mais simples) versus
$H_1 : \mu_1 \neq \mu_2$ (o modelo verdadeiro é (1.3.4))
é feito usando-se uma estatística F dada por :

$$F = \frac{S_0 - S_1}{S_1/(2K-2)}$$

onde K é o número de observações em cada amostra (neste caso K =10).

Sob a hipótese nula H_0, a distribuição da estatística F acima é uma distribuição F de Snedecor com 1 grau de liberdade no numerador e 2K-2 graus no denominador.

Neste exemplo, a estatística F é dada por :

$$F = \frac{(9.417 - 8.729)}{8.729 / 18} = 1.42$$

Este valor não é estatisticamente significante, e assim não podemos rejeitar a hipótese nula, isto é, não podemos afirmar que existe diferença significativa entre os pesos das plantas nos 2 grupos.

O teste usual de hipóteses neste caso utiliza uma distribuição t, e é equivalente ao teste F mostrado acima. A grande vantagem em usar uma estatística F está em sua generalidade - o mesmo procedimento pode ser utilizado, com pequenas alterações, na comparação de mais de duas médias.

Os modelos (1.3.4) e (1.3.6) podem ser facilmente colocados na forma matricial (1.3.3).

O modelo (1.3.4) tem a forma y = Xβ + ε onde:

$$y = \begin{pmatrix} y_{11} \\ y_{12} \\ \dots \\ y_{1K} \\ y_{21} \\ \dots \\ y_{2K} \end{pmatrix}, \beta = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \mathrm{X} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ \dots & \\ 1 & 0 \\ 0 & 1 \\ \dots & \\ 0 & 1 \end{pmatrix}, \varepsilon = \begin{pmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \dots \\ \varepsilon_{1K} \\ \varepsilon_{21} \\ \dots \\ \varepsilon_{2K} \end{pmatrix}$$

O modelo simplificado (1.3.4) tem a forma:

$$y = \begin{pmatrix} y_{11} \\ y_{12} \\ \dots \\ y_{1K} \\ y_{21} \\ \dots \\ y_{2K} \end{pmatrix}, \beta = (\mu), \mathrm{X} = \begin{pmatrix} 1 \\ 1 \\ \dots \\ 1 \\ 1 \\ \dots \\ 1 \end{pmatrix}, \varepsilon = \begin{pmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \dots \\ \varepsilon_{1K} \\ \varepsilon_{21} \\ \dots \\ \varepsilon_{2K} \end{pmatrix}$$

Famílias Exponenciais e Modelos Lineares Generalizados

Por muitas décadas, o modelo linear y = Xβ + ε foi a "ferramenta de trabalho" tradicional dos estatísticos. Este modelo engloba diversos modelos tradicionais, como regressão e análise de variância. Entretanto, este tipo de relação linear não é apropriada em muitas situações.

Por exemplo, suponha que desejamos modelar a probabilidade de compra de um certo eletrodoméstico como função de diversos indicadores, como renda, idade, estado civil, etc... A variável dependente é p, a probabilidade de compra do produto, e claramente é necessário fazer algum tipo de transformação, pois um modelo da forma p = Xβ + ε não é apropriado, já que p está restrito ao intervalo (0,1).

Os modelos lineares generalizados estendem os modelos lineares da forma Y = Xβ + ε em duas direções:

A densidade das observações Y não é mais Normal, e é um membro da família exponencial, definida abaixo.

A relação entre a média de Y, μ, e os preditores não é mais da forma μ = Xβ, e depende de uma função g(.) especificada tal que g(μ) = Xβ. Esta função g(.), chamada de *função de ligação* é uma função conhecida, contínua e diferenciável.

Definição 1.3.2. (Família Exponencial)

Dizemos que a variável aleatória Y tem distribuição na família exponencial (univariada) com parâmetro θ se a densidade de Y pode ser escrita como :

$$f(y) = \exp\big(a(y).b(\theta) + c(\theta) + d(y)\big) \qquad (1.3.7)$$

b(θ) é chamado de **parâmetro natural**, e em muitas distribuições usuais temos b(θ) = θ. Neste caso dizemos que (1.3.7) representa uma família exponencial natural.

Os próximos exemplos indicam que algumas das distribuições de probabilidade mais comuns são famílias exponenciais, isto é, podem ser colocadas na forma (1.3.7).

Se a densidade (1.3.7) envolver outros parâmetros além de θ, estes serão considerados como "nuisance parameters", e encarados como constantes fixas.

Exemplo 1.3.3. (Densidade Normal)

Seja Y ~ N(μ, σ^2) onde supomos σ^2 fixo e conhecido. A densidade de Y é dada por:

$$f(y) = \left(2\pi\sigma^2\right)^{-1/2}.\exp\left\{\frac{-(y-\mu)^2}{2\sigma^2}\right\}$$

$$f(y) = \left(2\pi\sigma^2\right)^{-1/2}.\exp\left\{\frac{-y^2 + 2\mu y - \mu^2}{2\sigma^2}\right\}$$

Igualando esta última expressão a (1.3.7) leva a:

b(θ) = θ = μ
a(y) = y/σ2
c(θ) = - μ2/2σ2 = - θ2/2σ2
d(y) = (-1/2).{ log(2πσ2) + y2/σ2 }

Exemplo 1.3.4. (Poisson)
Seja Y ~ Poisson(θ). Então:

$$f(y) = \frac{e^{-\theta}.\theta^y}{y!} = \exp\{-\theta + y.\log(\theta) - \log(y!)\}$$

Neste caso:

b(θ) = log(θ)
a(y) = y
c(θ) = -θ
d(y) = -log(y!)

Exemplo 1.3.5. (Binomial)
Seja Y ~ Bin(n,p). A densidade de Y pode ser escrita como :

$$f(y) = \binom{n}{y} p^y (1-p)^{n-y} =$$

$$= \exp\left\{\log\binom{n}{y} + y\log(p) + (n-y)\log(1-p)\right\} =$$

$$\exp\left\{\log\binom{n}{y} + y.\log\left(\frac{p}{1-p}\right) + n\log(1-p)\right\}$$

Neste caso:

$$\theta = \log\left(\frac{p}{1-p}\right) \Rightarrow p = \frac{e^\theta}{1+e^\theta}$$

é o parâmetro natural, e é chamado de "logit" de p. Também : a(y) = y, c(θ) = n.log(1-p) = - n.log(1+exp(θ)) e

$$d(y) = \log\binom{n}{y}$$

Note que θ (parâmetro natural) é uma função de p, e θ é um valor no intervalo (-∞, +∞). Então, faz sentido empregar um modelo linear relacionando θ às covariáveis, isto é, alguma coisa da forma:

$$\log\left(\frac{p}{1-p}\right) = X\beta + \varepsilon$$

pois a variável dependente neste modelo não está restrita ao intervalo (0,1).

Exemplo 1.3.6. (Gama)
Seja Y ~ Gama(α, β) onde α indica o parâmetro de forma e β é o parâmetro de escala. Suponha que α é conhecido. Então a densidade de Y é:

$$f(y) = \frac{y^{\alpha-1}.\beta^{\alpha}.e^{-\beta y}}{\Gamma(\alpha)} = \exp\left\{-\beta y - \log(\Gamma(\alpha)) + \alpha\log(\beta) + (\alpha-1)\log(y)\right\}$$

Neste caso:

a(y) = -y
b(θ) = β = θ
c(θ) = α.log(θ)
d(y) = (α-1) log(y)

Se uma distribuição é membro de uma família exponencial torna-se bastante fácil calcular os dois primeiros momentos de a(y). Isso é mostrado no teorema a seguir.

Teorema 1.3.7.
Seja Y uma variável aleatória com densidade na família exponencial dada por (1.3.7). Então :

E(a(Y)) = - c' (θ)/b'(θ)

e

$$VAR(a(Y)) = \frac{\left[b''(\theta)c'(\theta) - c''(\theta)b'(\theta)\right]}{\left(b'(\theta)\right)^3}$$

onde as derivadas são calculadas em relação a θ.

Demonstração
Considere o logaritmo da expressão (1.3.7) :

$$l = \log(f(y)) = a(y).b(\theta) + c(\theta) + d(y) \qquad (1.3.8)$$

O Score de Fisher (denotado por S) é a primeira derivada de l com relação a θ. Pode-se provar facilmente que:

E(S) = 0 e VAR(S) = E(S^2) = E(-dS/dθ) = E(-d^2l(θ)/dθ^2)

Aplicando estes resultados à log-verossimilhança (1.3.8) resulta em:

l(θ) = a(y).b(θ) + c(θ) + d(y)
S = dl/dθ = a(y).b' (θ) + c'(θ)
E(S) = 0 $\Rightarrow$ E{ a(y).b' (θ) + c'(θ) } = 0
E(a(y)) = - c'(θ) / b'(θ)

Também:
VAR(S) = E(- S') = E{ - a(y).b"(θ) - c"(θ) }

= - b''(θ).E(a(y)) - c''(θ) =
= - b''(θ){- c'(θ) / b'(θ) } - c''(θ) (1.3.9)

Mas :
VAR(S) = VAR{ a(y).b' (θ) + c'(θ) } = {b' (θ)}^{2}.VAR(a(y)) (1.3.10)

Igualando as expressões (1.3.9) e (1.3.10) resulta em:

$$VAR(a(Y)) = \frac{[b''(\theta)c'(\theta) - c''(\theta)b'(\theta)]}{(b'(\theta))^3}$$

No caso particular a(y) = y o resultado do teorema se reduz a:

μ = E(Y) = - c' (θ)/b'(θ)

e

$$VAR(Y) = \frac{[b''(\theta)c'(\theta) - c''(\theta)b'(\theta)]}{(b'(\theta))^3}$$

Note que a média e variância de Y são ambas funções de θ, e isso nos permite expressar a variância de Y como função de sua média μ.

Exemplo 1.3.8.
Seja Y ~ Poisson(θ), como no exemplo 1.3.4.. Então : a(y) = - y, b(θ) = log(θ), c(θ) = -θ, d(y) = -log(y!)

Assim:

b' (θ) = 1/θ
b'' (θ) = -1/θ^2
c'(θ) = -1
c''(θ) = 0

Usando os resultados do teorema encontramos:
E(a(Y)) = E(Y) = - c' (θ)/b'(θ) = - (-1)/(1/θ) = + θ

$$VAR(a(Y)) = \frac{[b''(\theta)c'(\theta) - c''(\theta)b'(\theta)]}{(b'(\theta))^3}$$

$$VAR(Y) = \frac{1}{(1/\theta)^3}.\left[\frac{-1}{\theta^2}.(-1) - 0\right] = \theta^3.\left[\frac{+1}{\theta^2}\right] = \theta$$

Nota
Suponha agora que Y_1, Y_2, ..., Y_n formam uma amostra aleatória de uma densidade na família exponencial. Então a verossimilhança pode ser escrita como:

$$L(\theta) = \exp\left[b(\theta).\sum_{i=1}^{n} a(y_i) + n.c(\theta) + \sum_{i=1}^{n} d(y_i)\right]$$

e a log-verossimilhança é:

$$l(\theta) = \left[b(\theta).\sum_{i=1}^{n} a(y_i) + n.c(\theta) + \sum_{i=1}^{n} d(y_i) \right]$$

Então $\Sigma a(y_i)$ é uma estatística suficiente para $b(\theta)$, e condensa toda a informação contida na amostra sobre o parâmetro θ. Este resultado será importante quando abordarmos a questão de estimação de parâmetros.

Definição 1.3.9. (Modelo Linear Generalizado)
Sejam $Y_1, Y_2, ..., Y_n$ uma coleção de variáveis aleatórias independentes cujas distribuições são membros de uma família exponencial na forma canônica, isto é : $a(Y_i) = Y_i$. Então a densidade conjunta de $Y_1, Y_2, ..., Y_n$ pode ser escrita como :

$$f(y_1, y_2, ..., y_n, \theta_1, ..., \theta_n) =$$
$$= \exp\left\{ \sum_{i=1}^{n} y_i.b(\theta_i) + \sum_{i=1}^{n} c(\theta_i) + \sum_{i=1}^{n} d(y_i) \right\} \qquad (1.3.11)$$

Note que, a princípio, os parâmetros naturais θ_i podem ser diferentes para cada Y_i.

As variáveis $Y_1, Y_2, ..., Y_n$ são variáveis de resposta, a serem explicadas a partir de um conjunto de p covariáveis (variáveis explicativas).

A relação entre cada Y_i e as p covariáveis é expressa através de :

$$\eta_i = g(\mu_i) = x_i^t.\beta \qquad (1.3.12)$$

onde :

$\mu_i = E(Y_i)$ é a média de cada variável dependente,

$\beta = (\beta_1, \beta_2, ..., \beta_p)^t$ é um vetor de parâmetros desconhecidos a serem estimados,

x_i^t é um vetor de dimensão 1 x p de variáveis explicativas,

g(.) é uma função conhecida, monótona e diferenciável, chamada de *função de ligação*. A função de ligação geralmente é não linear.

$\eta_i = x_i^t.\beta$ é chamado de preditor linear.

A equação (1.3.12) é chamada de equação de ligação, pois relaciona as componentes sistemática ($x_i^t.\beta$) e aleatória ($E(Y_i)$) do modelo.

Em resumo, um Modelo Linear Generalizado pode ser caracterizado por três componentes:

1- Uma variável de resposta Y_i (i=1,2,..,n) que tem distribuição na família exponencial,

2- Um conjunto de p-1 covariáveis e parâmetros a serem estimados β_0, β_1, β_2, ..., β_{p-1} que representam a componente sistemática do modelo,

3- Uma função de ligação g(.) conhecida, tal que : $g(\mu_i) = \eta_i = x_i^t.\beta$.

A função de ligação relaciona as componentes aleatória e sistemática do modelo.

Num Modelo Linear Generalizado a variância da variável de resposta Y não é constante, mas depende do valor da média. Isso representa uma extensão dos modelos de regressão usuais, onde a variância é constante. Note que, no caso da distribuição Normal, a média e a variância da densidade não apresentam qualquer relação. O mesmo não ocorre em outras densidades na família exponencial. Por exemplo, no caso Poisson, VAR(Y) = E(Y) = μ. Para a distribuição Binomial, E(Y) = μ = n.p e VAR(Y) = n.p(1-p) = μ.(1-μ/n), etc...

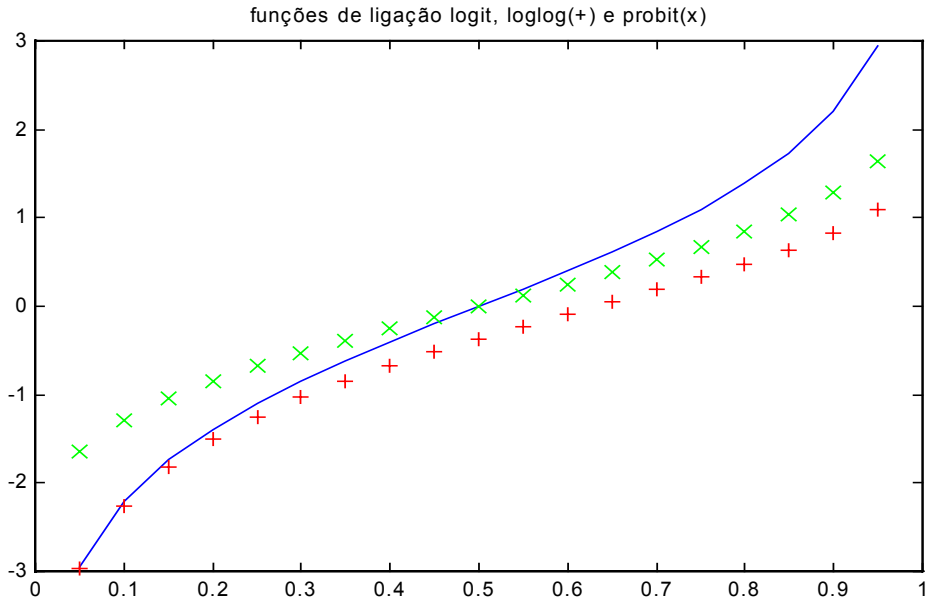
A escolha de diferentes funções de ligação leva a Modelos Lineares Generalizados com estruturas diferentes. A função de ligação deve ser escolhida de maneira compatível com a estrutura do erro e de maneira a tornar o MLG resultante facilmente interpretável. As ligações mais comuns são :

1- Potência : $\eta = g(\mu) = \mu^k$ onde k é um número real,
2- Logística : $\eta = g(\mu) = \log\{ \mu/(1-\mu) \}$,
3- Probit : $\eta = g(\mu) = \Phi^{-1}(\mu)$ onde Φ indica a função de distribuição N(0,1),
4- Log-log complementar : $\eta = g(\mu) = \log\{-\log(1- \mu) \}$.
5- Logaritmo : $\eta = g(\mu) = \log(\mu)$

As ligações logística, probit e log-log complementar são apropriadas no caso do modelo Binomial, pois transformam o intervalo (0,1) no conjunto de números reais, isto é, $\mu \in (0,1)$ é levado em $\eta = X\beta \in (-\infty, \infty)$. A ligação logaritmo é freqüentemente empregada na análise de dados categorizados (distribuição Poisson).

As ligações potência mais usuais ocorrem quando k = 1 (ligação identidade), k = 1/2 (raiz quadrada) e k = -1 (recíproco).

Exemplo 1.3.10.
Comparação das funções de ligação logística, probit e log-log complementar.
A seguir exibimos o gráfico das funções :
1-logit(μ) = $\log\{ \mu/(1-\mu) \}$
2- probit(μ) = $\Phi^{-1}(\mu)$
3- loglog(μ) = $\log\{-\log(1- \mu) \}$
onde μ é um número qualquer no intervalo (0,1).



Exemplo 1.3.11.
O modelo linear : $Y = X\beta + \varepsilon$ onde $Y = (Y_1, Y_2, ..., Y_n)^t$, $\beta = (\beta_1, \beta_2, ..., \beta_p)^t$ e $\varepsilon = (\varepsilon_1, \varepsilon_2, ..., \varepsilon_n)^t$ é um caso particular de um Modelo Linear Generalizado, onde a distribuição da variável dependente é Normal e a função de ligação é a identidade, isto é : $g(\mu_i) = \mu_i = E(Y_i) = X\beta$.

No modelo linear a estimativa do vetor de parâmetros β é feita diretamente, e não envolve cálculos recursivos, como em todos os outros Modelos Lineares Generalizados.

Exemplo 1.3.12. (Regressão Logística)
Seja $Y_i \sim Bin(n, p_i)$ para I=1, 2,..., n. Um modelo usual nesta situação é o modelo de regressão logística, no qual :

$$\log\left(\frac{p_i}{1-p_i}\right) = x_i^t.\beta \qquad (1.3.13)$$

De maneira equivalente podemos escrever :

$$\frac{p_i}{1-p_i} = \exp\left(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_{p-1} x_{p-1}\right)$$

Note que neste caso $\mu_i = E(Y_i) = n.p_i$ e então a equação (1.3.13) relaciona uma função não linear da média às covariáveis .

Um outro modelo possível (mas muito menos atraente) para dados binomiais é :

$$p_i = x_i^t.\beta$$

Por que este modelo não é atraente ? Note que p_i é uma probabilidade, e está restrito ao intervalo (0,1). Por outro lado, qualquer combinação linear da forma $x_i^t.\beta$ eventualmente apresentará valores fora deste intervalo, e assim não se pode garantir que esta função de ligação identidade gerará (em qualquer situação) valores estimados coerentes para p_i.

Como já mencionado antes (vide exemplo 1.3.10.), existem outras funções de ligação apropriadas para dados binomiais, como a log-log complementar e a probit.

Exemplo 1.3.13. (Modelo log-linear)
Suponha que $Y_i \sim$ Poisson(μ_i) onde i = 1,2,...n. Então os Y_i são contagens, e este modelo tem importância fundamental na análise de dados categorizados. O modelo log-linear tem a forma :

$$\eta_i = g(\mu_i) = \log(\mu_i) = x_i^t.\beta$$
$$\Rightarrow \mu_i = \exp(x_i^t.\beta)$$

Exemplo 1.3.14. (Modelo Gama)
Sejam $Y_i \sim$ Gama(r,β_i), (i=1,2,....,n) onde r é uma constante conhecida e a densidade está parametrizada como :

$$f(y) = \frac{1}{\Gamma(r).\beta^r}.y^{r-1}.\exp\left(-\frac{y}{\beta}\right)$$

Note que Y_i é uma variável aleatória não negativa. Neste caso $\mu_i = E(Y_i) = r.\beta$ e $VAR(Y_i) = r.\beta^2 = \mu^2/r$. Logo, a variância é uma função crescente da média. O coeficiente de variação de Y_i é :

$$CV = \frac{\sqrt{VAR(Y)}}{E(Y)} = \frac{\sqrt{(\mu^2/r)}}{\mu} = \frac{1}{\sqrt{r}}$$

Ou seja, é uma constante.

Um Modelo Linear Generalizado comum na prática para dados com distribuição Gama é o modelo log-linear dado por :

$$\eta_i = g(\mu_i) = \log(\mu_i) = x_i^t.\beta$$
$$\Rightarrow \mu_i = \exp(x_i^t.\beta)$$

Note que $\mu_i = E(Y_i)$ é sempre positivo, enquanto η_i, o preditor linear, pode ser qualquer número real. A função de ligação $\eta = g(\mu) = \log(\mu)$ deve parecer "natural", pois leva

números positivos (μ) em números reais (η). Um argumento análogo pode ser usado no caso de dados Poisson, onde a média μ é também um número positivo.