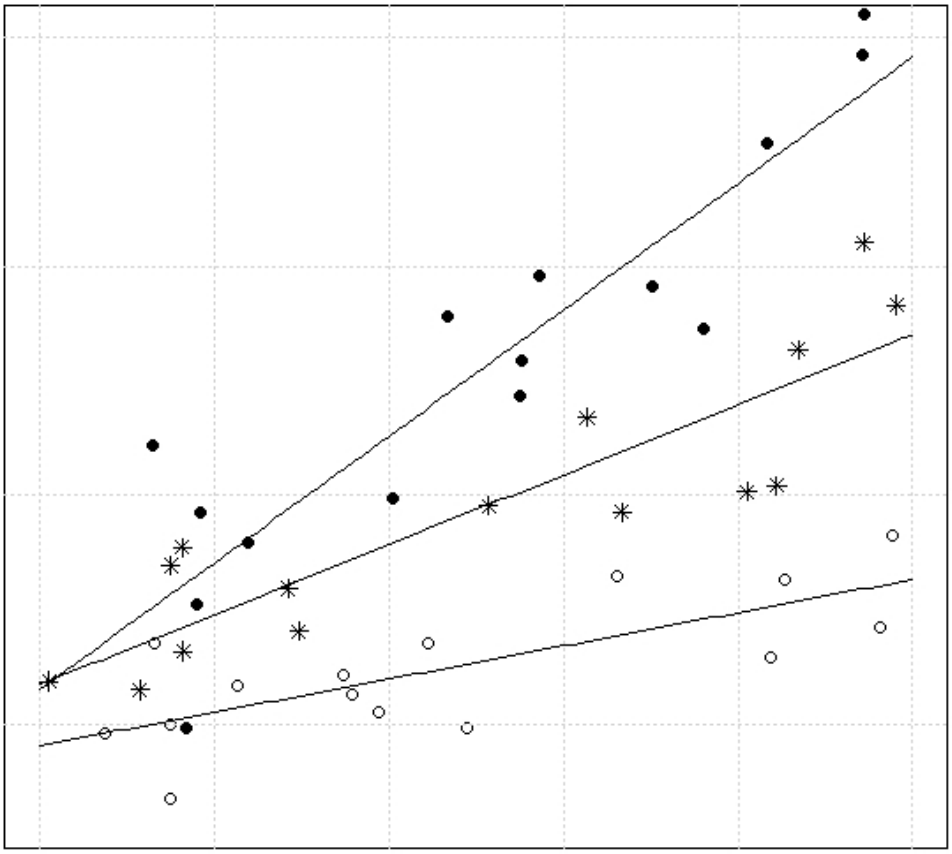
Notas de Aula
Modelos Lineares I - GET00138



Jessica Kubrusly

Departamento de Estatística
Instituto de Matemática e Estatística
Universidade Federal Fluminense

Julho de 2014

Sumário

Capítulo 1

Regressão Linear Simples

1.1 O Modelo de Regressão Linear Simples

Dizemos que existe uma **relação funcional** entre duas variáveis x e y se existe uma função f tal que $y = f(x)$. Nesse caso, para cada valor de x existe um único valor de y tal que $y = f(x)$.

Exemplo 1.1.1 *Considere y o valor total arrecadado com as vendas de um certo produto e x quantidade de produtos vendidos. Se o produto em questão custa R$ 0,50 temos $y = 0,5 \times x$. Nesse caso, se conhecermos o valor de x sabemos exatamente qual será o valor de y.*

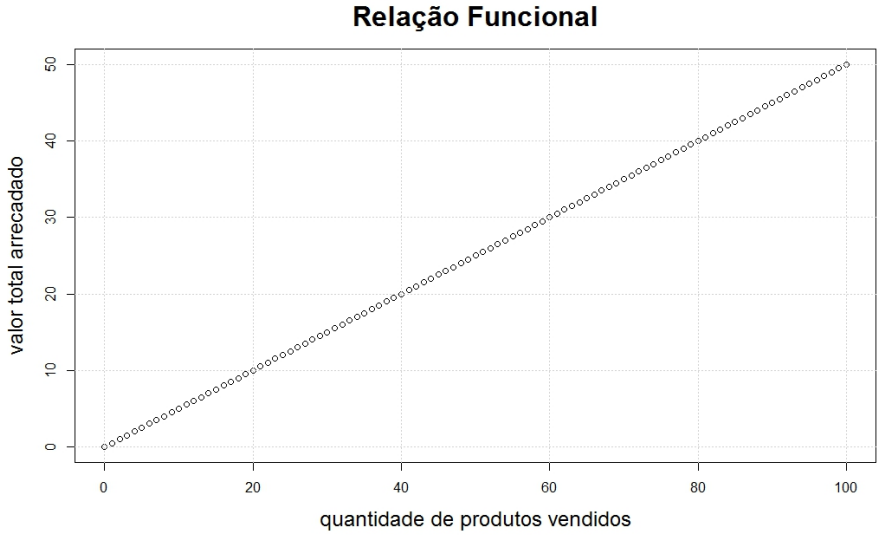


Figura 1.1: Exemplo de relação funcional

Dizemos que existe uma **relação estatística** entre duas variáveis x e y se para cada valor de x existe uma distribuição de probabilidade para y. Nesse caso, y é uma variável aleatória e para cada valor de x podem existir diferentes valores para y.

Exemplo 1.1.2 *A fim de analisar o comportamento dos consumidores 60 famílias foram entrevistadas e nessa entrevista foi perguntado sobre a renda familiar (x) e o total gasto com alimentação no último mês (y). O objetivo desse estudo é tentar entender a relação entre as variáveis x e y assim definidas. Para isso os pontos (x, y) coletado na entrevista forma plotado no plano xy, como mostra a Figura 1.2.*

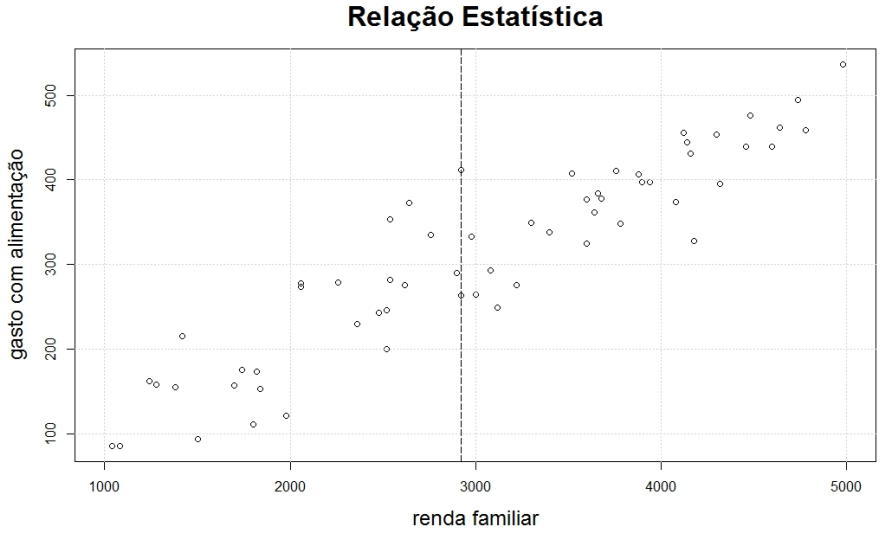


Figura 1.2: Exemplo de relação estatística - comportamento do consumidor

Veja que podemos destacar pelo menos duas famílias com a mesma renda, porém com diferentes gastos com alimentação no último mês. Isso já indica que a relação entre x e y não é funcional. Apensar de não ser possível estabelecer uma relação funcional entre x e y podemos observar que, em média, quanto maior é a renda da família maior é o gasto com alimentação.

Exemplo 1.1.3 *Um médico mediu os níveis de esteróide no plasma de mulheres saudáveis entre 8 e 25 anos. É sugestivo estudar a relação (estatística) entre a idade da paciente (x) e o nível de esteróide no plasma (y). Para isso os pontos (x, y) coletado pelo médico forma plotado no plano xy, como mostra a Figura 1.3.*

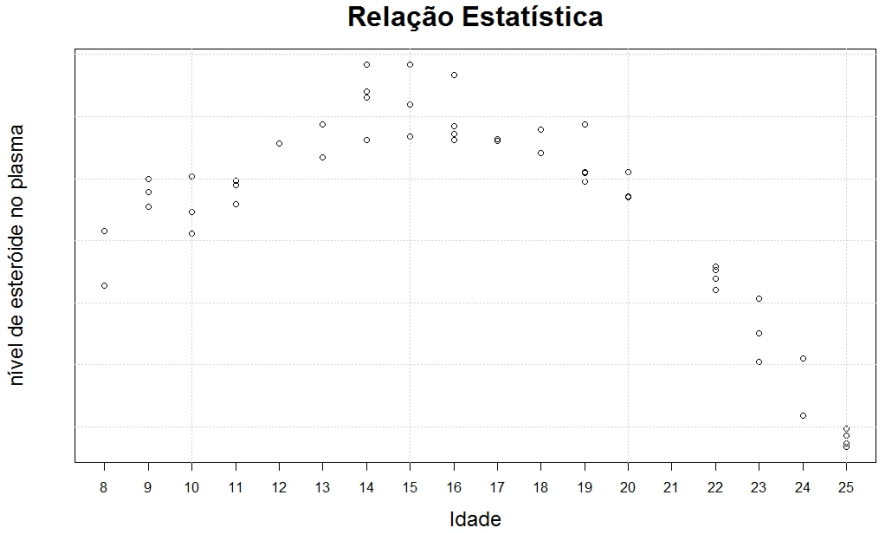


Figura 1.3: Exemplo de relação estatística - nível de esteróide no plasma

Veja que na amostra de pacientes existem mulheres com mesma idade e diferentes níveis de esteróide no plasma.

A partir desse estudo pode-se observar que o nível de esteroide cresce quando as pacientes aumentam de idade entre 8 e 15 anos. Depois dessa idade os níveis de esteroide começam a cair.

Um modelo de regressão é a formalização de uma relação estatística entre duas variáveis: x e y. Neste curso vamos sempre supor que x não é variável aleatória. Já a variável y será variável aleatória, como veremos em breve.

Em todo modelo de regressão são considerados os seguintes postulados:

- Existe uma distribuição de probabilidade para y para cada valor da variável x, uma vez que supomos existir uma relação estatística entre x e y.
- A média de y varia de forma sistemática em relação à x.

Voltando ao Exemplo 1.1.2, podemos pensar que para uma determinada renda familiar x existe uma distribuição de probabilidade para o gasto com alimentação y. Isto é, fixada a renda de uma família em x podemos supor que o gasto com alimentação dessa família, y, é uma variável aleatória. Além disso, a média dessa variável aleatória varia de acordo com o valor de x. Aparentemente quanto maior a renda de uma família maior será o gasto médio dessa família com alimentação, ou seja, quanto maior x maior será $\mathrm{E}[y|x]$.

Definição 1.1.4 *O* **Modelo de Regressão Linear Simples** *é o modelo que define uma relação estatística linear entre uma variável independente x, também chamada de variável preditiva, e uma variável dependente y, denominada variável resposta do modelo. A suposição básica desse modelo é que a média da distribuição de y varia de forma linear com x. Essa relação pode ser estabelecido por:*

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad E[\varepsilon_i] = 0,\ Var(\varepsilon_i) = \sigma^2 \ e\ \ Cov(\varepsilon_i, \varepsilon_j) = 0 \ \ \forall \ \ i \neq j \qquad (1.1)$$

onde, x_i é o valor para a i-ésima observação da variável independente x, y_i é o valor para a i-ésima observação da variável resposta y e ε_i é o erro aleatório para a i-ésima observação. Além disso, β_0, β_1 e σ^2 são os parâmetros do modelo.

Algumas observações:

1. Veja que ε_i é variável aleatória. Logo $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ também é variável aleatória.
2. $E[y_i|x_i] = E[\beta_0 + \beta_1 x_i + \varepsilon_i|x_i] = \beta_0 + \beta_1 x_i + \mathrm{E}[\varepsilon_i|x_i] = \beta_0 + \beta_1 x_i$, uma vez que $E[\varepsilon_i] = 0$. Ou seja, a média da variável aleatória y_i varia de forma linear com x_i.
3. $Var(y_i|x_i) = Var(\beta_0 + \beta_1 x_i + \varepsilon_i|x_i) = Var(\varepsilon_i|x_i) = \sigma^2$, pois $Var(\varepsilon_i) = \sigma^2$ independente de x_i.
4. O valor observado de y_i difere da reta de regressão $\beta_0 + \beta_1 x_i$ por uma quantidade igual ao termo aleatório ε_i.
5. Os erros aleatórios não são correlacionados, isto é, o erro de uma observação i não é influenciado pelos erros de outras observações.
6. Se $i \neq j$, $Cov(y_i, y_j) = Cov(\beta_0 + \beta_1 x_i + \varepsilon_i, \beta_0 + \beta_1 x_j + \varepsilon_j) = Cov(\varepsilon_i, \varepsilon_j) = 0$. Ou seja, y_i e y_j não são correlacionados sempre que $i \neq j$.

Exemplo 1.1.5 *Continuando o exemplo 1.1.2, veja que o gráfico de dispersão apresentado na Figura 1.2 sugere que em média y varia de forma linear com x. Dessa forma o Modelo de Regressão Linear Simples parece adequado para descrever a relação entre o gasto com alimentação das famílias (y) e a renda familiar (x).*

Suponha que para esses dados seja definido o seguinte modelo:

$$y_i = 10,3 + 0,1x_i + \varepsilon_i$$

onde x_i é a renda familiar da i-ésima família entrevistada e y_i é o total gasto com alimentação dessa família no último mês. Isso significa que uma família com renda de R$ 3.000,00 gasta em média $10,3 + 0,1 \times 3.000$ =R$ 310,30 com alimentação por mês. Veja Figura 1.4. Veja na Figura 1.4 que para cada valor de x os valores médios da variável

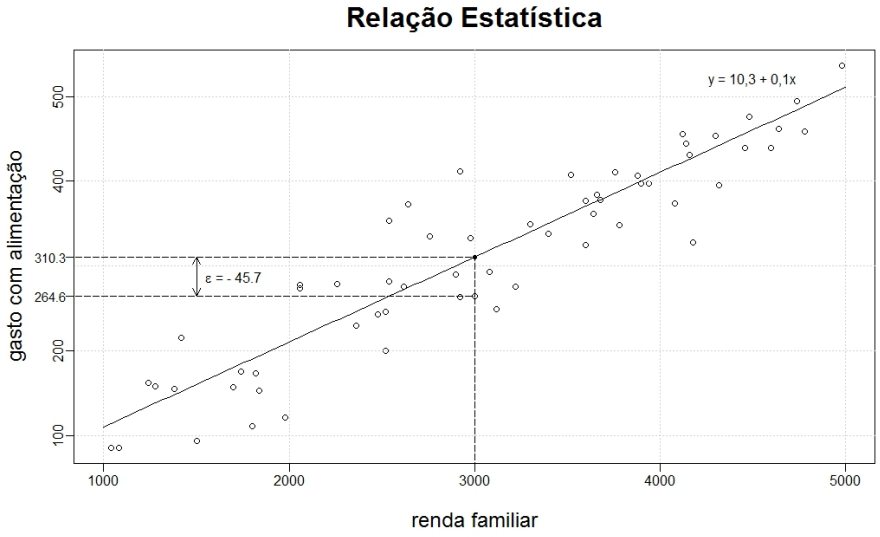


Figura 1.4: Modelo de Regressão Linear - comportamento do consumidor

y são aqueles definidos pela reta, chamada de reta de regressão. Veja também que uma das famílias entrevistadas tem renda média de R$ 3.000,00 e gastou no último mês R$ 264,6 com alimentação. Logo, para essa família, o erro cometido pelo modelo foi de ε = R$ 264,6 − R$ 310,3 = − R$ 45,7.

1.2 Estimadores para β_0 e β_1

Os parâmetros β_0 e β_1 do Modelo de Regressão Linear definem a reta de regressão: $\mathrm{E}[y|x] = \beta_0 + \beta_1 x$, que descreve a relação ente x e $E[y|x]$. O parâmetro β_0 é o coeficiente linear dessa reta e o parâmetro β_1 o coeficiente angular. É muito comum interpretar os dados a partir dos valores desses dois parâmetros. Nesse caso o parâmetro β_0 indica a média da variável resposta ao nível zero, isto é, o valor de $\mathrm{E}[y|x = 0]$. Já o parâmetro β_1 indica o acréscimo (ou decréscimo) na média da variável resposta y quando a variável preditiva x aumenta em uma unidade.

Nessa seção veremos como estimar β_0 e β_1 a partir de uma amostra de (x, y). O que será feito é buscar a reta que melhor se ajusta aos pontos (x, y) definidos pela amostra recolhida. Para isso suponha que o modelo de regressão linear seja adequado para representar a relação entre as variáveis x e y. Suponha também que conhecemos uma amostra de tamanho n destas variáveis: $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$.

1.2.1 Estimadores por Mínimos Quadrados

O estimador para (β_0, β_1) por mínimos quadrados é aquele que minimiza a soma dos quadrados dos erros ε_i, isto é, vamos buscar a reta que melhor se ajusta nos pontos da amostra.

A soma dos quadrados dos erros pode ser definida por:

$$Q = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

Veja que Q é função de β_0 e β_1. Queremos então encontrar os valores de β_0 e β_1 que minimiza Q, ou seja,

$$(\hat{\beta}_0, \hat{\beta}_1) = \arg\min Q(\beta_0, \beta_1) = \arg\min \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

Para resolver esse problema vamos procurar os pontos críticos de Q: (i) derivar Q em relação à β_0 e igualar a zero; (ii) derivar Q em relação à β_1 e igualar a zero; (iii) resolver o sistema linear formado pelas duas equações.

Como Q é uma função convexa, isto é, não precisamos usar o a condição de segunda ordem para verificar que os pontos críticos são realmente pontos de mínimo. Funções convexas só assumem ponto de mínimo, nunca de máximo ou sela. Então vamos às contas.

Primeiro derivamos Q em relação à β_0:

$$\frac{\partial Q}{\partial \beta_0} = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) = -2 \left(\sum_{i=1}^{n} y_i - n\beta_0 - \beta_1 \sum_{i=1}^{n} x_i \right)$$

Em seguida igualamos o resultado à zero:

$$\frac{\partial Q}{\partial \beta_0} = 0 \Rightarrow -2 \left(\sum_{i=1}^{n} y_i - n\beta_0 - \beta_1 \sum_{i=1}^{n} x_i \right) = 0 \Rightarrow \sum_{i=1}^{n} y_i = n\beta_0 + \beta_1 \sum_{i=1}^{n} x_i$$

E assim chegamos na primeira equação do nosso sistema.

$$\sum_{i=1}^{n} y_i = n\beta_0 + \beta_1 \sum_{i=1}^{n} x_i \tag{1.2}$$

Agora derivamos Q em relação à β_1:

$$\frac{\partial Q}{\partial \beta_1} = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)\, x_i = -2 \left(\sum_{i=1}^{n} y_i x_i - \beta_0 \sum_{i=1}^{n} x_i - \beta_1 \sum_{i=1}^{n} x_i^2 \right)$$

Em seguida igualamos o resultado à zero:

$$\frac{\partial Q}{\partial \beta_1} = 0 \Rightarrow -2 \left(\sum_{i=1}^{n} y_i x_i - \beta_0 \sum_{i=1}^{n} x_i - \beta_1 \sum_{i=1}^{n} x_i^2 \right) = 0 \Rightarrow \sum_{i=1}^{n} y_i x_i = \beta_0 \sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} x_i^2$$

E assim chegamos na segunda equação do nosso sistema.

$$\sum_{i=1}^{n} y_i x_i = \beta_0 \sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} x_i^2 \tag{1.3}$$

Logo, o sistema a ser resolvido é:

$$\left\{ \begin{array}{rcl} \sum_{i=1}^{n} y_i & = & n\beta_0 + \beta_1 \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} y_i x_i & = & \beta_0 \sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} x_i^2 \end{array} \right.$$

Que pode ser escrito em sua forma matricial:

$$\begin{pmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} y_i x_i \end{pmatrix}$$

Para solucionar o sistema basta inverter a matriz e fazer as contas.

$$\begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = \begin{pmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{pmatrix}^{-1} \begin{pmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} y_i x_i \end{pmatrix}$$

$$\begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = \frac{1}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \begin{pmatrix} \sum_{i=1}^{n} x_i^2 & -\sum_{i=1}^{n} x_i \\ -\sum_{i=1}^{n} x_i & n \end{pmatrix} \begin{pmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} y_i x_i \end{pmatrix}$$

Geralmente nos livros os estimadores são apresentados na seguinte forma,

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{1.4}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \tag{1.5}$$

que é a solução da operação matricial descrita acima seguida por algumas manipulações algébricas.

1.2.2 Estimadores por Máxima Verossimilhança

Para encontrar os estimadores por mínimos quadrados não precisamos definir a distribuição de ε e nem da variável resposta y. Mas se quisermos encontrar os estimadores por mínimos quadrados é preciso conhecê-las. Por isso antes de encontrar os estimadores por MV vamos definir o modelo normal.

Definição 1.2.1 *O* **Modelo de Regressão Linear Normal Simples** *é um Modelo de Regressão Linear Simples em que os erros ε_i são normalmente distribuídos.*

Como consequência temos:

- $\varepsilon_i \sim N(0, \sigma^2)$;
- ε_i e ε_j são variáveis aleatórias independentes para $i \neq j$;
- $y_i \sim Normal(\beta_0 + \beta_1 x_i, \sigma^2)$;
- y_i e y_j são variáveis aleatórias independentes para $i \neq j$.

Agora que já temos distribuição para y_i podemos encontrar os estimadores de máxima verossimilhança. O primeiro passo é encontrar a função de máxima verossimilhança em função dos parâmetros do modelo.

$$L(\beta_0, \beta_1, \sigma^2) = \prod_{i=1}^{n} f_{y_i}(y_i|\beta_0, \beta_1, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i-\beta_0-\beta_1 x_i)^2}{2\sigma^2}} = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n e^{-\sum_{i=1}^{n}\frac{(y_i-\beta_0-\beta_1 x_i)^2}{2\sigma^2}}$$

Veja que se estamos considerando as variáveis β_0 e β_1, fixando σ, maximizar L é o mesmo que minimizar $\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$, ou seja, recaímos no problema do estimador por mínimos quadrados. Logo, os estimadores definidos nas Equações 1.4 e 1.5 são também os estimadores por máxima verossimilhança.

1.3 Estimador para σ^2

Primeiro vamos encontrar o estimador de máxima verissimilhança para σ^2. Escrevendo a função de verossimilhança somente em função de σ^2, considerando que os estimadores de MV para β_0 e β_1 já foram encontrados, temos:

$$L(\sigma^2) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n e^{-\sum_{i=1}^{n}\frac{(y_i-\hat{\beta}_0-\hat{\beta}_1 x_i)^2}{2\sigma^2}}$$

A função de log-verossimilhança será definida por:

$$l(\sigma^2) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

Derivando l em relação à σ^2 e igualando a zero para encontrar o seu máximo temos:

$$\frac{\partial l}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 = -\frac{1}{2\sigma^2}\left(n - \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2\right)$$

$$\frac{\partial l}{\partial \sigma^2} = 0 \Rightarrow -\frac{1}{2\sigma^2}\left(n - \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2\right) = 0 \Rightarrow n - \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 = 0$$

Logo,

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2}{n} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}$$

onde $\hat{y}_i = \hat{\beta}_0 - \hat{\beta}_1 x_i$ é o estimador para a variável resposta ao nível x_i.

Em algumas aulas veremos que este estimador é tendencioso e que $E[\hat{\sigma}^2] = \frac{n-2}{n}\sigma^2$, por isso este não será o estimador usado. Usaremos a sua versão não-tendenciosa.

O estimador para o parâmetro σ^2 adotado por nós será

$$MSE = \frac{n}{n-2}\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}. \tag{1.6}$$

Verifique que este é um estimador não-tendencioso.

1.4 Inferências para β_1

Nessa seção vamos analisar o estimador $\hat{\beta}_1$ e a partir dele obter mais informações sobre o parâmetro β_1, como por exemplo, intervalos de confiança e testes de hipóteses.

1.4.1 Distribuição Amostral de $\hat{\beta}_1$

Veremos nessa seção a demonstração da Proposição 1.4.1 enunciada a seguir.

Proposição 1.4.1 *Considerando o Modelo Linear Normal seja*

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}.$$

Então,

(i) $\hat{\beta}_1$ *é combinação linear de* $\{y_i\}_{i=1}^n$

(ii) $\hat{\beta}_1$ *é variável aleatória normal*

(iii) $E[\hat{\beta}_1] = \beta_1$

(iv) $Var(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$

Demonstração:

(i) Queremos mostrar que $\hat{\beta}_1 = \sum_{i=1}^n k_i y_i$.

$$\begin{aligned}
\hat{\beta}_1 &= \frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) \\
&= \frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2} \left(\sum_{i=1}^n x_i y_i - \bar{y} \cancel{\sum_{i=1}^n x_i} - \bar{x} \sum_{i=1}^n y_i + \cancel{n\bar{x}\bar{y}} \right) \\
&= \frac{\sum_{i=1}^n x_i y_i - \bar{x} \sum_{i=1}^n y_i}{\sum_{i=1}^n (x_i - \bar{x})^2} \\
&= \frac{\sum_{i=1}^n (x_i - \bar{x})\, y_i}{\sum_{i=1}^n (x_i - \bar{x})^2} \\
&= \sum_{i=1}^n \frac{(x_i - \bar{x})\, y_i}{\sum_{i=1}^n (x_i - \bar{x})^2} \\
&= \sum_{i=1}^n k_i y_i \text{ , com } k_i = \frac{(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2}.
\end{aligned}$$

□

(ii) Como $y_i \sim$ Normal e combinação linear de normais também é v.a. Normal, então $\hat{\beta}_1 \sim$ Normal. □

(iii) Queremos calcular $E[\hat{\beta}_1]$.

$$E[\hat{\beta}_1] = E[\sum_{i=1}^{n} k_i y_i] = \sum_{i=1}^{n} E[k_i y_i] = \sum_{i=1}^{n} k_i E[y_i] = \sum_{i=1}^{n} k_i(\beta_0+\beta_1 x_i) = \beta_0 \sum_{i=1}^{n} k_i + \beta_1 \sum_{i=1}^{n} k_i x_i$$

Primeiro veja que

$$\sum_{i=1}^{n} k_i = \sum_{i=1}^{n} \frac{(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n} x_i - n\bar{x}}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = 0.$$

Agora veja que

$$\sum_{i=1}^{n} k_i x_i = \sum_{i=1}^{n} \frac{(x_i - \bar{x})\, x_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n}(x_i^2 - \bar{x}x_i)}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

Contas no numerador:

$$\sum_{i=1}^{n} \left(x_i^2 - \bar{x}x_i\right) = \sum_{i=1}^{n} x_i^2 - \bar{x} \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2.$$

Contas no denominador:

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n}(x_i^2 + \bar{x}^2 - 2x_i\bar{x}) = \sum_{i=1}^{n} x_i^2 + n\bar{x}^2 - 2\bar{x} \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2.$$

Então,

$$\sum_{i=1}^{n} k_i x_i = \frac{\sum_{i=1}^{n}(x_i^2 - \bar{x}x_i)}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = 1.$$

Assim concluímos que

$$E[\hat{\beta}_1] = \beta_0 \sum_{i=1}^{n} k_i + \beta_1 \sum_{i=1}^{n} k_i x_i = \beta_1.$$

□

(iv) Para terminar vamos calcular a variância de $\hat{\beta}_1$.

$$Var\left(\hat{\beta}_1\right) = Var\left(\sum_{i=1}^{n} k_i y_i\right) = \sum_{i=1}^{n} Var\left(k_i y_i\right) = \sum_{i=1}^{n} k_i^2 Var\left(y_i\right) = \sigma^2 \sum_{i=1}^{n} k_i^2.$$

Veja que $\sum_{i=1}^{n} k_i^2 = \sum_{i=1}^{n} \frac{(x_i-\bar{x})^2}{\left(\sum_{i=1}^{n}(x_i-\bar{x})^2\right)^2} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{\left(\sum_{i=1}^{n}(x_i-\bar{x})^2\right)^2} = \frac{\cancel{\sum_{i=1}^{n}(x_i-\bar{x})^2}}{\left(\sum_{i=1}^{n}(x_i-\bar{x})^2\right)\cancel{\sum_{i=1}^{n}(x_i-\bar{x})^2}} = \frac{1}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$.

□

1.4.2 Intervalo de Confiança para β_1

Agora que já conhecemos a distribuição amostral de $\hat{\beta}_1$ e sabemos que este é um estimador não tendencioso para β_1 podemos buscar uma quantidade pivotal para β_1 a fim de criar um intervalo de confiança deste parâmetro. Para simplificar a notação vamos usar $S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2$.

Como $\hat{\beta}_1 \sim N(\beta_1, \sigma^2/S_{xx}) \Rightarrow \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\sigma^2/S_{xx}}} \sim N(0,1)$. Mas ainda não temos uma quantidade pivotal, pois a transformação depende do parâmetro desconhecido σ^2.

Para construir a quantidade pivotal vamos precisar do resultado do Teorema 1.4.2 enunciado a seguir. Sua demonstração sera omitida.

Teorema 1.4.2 *Suponha o modelo de regressão linear. Então,*

(i) $(n-2)MSE/\sigma^2 \sim \chi^2_{n-2}$

(ii) MSE *é v.a. independente de* $\hat{\beta}_1$ *e* $\hat{\beta}_0$.

Baseado nesse teorema podemos mostrar a Proposição 1.4.3 a seguir.

Proposição 1.4.3

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{MSE/S_{xx}}} \sim t_{n-2}.$$

Demonstração:

Seja $Z = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\sigma^2/S_{xx}}} \sim N(0,1)$ e $X = (n-2)MSE/\sigma^2 \sim \chi^2_{n-2}$. Sabemos que $Z/\sqrt{X/(n-2)} \sim t_{n-2}$. Ou seja,

$$\frac{Z}{\sqrt{X/(n-2)}} = \frac{\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\sigma^2/S_{xx}}}}{\sqrt{(n-2)MSE/\sigma^2(n-2))}} = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{MSE/S_{xx}}} \sim t_{n-2}$$

□

Agora já temos uma quantidade pivotal e podemos construir um intervalo de confiança para β_1.

O intervalo de confiança para β_1 com confiabilidade de $1-\alpha$ é definido por:

$$\hat{\beta}_1 \pm t_{1-\frac{\alpha}{2},n-2}\sqrt{\frac{MSE}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}. \tag{1.7}$$

Deixo a dedução da Equação 1.7 também como exercício.

1.4.3 Testes de Hipótese para β_1

Uma questão relevante em modelos lineares é se a variável preditiva x realmente influencia a média da variável resposta de forma linear, indicando que o modelo linear é adequado para descrever a relação entre essas duas variáveis. Essa questão pode ser formulada a partir de um teste de hipótese:

$$H_0 : \beta_1 = 0 \qquad H_1 : \beta_1 \neq 0.$$

O que esse teste verifica é se os valores das observações de y crescem (ou decrescem) conforme x aumenta de valor. PASSAR ALGUNS EXEMPLOS DE GRAFICOS EM QUE O TESTE SERIA ACEITO E REJEITADO.

Dessa forma, se o teste acima não for rejeitado temos fortes indícios para acreditar que o modelo linear não é adequado para descrever a relação entre x e y.

Para testar esse teste podemos usar o intervalo de confiança definido acima e rejeitar H_0 se o intervalo não contiver o zero ou aceitar caso ele contenha. Mas vamos formalizar isso a partir de uma estatística de teste.

Seja

$$T = \frac{\hat{\beta}_1}{\sqrt{MSE/S_{xx}}}. \tag{1.8}$$

Veja que sob H_0 $T \sim t_{n-2}$. Então a regra de decisão do nosso teste será:

- H_0 será rejeitada se $|T| \geq t_{1-\frac{\alpha}{2},n-2}$
- H_0 não será rejeitada se $|T| < t_{1-\frac{\alpha}{2},n-2}$

Em geral vamos usar o pvalor deste teste para tomar a decisão. Então H_0 será rejeitada se o p-valor do teste for pequeno e aceita caso contrário.

E o que significa aceitar H_0? Significa concluir que o modelo linear não é adequado.

O teste que acabamos de descrever é o mais comum entre os testes para β_1, mas nada impede de queremos testar outras hipóteses. Vejamos mais algumas agora.

Suponha que queremos teste as seguintes hipótese

$$H_0 : \beta_1 \leq 0 \qquad H_1 : \beta_1 > 0.$$

Nesse caso vamos usar a mesma estatística de teste T definida na equação 1.8, a única diferença é que o teste em vez de ser bilateral será unilateral. A regra de decisão será:

- H_0 será rejeitada se $T \geq t_{1-\alpha,n-2}$
- H_0 não será rejeitada se $T < t_{1-\alpha,n-2}$

E se as hipóteses a serem testadas forem

$$H_0 : \beta_1 = \beta_1^* \qquad H_1 : \beta_1 \neq \beta_1^*,$$

teremos que criar uma nova estatística de teste. Nesse caso a estatística de teste terá que mudar para aquela aquela apresentada na Equação 1.9.

$$T = \frac{\hat{\beta}_1 - \beta_1^*}{\sqrt{MSE/S_{xx}}}. \tag{1.9}$$

Como sob H_0 $T \sim t_{n-2}$ e este teste também é bilateral, as regras de decisão são as mesmas para o caso de igual ou diferente de zero.

Para terminar, se queremos testar

$$H_0 : \beta_1 \leq \beta_1^* \qquad H_1 : \beta_1 > \beta_1^*,$$

a estatística usada será a apresentada na equação 1.9 e a regra de decisão são as mesmas para o caso de menor ou maior que zero.

1.5 Inferências para β_0

Agora vamos fazer o mesmo para o parâmetro β_0.

1.5.1 Distribuição Amostral de $\hat{\beta}_0$

Já definimos $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$ e já vimos que $\hat{\beta}_1 \sim N(\beta_1, \sigma^2/S_{xx})$. Com esses resultados já podemos provar a seguinte proposição:

Proposição 1.5.1 *Seja $\hat{\beta}_0$ o estimador por mínimos quadrados para β_0 em um modelo de regressão linear normal simples. Então é verdade que:*

(i) $\hat{\beta}_0$ *é combinação linear de* $\{y_i\}_{i=1}^n$

(ii) $\hat{\beta}_0$ *é variável aleatória normal*

(iii) $E[\hat{\beta}_0] = \beta_0$

(iv) $Var(\hat{\beta}_0) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right)$

Demonstração:

(i) Já sabemos que $\hat{\beta}_1 = \sum_{i=1}^n k_i y_i$. Então podemos escrever:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} = \sum_{i=1}^n \frac{y_i}{n} - \sum_{i=1}^n \bar{x}k_i y_i = \sum_{i=1}^n \frac{y_i}{n} - \bar{x}k_i y_i = \sum_{i=1}^n \left(\frac{1}{n} - \bar{x}k_i\right) y_i.$$

Logo, $\hat{\beta}_0$ é combinação linear de $\{y_i\}_{i=1}^n$.

□

(ii) Como $y_i \sim$ Normal e combinação linear de normais também é v.a. Normal, então $\hat{\beta}_0 \sim$ Normal. □

(iii) Queremos calcular $E[\hat{\beta}_0]$.

$$E[\hat{\beta}_0] = E\left[\bar{y} - \hat{\beta}_1\bar{x}\right] = E\left[\sum_{i=1}^n \frac{y_i}{n}\right] - \bar{x}E\left[\hat{\beta}_1\right] = \sum_{i=1}^n \frac{1}{n} E\left[y_i\right] - \bar{x}\beta_1 = \sum_{i=1}^n \frac{1}{n}\left(\beta_0 + \beta_1 x_i\right) - \bar{x}\beta_1 = \beta_0$$

□

(iv)

$$\begin{aligned} Var(\hat{\beta}_0) &= Var\left(\sum_{i=1}^{n}\left(\frac{1}{n}-\bar{x}k_i\right)y_i\right) = \sum_{i=1}^{n} Var\left(\left(\frac{1}{n}-\bar{x}k_i\right)y_i\right) = \sum_{i=1}^{n}\left(\frac{1}{n}-\bar{x}k_i\right)^2 Var\left(y_i\right) \\ &= \sum_{i=1}^{n}\left(\frac{1}{n}-\bar{x}k_i\right)^2\sigma^2 = \sigma^2\sum_{i=1}^{n}\left(\frac{1}{n^2}+\bar{x}^2k_i^2-2\frac{1}{n}\bar{x}k_i\right) \\ &= \sigma^2\left(\frac{n}{n^2}+\bar{x}^2\sum_{i=1}^{n}k_i^2-2\frac{1}{n}\bar{x}\sum_{i=1}^{n}k_i\right) = \sigma^2\left(\frac{1}{n}+\frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right) \end{aligned}$$

□

1.5.2 Intervalo de Confiança para β_0

Assim como foi feito para β_1, vamos construir um intervalo de confiança para β_0. Para isso vamos precisar de uma quantidade pivotal para o parâmetro β_0, que é apresentado na Proposição 1.5.2 a seguir.

Proposição 1.5.2

$$\frac{\hat{\beta}_0-\beta_0}{\sqrt{MSE\left(\frac{1}{n}+\frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right)}} \sim t_{n-2}.$$

A demonstração da Proposição 1.5.2 será deixada como exercício. Veja que ela é análoga a demonstração da Proposição 1.4.3.

O intervalo de confiança para β_0 com confiabilidade de $1-\alpha$ é definido por:

$$\hat{\beta}_0 \pm t_{1-\frac{\alpha}{2},n-2}\sqrt{MSE\left(\frac{1}{n}+\frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right)}. \tag{1.10}$$

Deixo a dedução da Equação 1.10 também como exercício.

1.5.3 Teste de Hipótese para β_0

Outra questão relevante em modelos lineares é se a média da variável resposta ao nível zero é ou não igual a zero. Por exemplo, suponha que x_i seja a o número de funcionários no dia i em uma fábrica e y_i o gasto com energia da fábrica no dia i. Nesse caso é razoável itemar se $E[y_i] = 0$ quando $x_i = 0$, isto é, nenhum funcionário trabalho.

Mas quem é a média da variável resposta ao nível zero? Como $E[y_i] = \beta_0 + \beta_1 x_i$, a média da variável resposta ao nível zero é β_0. Nesse caso o nosso itemamento é se $\beta_0 = 0$. Então podemos transformar essa pergunta em um teste de hipótese a fim de nos auxiliar na tomada de decisão. As hipóteses a serem testadas serão:

$$H_0 : \beta_0 = 0 \qquad H_1 : \beta_0 \neq 0.$$

Para testá-las vamos usar a estatística de teste T:

$$T = \frac{\hat{\beta}_0}{\sqrt{MSE\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}}. \tag{1.11}$$

Veja que sob H_0 $T \sim t_{n-2}$. Então a regra de decisão do nosso teste será:

- H_0 será rejeitada se $|T| \geq t_{1-\frac{\alpha}{2},n-2}$
- H_0 não será rejeitada se $|T| < t_{1-\frac{\alpha}{2},n-2}$

Em geral vamos usar o p-valor deste teste para tomar a decisão. Então H_0 será rejeitada se o p-valor do teste for pequeno e aceita caso contrário.

E o que significa aceitar H_0? Significa aceitar que a reta de regressão passa pela origem, ou seja, que $\beta_0 = 0$, ou seja, que a média da variável resposta ao nível zero é zero. Para cada problema isso terá uma interpretação que sempre deverá ser explicitada.

O teste que acabamos de descrever é o mais comum entre os testes para β_0, mas nada impede de queremos testar outras hipóteses. Vejamos mais algumas agora.

Suponha que queremos teste as seguintes hipótese

$$H_0 : \beta_0 \leq 0 \qquad H_1 : \beta_0 > 0.$$

Nesse caso vamos usar a mesma estatística de teste T definida na equação 1.11, a única diferença é que o teste em vez de ser bilateral será unilateral. A regra de decisão será:

- H_0 será rejeitada se $T \geq t_{1-\alpha,n-2}$
- H_0 não será rejeitada se $T < t_{1-\alpha,n-2}$

E se as hipóteses a serem testadas forem

$$H_0 : \beta_0 = \beta_0^* \qquad H_1 : \beta_0 \neq \beta_0^*,$$

teremos que criar uma nova estatística de teste. Nesse caso a estatística de teste terá que mudar para aquela aquela apresentada na Equação 1.12.

$$T = \frac{\hat{\beta}_0 - \beta_0^*}{\sqrt{MSE\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}}. \tag{1.12}$$

Como sob H_0 $T \sim t_{n-2}$ e este teste também é bilateral, as regras de decisão são as mesmas para o caso de igual ou diferente de zero.

Para terminar, se queremos testar

$$H_0 : \beta_0 \leq \beta_0^* \qquad H_1 : \beta_0 > \beta_0^*,$$

a estatística usada será a apresentada na equação 1.12 e a regra de decisão são as mesmas para o caso de menor ou maior que zero.

1.6 Teorema de Gauss-Markov

O Teorema de Gauss-Markov nos fornece propriedades importantes para os estimadores de β_0 e β_1.

Teorema 1.6.1 - Teorema de Gauss-Markov
Os estimadores de mínimos quadrados (ou máxima verossimilhança) para β_0 e β_1 do modelo de regressão linear são não-tendenciosos e têm variância mínima dentre todos os estimadores lineares não-tendenciosos.

Antes de demostrar o teorema acima vamos entender o que ele quer dizer.

- Primeiro ele diz que os estimadores são não-tendenciosos, isso quer dizer que $E[\hat{\beta}_0] = \beta_0$ e $E[\hat{\beta}_1] = \beta_1$.
- A segunda afirmação diz que tais estimadores têm variância mínima entre todos os estimadores lineares não-tendenciosos, isso quer dizer que: (i) trata-se de um estimador linear, isto é, $\hat{\beta}_1 = \sum_{i=1}^{n} k_i y_i$; (ii) Entre todos os estimadores desse tipo $\hat{\beta}_1$ é o com menor variância.

Veja que o primeiro ponto do Teorema já foi demonstrado quando encontramos as distribuições amostrais de $\hat{\beta}_1$ e $\hat{\beta}_0$. Então apenas falta demostrar o segundo ponto.

Demonstração: (só para β_1)

Já vimos que $E[\hat{\beta}_1] = \beta_1$, então só falta verificar que $Var(\hat{\beta}_1) \leq Var(\hat{b}_1)$ qualquer que seja $\hat{b}_1$ que satisfaça:

- $E[\hat{b}_1] = \beta_1$
- $\hat{b}_1 = \sum_{i=1}^{n} c_i y_i$

Nesse caso podemos afirmar que

$$E\left[\hat{b}_1\right] = E\left[\sum_{i=1}^{n} c_i y_i\right] = \sum_{i=1}^{n} c_i E\left[y_i\right] = \sum_{i=1}^{n} c_i \left(\beta_0 + \beta_1 x_i\right) = \beta_0 \sum_{i=1}^{n} c_i + \beta_1 \sum_{i=1}^{n} c_i x_i$$

Como $E[\hat{b}_1] = \beta_1$ podemos escrever $\beta_0 \sum_{i=1}^{n} c_i + \beta_1 \sum_{i=1}^{n} c_i x_i = \beta_1 \; \forall \; x_i \; \Rightarrow \sum_{i=1}^{n} c_i = 0$ e $\sum_{i=1}^{n} c_i x_i = 1$.

Sem perda de generalidade, vamos supor $c_i = k_i + d_i$, onde $k_i = \frac{(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$. Nesse caso podemos afirmar que

$$\begin{aligned} Var(\hat{b}_i) &= Var(\sum_{i=1}^{n}(k_i + d_i)y_i) = \sum_{i=1}^{n}(k_i + d_i)^2 Var(y_i) = \sigma^2 \sum_{i=1}^{n}(k_i + d_i)^2 \\ &= \sigma^2 \sum_{i=1}^{n}(k_i^2 + d_i^2 + 2k_i d_i) = \sigma^2 \left(\sum_{i=1}^{n} k_i^2 + \sum_{i=1}^{n} d_i^2 + 2\sum_{i=1}^{n} k_i d_i\right) \end{aligned}$$

Veja que $\sum_{i=1}^{n} k_i d_i = 0$:

$$\begin{aligned}
\sum_{i=1}^{n} k_i d_i &= \sum_{i=1}^{n} k_i(c_i - k_i) = \sum_{i=1}^{n} k_i c_i - \sum_{i=1}^{n} k_i^2 \\
&= \sum_{i=1}^{n} c_i \frac{(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} - \frac{1}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n} c_i (x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} - \frac{1}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\
&= \frac{\sum_{i=1}^{n} c_i x_i - \bar{x}\sum_{i=1}^{n} c_i - 1}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{1 - 0 - 1}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = 0
\end{aligned}$$

Então,

$$Var(\hat{b}_i) = \sigma^2 \left(\sum_{i=1}^{n} k_i^2 + \sum_{i=1}^{n} d_i^2 \right) = Var(\hat{\beta}_1) + \sigma 2 \sum_{i=1}^{n} d_i^2 > Var(\hat{\beta}_1)$$

uma vez que $\sigma 2 \sum_{i=1}^{n} d_i^2 > 0$.

□

1.7 Inferências para a Variável Resposta

1.7.1 Distribuição Amostral de $\hat{y}_i$

Proposição 1.7.1 *Seja $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$. Então é verdade que:*

(i) $\hat{y}_i$ *é combinação linear de* $\{y_i\}_{i=1}^{n}$

(ii) $\hat{y}_i$ *é variável aleatória normal*

(iii) $E[\hat{y}_i] = \beta_0 + \beta_1 x_i$

(iv) $Var(\hat{y}_i) = \sigma^2 \left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{k=1}^{n}(x_k - \bar{x})^2} \right)$

Demonstração:

(i) Já vimos que $\hat{\beta}_1 = \sum_{j=1}^{n} k_j y_j$ e $\hat{\beta}_0 = \sum_{j=1}^{n} \left(\frac{1}{n} - \bar{x} k_j \right) y_j$ com $k_j = \frac{(x_j - \bar{x})}{\sum_{k=1}^{n}(x_k - \bar{x})^2}$. Então podemos escrever:

$$\begin{aligned}
\hat{y}_i &= \hat{\beta}_0 + \hat{\beta}_1 x_i = \left(\sum_{j=1}^{n} \left(\frac{1}{n} - \bar{x} k_j \right) y_j \right) + \left(\sum_{j=1}^{n} k_j y_j \right) x_i \\
&= \left(\sum_{j=1}^{n} \left(\frac{1}{n} - \bar{x} k_j \right) y_j \right) + \left(\sum_{j=1}^{n} x_i k_j y_j \right) = \sum_{j=1}^{n} \left(\left(\frac{1}{n} - \bar{x} k_j \right) y_j + x_i k_j y_j \right) \\
&= \sum_{j=1}^{n} \left(\frac{1}{n} - \bar{x} k_j + x_i k_j \right) y_j = \sum_{j=1}^{n} \left(\frac{1}{n} + (x_i - \bar{x}) k_j \right) y_j = \sum_{j=1}^{n} a_{ij} y_j
\end{aligned}$$

□

(ii) Como $y_i \sim$ Normal e combinação linear de normais também é v.a. Normal, então $\hat{y}_i \sim$ Normal.

□

(iii) $E[\hat{y}_i] = E[\hat{\beta}_0 + \hat{\beta}_1 x_i] = \beta_0 + \beta_1 x_i$.

□

(iv) Usando as propriedades da variância temos $Var\left(\hat{y}_i\right) = Var\left(\sum_{j=1}^n a_{ij} y_j\right) = \sum_{j=1}^n a_{ij}^2 Var(y_j) = \sigma^2 \sum_{j=1}^n a_{ij}^2$. Onde,

$$\begin{aligned}\sum_{j=1}^n a_{ij}^2 &= \sum_{j=1}^n \left(\frac{1}{n} + (x_i - \bar{x})\, k_j\right)^2 = \sum_{j=1}^n \left(\frac{1}{n^2} + (x_i - \bar{x})^2 k_j^2 + 2\frac{1}{n}(x_i - \bar{x})\, k_j\right) \\ &= \frac{n}{n^2} + (x_i - \bar{x})^2 \sum_{j=1}^n k_j^2 + 2\frac{1}{n}(x_i - \bar{x}) \sum_{j=1}^n k_j = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{k=1}^n (x_k - \bar{x})^2}\end{aligned}$$

uma vez que $\sum_{j=1}^n k_j = 0$ e $\sum_{j=1}^n k_j^2 = \frac{1}{\sum_{k=1}^n (x_k - \bar{x})^2}$.

□

1.7.2 Intervalo de Confiança para $E[y_i]$

Queremos encontrar o IC para a média da variável resposta no nível x_i, ou seja, queremos um IC para o parâmetro $\mu_i = E[y_i] = \beta_0 + \beta_1 x_i$.

Para construir uma quantidade pivotal para μ_i podemos usar a distribuição amostral de $\hat{y}_i$.

Proposição 1.7.2

$$\frac{\hat{y}_i - \mu_i}{\sqrt{MSE\left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{k=1}^n (x_k - \bar{x})^2}\right)}} \sim t_{n-2}.$$

A demonstração da Proposição 1.7.2 será deixada como exercício. Veja que ela é análoga a demonstração da Proposição 1.4.3.

Então o intervalo de confiança para $\mu_i = E[y_i]$ com confiabilidade de $1 - \alpha$ é definido por:

$$\hat{y}_i \pm t_{1-\frac{\alpha}{2}, n-2} \sqrt{MSE\left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{k=1}^n (x_k - \bar{x})^2}\right)}. \tag{1.13}$$

Deixo a dedução da Equação 1.13 também como exercício.

Vejamos agora algumas observações sobre o resultado encontrado:

- Para cada nível x_i temos um intervalo diferente, isso porque a variável resposta tem uma distribuição para cada valor da variável preditiva x_i.
- Quando temos $x_i = 0$ esse intervalo de confiança coincide com o intervalo de confiança para β_0. Esse era um resultado esperado? Por que?
- Quanto mais perto de $\bar{x}$ está x_i menor a amplitude do intervalo, ou seja, menor a incerteza sobre o parâmetro $\mu_i = E[y_i]$.

1.7.3 Intervalo de Predição para uma Nova Observação

Agora o cenário é o seguinte, depois de recolhida a amostra de tamanho n e ajustado o modelo de regressão linear normal simples queremos prever qual será o valor da variável resposta y se a variável preditiva assumir um valor específico x. Para isso vamos usar uma estimativa intervalar e encontrar um intervalo de confiança para o valor da variável resposta de uma nova observação.

Veja que isso é diferente do que foi feito na Seção 1.7.2. Na Seção 1.7.2 encontramos um intervalo de confiança para a média da variável resposta para um dado nível x_i, isto é, para $\mu_i = E[y_i] = \beta_0 + \beta_1 x_i$. Agora queremos um intervalo de confiança para o valor da variável resposta para um dado nível x_{novo} fora da amostra, ainda não observado. Isto é, queremos um intervalo de confiança para y_{novo}.

Já vimos que $y_{novo} \sim N(\beta_0+\beta_1 x_{novo}, \sigma^2)$ e que $\hat{y}_{novo} \sim N(\beta_0+\beta_1 x_{novo}, \sigma^2 \left(\frac{1}{n} + \frac{(x_{novo}-\bar{x})^2}{\sum_{k=1}^n (x_k-\bar{x})^2}\right))$ para qualquer nível x_{novo}. Observe que y_{novo} e $\hat{y}_{novo}$ são variáveis aleatórias independentes, uma vez que $\hat{y}_{novo}$ é combinação linear de $\{y_i\}$ dentro da amostra e y_{novo} é uma nova observação e por isso não faz parte da amostra. Então podemos afirmar que,

$$y_{novo} - \hat{y}_{novo} \sim N\left(0, \sigma^2\left(1 + \frac{1}{n} + \frac{(x_{novo}-\bar{x})^2}{\sum_{k=1}^n (x_k-\bar{x})^2}\right)\right).$$

Desse resultado segue que

$$\frac{y_{novo} - \hat{y}_i}{\sqrt{\sigma^2\left(1 + \frac{1}{n} + \frac{(x_i-\bar{x})^2}{\sum_{k=1}^n (x_k-\bar{x})^2}\right)}} \sim N(0,1).$$

Usando o resultado do Teorema 1.4.2 é possível demostrar a Proposição 1.7.3 a seguir. Faça a demostração desta proposição como exercício.

Proposição 1.7.3

$$\frac{y_{novo} - \hat{y}_{novo}}{\sqrt{MSE\left(1 + \frac{1}{n} + \frac{(x_{novo}-\bar{x})^2}{\sum_{k=1}^n (x_k-\bar{x})^2}\right)}} \sim t_{n-2}$$

A partir do resultado da Proposição 1.7.3 podemos construir um intervalo e confiança para y_{novo}. Esse intervalo será:

$$\hat{y}_{novo} \pm t_{1-\frac{\alpha}{2},n-2}\sqrt{MSE\left(1 + \frac{1}{n} + \frac{(x_i-\bar{x})^2}{\sum_{k=1}^n (x_k-\bar{x})^2}\right)} \tag{1.14}$$

Alguns comentários sobre esse resultado:

- Assim como o IC para μ_i o centro dele é $\hat{y}$.
- A variância desse IC é bem maior que a variância do IC para μ_i, você acha isso razoável?
- Assim como para o IC para μ_i, esse intervalo tem menor amplitude quando x_{novo} está mais perto de $\bar{x}$.
- E se x_{novo} está muito longe de $\bar{x}$ isso significa que a variância é muito grande. Na prática que isso significa?

1.8 $E[\hat{\sigma}^2]$ e $E[MSE]$

Na Seção 1.3 encontramos o estimador de máxima verossimilhança para σ^2, $\hat{\sigma}^2 = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n}$. Mas ficou faltando mostra que este estimador é tendencioso e por isso vamos usar o MSE, que é não-tendencioso. Esse é o resultado da Proposição 1.8.1 a seguir.

Proposição 1.8.1 *Seja* $\hat{\sigma}^2 = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n}$ *o estimado de máxima verossimilhança para* σ^2 *e seja* $MSE = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n-2}$. *Então,*

(i) $E\left[\hat{\sigma}^2\right] = \frac{n-2}{n}\sigma^2$

(ii) $E\left[MSE\right] = \sigma^2$

Demonstração:

(i)

$$E\left[\hat{\sigma}^2\right] = E\left[\frac{\sum_{i=1}^n \left(y_i - \hat{y}_i\right)^2}{n}\right] = \frac{1}{n}\sum_{i=1}^n E\left[\left(y_i - \hat{y}_i\right)^2\right] = \frac{1}{n}\sum_{i=1}^n Var\left(y_i - \hat{y}_i\right),$$

pois $E\left[y_i - \hat{y}_i\right] = 0$. Continuando,

$$E\left[\hat{\sigma}^2\right] = \frac{1}{n}\sum_{i=1}^n Var\left(y_i - \hat{y}_i\right) = \frac{1}{n}\sum_{i=1}^n \left(Var(y_i) + Var(\hat{y}_i) - 2Cov(y_i, \hat{y}_i)\right)$$

Já conhecemos $Var(y_i) = \sigma^2$ e $Var(\hat{y}_i) = \sigma^2\left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{k=1}^n (x_k - \bar{x})^2}\right)$, falta apenas encontrar $Cov(y_i, \hat{y}_i)$.

$$Cov(y_i, \hat{y}_i) = Cov(y_i, \sum_{j=1}^n a_{ij} y_j) = \sum_{j=1}^n a_{ij} Cov(y_i, y_j) = a_{ii}\sigma^2,$$

pois se $i \neq j$, y_i e y_j são v.a. independentes, logo $Cov(y_i, y_j) = 0$ para $i \neq j$. E se $i = j$, $Cov(y_i, y_j) = Cov(y_i, y_i) = Var(y_i) = \sigma^2$. Logo,

$$Cov(y_i, \hat{y}_i) = \left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{k=1}^n \left(x_k - \bar{x}\right)^2}\right)\sigma^2 \tag{1.15}$$

Então, continuando,

$$
\begin{aligned}
E\left[\hat{\sigma}^2\right] &= \frac{1}{n}\left(\sum_{i=1}^{n} Var(y_i) + \sum_{i=1}^{n} Var(\hat{y}_i) - 2\sum_{i=1}^{n} Cov(y_i, \hat{y}_i)\right) \\
&= \frac{1}{n}\left(n\sigma^2 + \sum_{i=1}^{n}\sigma^2\left(\frac{1}{n} + \frac{(x_i-\bar{x})^2}{\sum_{k=1}^{n}(x_k-\bar{x})^2}\right) - 2\sum_{i=1}^{n}\left(\frac{1}{n} + \frac{(x_i-\bar{x})^2}{\sum_{k=1}^{n}(x_k-\bar{x})^2}\right)\sigma^2\right) \\
&= \frac{\sigma^2}{n}\left(n+1+\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{\sum_{k=1}^{n}(x_k-\bar{x})^2} - 2\sum_{i=1}^{n}\left(\frac{1}{n} + \frac{(x_i-\bar{x})^2}{\sum_{k=1}^{n}(x_k-\bar{x})^2}\right)\right) \\
&= \frac{\sigma^2}{n}\left(n+1+1-2\frac{n}{n} - 2\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{\sum_{k=1}^{n}(x_k-\bar{x})^2}\right) \\
&= \frac{\sigma^2}{n}(n+1+1-2-2) \\
&= \frac{\sigma^2}{n}(n-2)
\end{aligned}
$$

□

(i) Veja que $MSE = \frac{n}{n-2}\hat{\sigma}^2$. Então,

$$E\left[MSE\right] = E\left[\frac{n}{n-2}\hat{\sigma}^2\right] = \frac{n}{n-2}E\left[\hat{\sigma}^2\right] = \frac{n}{n-2}\frac{n-2}{n}\sigma^2 = \sigma^2.$$

□

1.9 Regressão pela Origem

Nessa seção vamos definir o modelo de regressão linear onde temos $\beta_0 = 0$, ou seja, β_0 não é mais um parâmetro desconhecido do modelo. Vamos também encontrar as estimativas para os demais parâmetros nesse caso.

Definição 1.9.1 *O* **Modelo de Regressão Linear Simples pela Origem** *é um Modelo de Regressão Linear Simples entre x e y tal que* $\mathrm{E}[y \mid x = 0] = 0$. *Essa relação pode ser estabelecido por:*

$$y_i = \beta_1 x_i + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2) \text{ independentes.} \tag{1.16}$$

onde, x_i = o valor da variável independente na i-ésima observação, y_i = o valor da variável resposta na i-ésima observação e ε_i o erro aleatório para a i-ésima observação. β_1 e σ^2 são parâmetros do modelo

1.9.1 Estimador para β_1

Veja que nesse caso a soma dos erros ao quadrado passa a ser definida por:

$$Q(\beta_1) = \sum_{i=1}^{n}\varepsilon_i^2 = \sum_{i=1}^{n}(y_i - \beta_1 x_i)^2$$

Então queremos o arg min da função Q de uma variável só. Para isso basta derivar e igualar a zero.

$$\frac{dQ}{d\beta_1} = -2\sum_{i=1}^{n}(y_i - \beta_1 x_i)\, x_i = -2\left(\sum_{i=1}^{n} y_i x_i - \beta_1 \sum_{i=1}^{n} x_i^2\right)$$

Em seguida igualamos o resultado à zero:

$$\frac{dQ}{d\beta_1} = 0 \Rightarrow -2\left(\sum_{i=1}^{n} y_i x_i - \beta_1 \sum_{i=1}^{n} x_i^2\right) = 0 \Rightarrow \hat{\beta}_1 = \frac{\sum_{i=1}^{n} y_i x_i}{\sum_{i=1}^{n} x_i^2}$$

Então o estimado para β_1 por mínimos quadrados é definido pela Equação 1.17 abaixo.

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} y_i x_i}{\sum_{i=1}^{n} x_i^2} \tag{1.17}$$

Esse também é o estimador por máxima verissimilhança. Verifique.

1.9.2 Estimador para σ^2

O estimador para σ^2 nesse caso será:

$$MSE = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-1}. \tag{1.18}$$

Não vamos mostrar esse resultado, mas equivalente ao Teorema 1.4.2 para o caso da Regressão pela origem temos o seguinte teorema:

Teorema 1.9.2 *Suponha o modelo de regressão linear pela origem. Então,*

(i) $(n-1)MSE/\sigma^2 \sim \chi^2_{n-1}$

(ii) MSE *é v.a. independente de* $\hat{\beta}_1$.

A partir dele será possível definir as quantidades pivotais para os IC e as estatísticas de testes para os testes de hipóteses para o modelo de regressão simples pela origem.

1.9.3 Inferências para β_1

Veja na Equação 1.17 que nesse caso já é claro que $\hat{\beta}_1$ é combinação linear de $\{y\}$, logo $\hat{\beta}_1$ é variável aleatória normal. Vamos calcular sua média e sua variância.

$$\mathrm{E}[\hat{\beta}_1] = \mathrm{E}\left[\frac{\sum_{i=1}^{n} y_i x_i}{\sum_{i=1}^{n} x_i^2}\right] = \frac{\sum_{i=1}^{n} \mathrm{E}\left[y_i\right] x_i}{\sum_{i=1}^{n} x_i^2} = \frac{\sum_{i=1}^{n} (\beta_1 x_i) x_i}{\sum_{i=1}^{n} x_i^2} = \beta_1 \frac{\sum_{i=1}^{n} x_i^2}{\sum_{i=1}^{n} x_i^2} = \beta_1.$$

$$\mathrm{Var}(\hat{\beta}_1) = \mathrm{Var}\left(\frac{\sum_{i=1}^{n} y_i x_i}{\sum_{i=1}^{n} x_i^2}\right) = \frac{\sum_{i=1}^{n} \mathrm{Var}(y_i) x_i^2}{\left(\sum_{i=1}^{n} x_i^2\right)^2} = \frac{\sigma^2 \sum_{i=1}^{n} x_i^2}{\left(\sum_{i=1}^{n} x_i^2\right)^2} = \frac{\sigma^2}{\sum_{i=1}^{n} x_i^2}.$$

Logo, $\hat{\beta}_1 \sim N(\beta_1, \frac{\sigma^2}{\sum_{i=1}^n x_i^2})$. Juntando esse resultado com o do Teorema 1.9.2 é possível encontrar a quantidade pivotal

$$T = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{MSE}{\sum_{i=1}^n x_i^2}}} \sim t_{n-1}$$

e assim definir um intervalo de confiança para β_1:

$$\hat{\beta}_1 \pm t_{1-\frac{\alpha}{2},n-1}\sqrt{\frac{MSE}{\sum_{i=1}^n x_i^2}}. \tag{1.19}$$

De forma análoga, para testar as hipóteses $H_0 : \beta_1 = 0$ contra $H_1 : \beta_1 \neq 0$ podemos usar a estatística de teste

$$T = \frac{\hat{\beta}_1}{\sqrt{\frac{MSE}{\sum_{i=1}^n x_i^2}}}$$

que sob H_0 tem distribuição t_{n-1}.

1.9.4 Inerências para a variável resposta

No caso da regressão pela origem a estimativa pontual para a média da variável resposta no nível $x = x_i$ é definida por

$$\mu_i = \mathrm{E}[y_i] = \beta_1 x_i.$$

Um estimador pontual para esse parâmetro pode ser definido por

$$\hat{y}_i = \hat{\beta}_1 x_i = \sum_{j=1}^n \frac{x_i x_j}{\sum_{k=1}^n x_k^2} y_j.$$

Então $\hat{y}_i$ é variável aleatória normal. Vamos encontrar a sua média e sua variância.

$$\mathrm{E}\left[\hat{y}_i\right] = \mathrm{E}\left[\sum_{j=1}^n \frac{x_i x_j}{\sum_{k=1}^n x_k^2} y_j\right] = \sum_{j=1}^n \frac{x_i x_j}{\sum_{k=1}^n x_k^2} \mathrm{E}\left[y_j\right] = \sum_{j=1}^n \frac{x_i x_j}{\sum_{k=1}^n x_k^2} \left(\beta_1 x_j\right) = \beta_1 x_i \frac{\sum_{j=1}^n x_j^2}{\sum_{k=1}^n x_k^2} = \beta_1 x_i$$

$$\mathrm{Var}\left(\hat{y}_i\right) = \mathrm{Var}\left(\sum_{j=1}^n \frac{x_i x_j}{\sum_{k=1}^n x_k^2} y_j\right) = \sum_{j=1}^n \left(\frac{x_i x_j}{\sum_{k=1}^n x_k^2}\right)^2 \mathrm{Var}\left(y_j\right) = \sigma^2 x_i^2 \frac{\sum_{j=1}^n x_j^2}{\left(\sum_{k=1}^n x_k^2\right)^2} = \sigma^2 \frac{x_i^2}{\sum_{k=1}^n x_k^2}$$

Logo, $\hat{y}_i \sim N(\beta_1 x_i, \sigma^2 \frac{x_i^2}{\sum_{k=1}^n x_k^2})$. Juntando esse resultado com o do Teorema 1.9.2 é possível encontrar a quantidade pivotal para μ_i

$$T = \frac{\hat{\beta}_1 x_i - \mu_i}{\sqrt{\frac{x_i^2}{\sum_{k=1}^n x_k^2}}} \sim t_{n-1}$$

e assim definir um intervalo de confiança para μ_i:

$$\hat{\beta}_1 x_i \pm t_{1-\frac{\alpha}{2},n-1}\sqrt{MSE\frac{x_i^2}{\sum_{k=1}^n x_k^2}}. \tag{1.20}$$

1.10 Resíduos na Regressão Linear Simples

O i-ésimo resíduo em um modelo de regressão linear é definido como a diferença entre o valor observado e o valor ajustado da variável resposta. Ou seja,

$$e_i = y_i - \hat{y}_i \tag{1.21}$$

Veja a diferença entre o erro e o resíduo na regressão linear.

$$\begin{aligned} \text{Erro:} \quad & \varepsilon_i = y_i - (\beta_0 + \beta_1 x_i) \sim N(0, \sigma^2) \\ \text{Resíduo:} \quad & e_i = y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right) \sim N(0, ?) \end{aligned}$$

Veremos mais a frente, na Seção 3.5.1, que $\text{Var}(e_i) \neq \sigma^2$ e que e_i e e_j não são independentes para $i \neq j$. Mas apesar disso, para amostras grandes, isto é n grande, podemos considerar aproximadamente

$$e_i \sim N\left(0, \sigma^2\right), \text{Cov}(e_i, e_j) = 0 \text{ para } i \neq j.$$

Mas é importante saber que isso é uma aproximação e na verdade os erros e os resíduos são variáveis aleatórias diferentes e inclusive com distribuições diferentes, apesar de serem assintoticamente identicamente distribuídas.

Baseado na distribuição amostral aproximada para os resíduos é mais comum usar uma padronização dos resíduos do que os próprios resíduos. Essa padronização é chamada de resíduo padronizado e definido por:

$$e_i^* = \frac{e_i - \text{E}[e_i]}{\sqrt{MSE}} = \frac{y_i - \hat{y}_i}{\sqrt{MSE}} \tag{1.22}$$

Veja que mantendo a aproximação para n grande podemos dizer que e_i^* é variável aleatória próxima de $N(0, 1)$.

1.10.1 Análise dos Resíduos - Diagnóstico no Modelo de Regressão Linear

Primeiro veja que, fazendo as devidas aproximações, podemos considerar que cada resíduo e_i^* é uma variável aleatória $N(0, 1)$. Veja que essa é uma variável aleatória para a qual conseguimos observar uma amostra de tamanho n: $\{e_1^*, e_2^*, \ldots, e_n^*\}$. Isso não acontece com os erros por exemplo, os erros são variáveis aleatórias que não observamos.

Então, se fizermos um gráfico de $e_i^* \times x_i$ ou $e_i^* \times \hat{y}_i$ esperamos encontrar pontos aleatoriamente distribuídos em torno do zero, sem nenhum padrão de crescimento, decrescimento ou aumento de amplitude. Um exemplo desse do gráfico esperado para $e_i^* \times x_i$ ou $e_i^* \times \hat{y}_i$ encontra-se na Figura 1.5.

Se esse padrão não for verificado temos um indicativo de que alguma(s) suposição feita pelo modelo está sendo violada. Só para lembrar, as suposições feitas no Modelo de Regressão Linear são:

1. Existe uma relação linear entre x_i e $\text{E}[y_i]$;
2. Os erros tem variância constante σ^2, isto é, não depende de x_i;
3. Os erros são normalmente distribuídos;

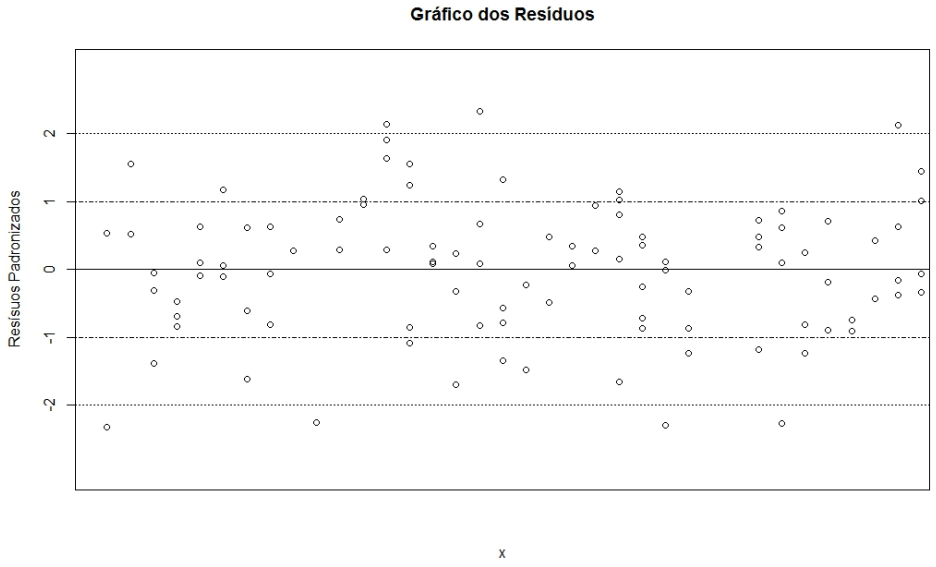


Figura 1.5: Padrão Esperado no Gráfico dos Resíduos

4. Os erros são variáveis aleatórias independentes.

Então, a partir da observação dos resíduos, seremos capazes de diagnosticar alguns problemas no Modelo de Regressão Linear. São eles:

1. Não linearidade: ausência de uma relação linear entre x_i e $\mathrm{E}[y_i]$;
2. Heterocedasticidade: erros com variância σ^2 não constante;
3. Não normalidade: erros (ou y) não são normalmente distribuídos;

Não-linearidade

Suponha que a relação entre x_i e $\mathrm{E}[y_i]$ não seja linear, mas rejeitamos a hipótese $\beta_1 = 0$ no teste t. Isso significa que existe um padrão de crescimento ou decrescimento de y conforme x cresce, mas esse padrão não é linear, veja o gráfico da esquerda da Figura 1.6. Nesse caso o gráfico dos resíduos versus x ou $\hat{y}$ não será como na Figura 1.5 e sim como o gráfico da direita da Figura 1.6.

Então sempre que encontramos um dos padrões apresentados na Figura 1.7 para os gráficos de $e_i^* \times x_i$ ou $e_i^* \times \hat{y}_i$ vamos diagnosticar a existência de não-linearidade. Nesse momento ainda não estamos preocupados com as medidas de correção, somente com o diagnóstico.

Heterocedasticidade

Suponha que a variância σ^2 seja não constante em função de x, então conforme x muda os resíduos que tem distribuição $N(0, \sigma^2)$ vão ficar com maior ou menor variabilidade e consequentemente o mesmo acontece com os resíduos padronizados, que simplesmente trata-se dos resíduos dividido por uma constate. Veja no gráfico da esquerda da Figura 1.8 um exemplo de uma amostra com variância não constante e no gráfico da direita os resíduos para o modelo linear em função de x.

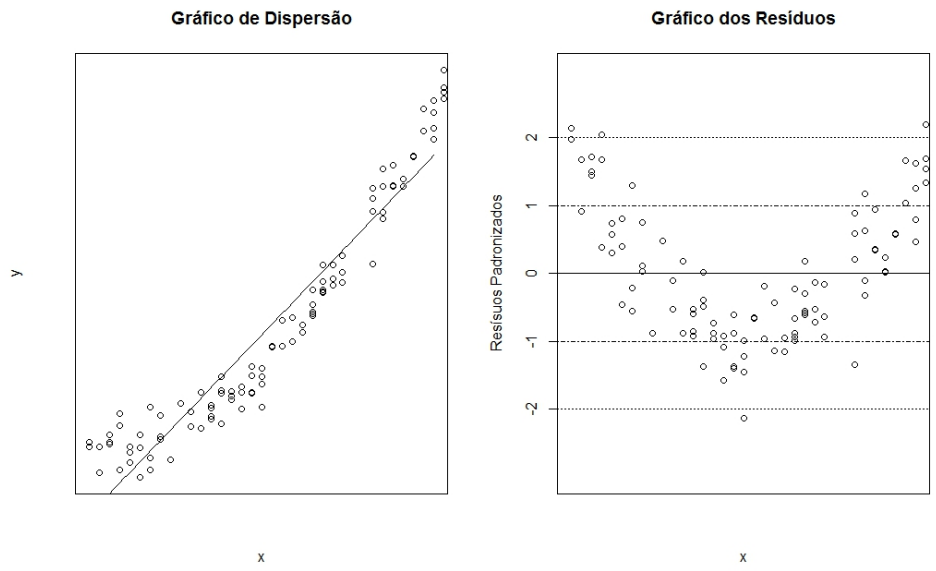


Figura 1.6: Gráfico dos Resíduos para Não Linearidade

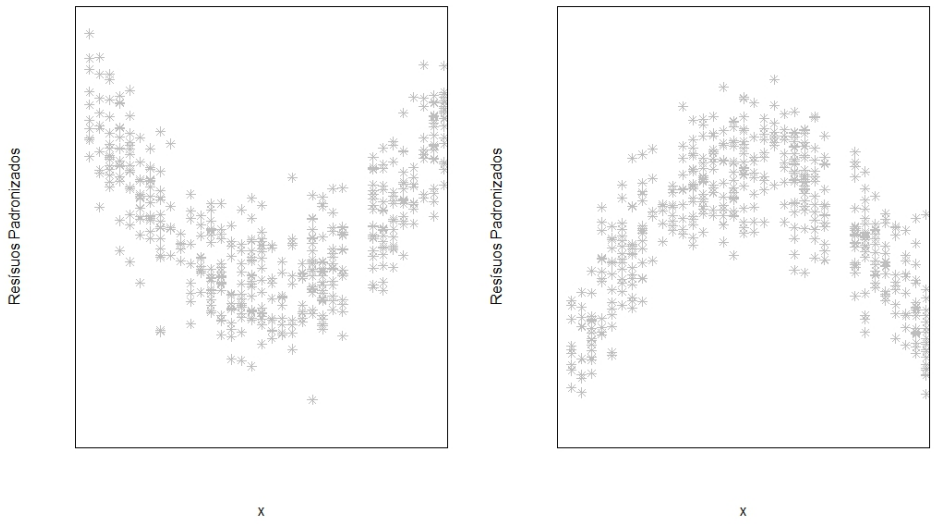


Figura 1.7: Padrões de Não Linearidade

Então sempre que encontramos um dos padrões apresentados na Figura 1.9 para os gráficos de $e_i^* \times x_i$ ou $e_i^* \times \hat{y}_i$ vamos diagnosticar a existência de Heterocedasticidade, isto é, variância não constante.

Para detectar a Heterocedasticidade é muito comum usar também os seguintes gráficos: $|e_i^*| \times x_i$ ou $(e_i^*)^2 \times x_i$, como mostra a Figura 1.10. Nesses gráficos o crescimento ou decrescimento da variabilidade dos resíduos com a mudança de x fica mais acentuado o que facilita o diagnóstico.

Caso haja uma desconfiança sobre a heterocedasticidade existem alguns testes que podem ajudar nesse diagnóstico. Esses testes verificam as hipóteses

$$H_0 : \text{variância constante} \qquad H_1 : \text{variância não constante.}$$

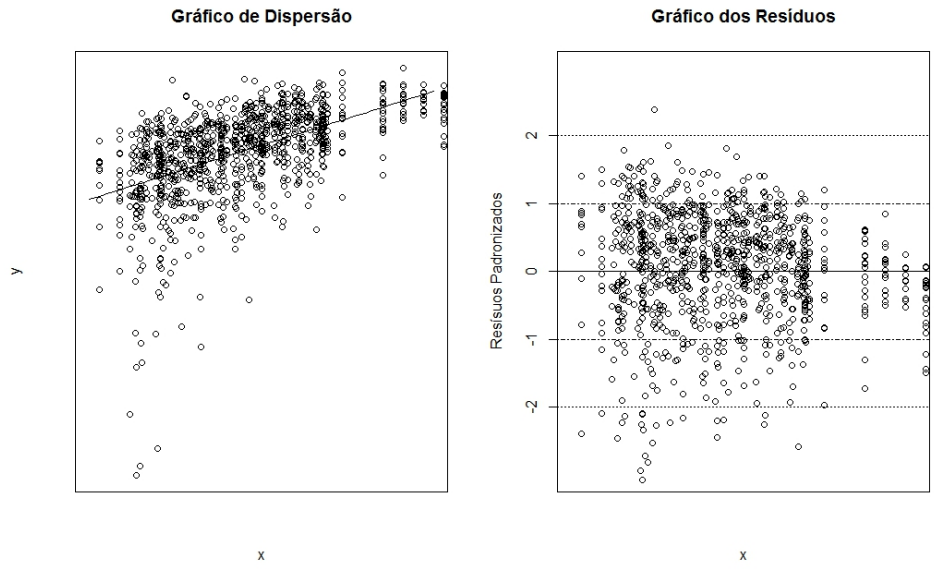


Figura 1.8: Gráfico dos Resíduos para Heterocedasticidade

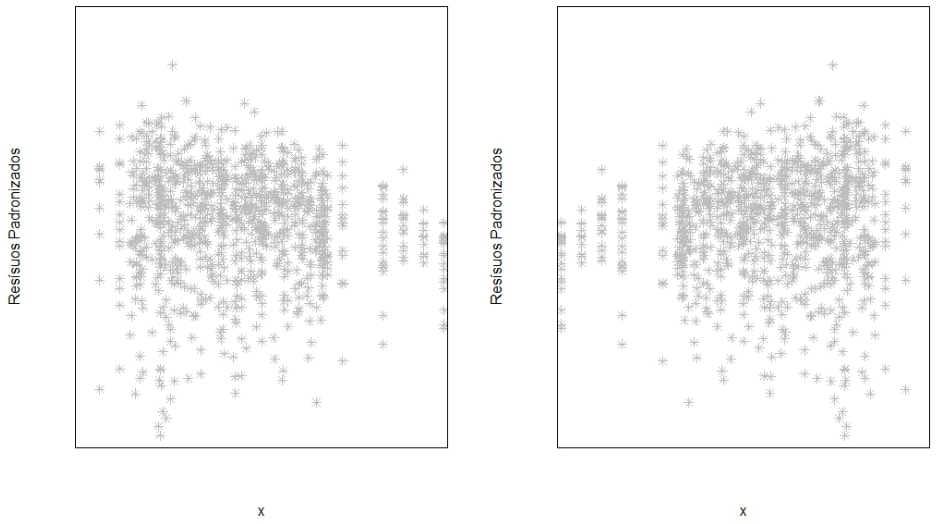


Figura 1.9: Padrões de Heterocedasticidade

Como exemplo podemos citar *Brown-Forsythe Test* e *Breusch-Pagan Test*. Para saber mais sobre esses dois testes veja a seção 3.6 de [Kutner et al., 2005]. O *Breusch-Pagan Test* está no pacote `lmtest` do R e pode ser rodado a partir do comando `bptest`.

Ainda não estamos preocupados com as medidas de correção, somente com o diagnóstico, mais a frente veremos como resolver esse problema.

Não normalidade

Para detectar a não normalidade dos erros vamos verificar se os resíduos são aparentemente normais. Lembre-se de que chegamos a conclusão de que os resíduos são normais partindo da hipótese de que os erros são normais, ou seja, se verificarmos que os resíduos

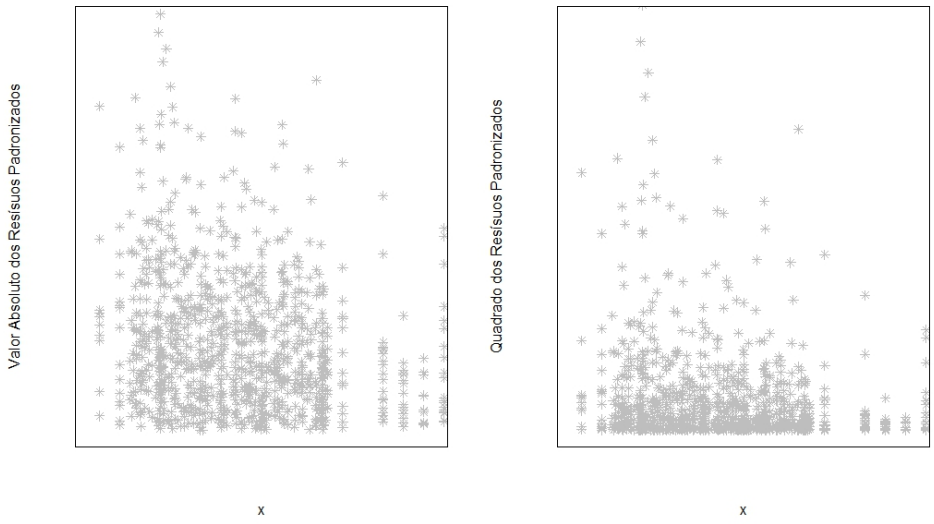


Figura 1.10: Outros Gráficos para Diagnosticar Heterocedasticidade

não são normais provavelmente a hipótese de que os erros são normais não é verdadeira.

Para verificar se a amostra de resíduos segue uma distribuição normal vamos usar o comando `qqnorm` do R, que plota os quantis amostrais versus os quantis teóricos. Se a amostra é de $N(0,1)$ então os pontos devem seguir a reta identidade, por isso é importante nesse gráfico usar sempre os resíduos padronizados. Na Figura 1.11 temos um exemplo do gráfico `qqnorm` para um modelo bem ajustado, isto é, para os resíduos padronizados normalmente distribuídos.

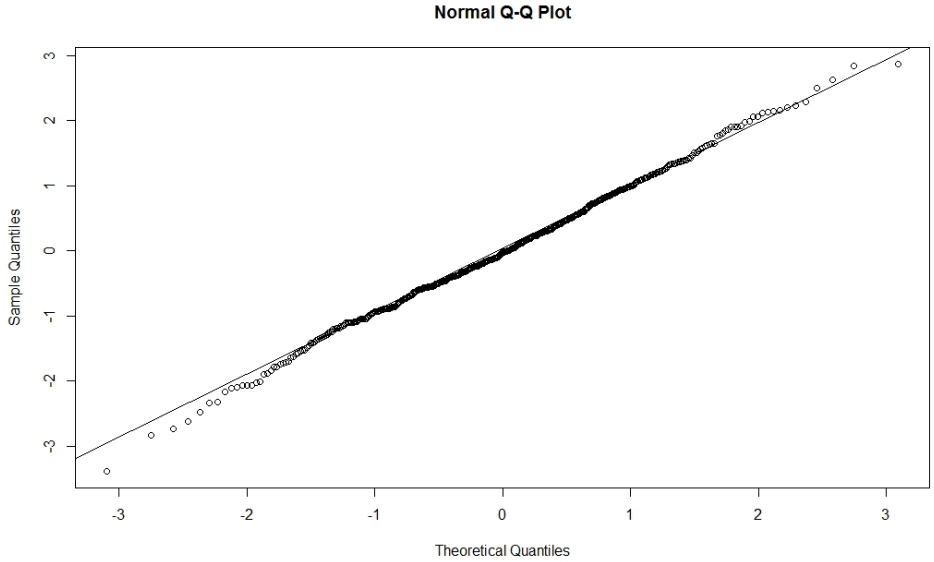


Figura 1.11: Padrão esperado no `qqnorm`

Quando encontramos no gráfico `qqnorm` um dos padrões apresentados na Figura 1.12 vamos diagnosticar não normalidade. O primeiro deles, à esquerda, indica que os erros têm distribuição assimétrica para a direita. O segundo, gráfico do meio, indica que os erros têm distribuição assimétrica para a esquerda. O gráfico da direita indica que os erros tem distribuição simétrica mas com caudas mais pesadas que as da normal.

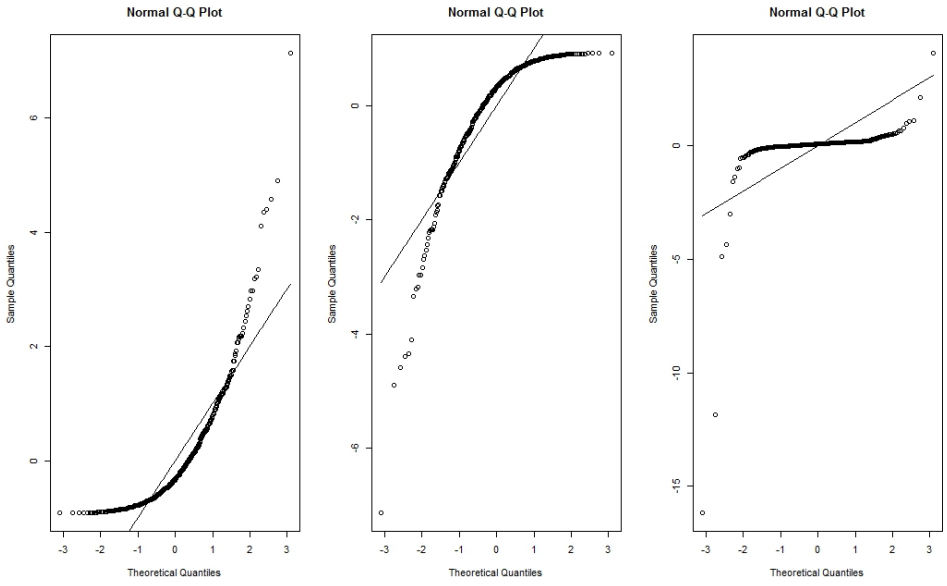


Figura 1.12: Padrões de Não Normalidade

Caso haja uma desconfiança sobre a não normalidade ainda é possível fazer alguns testes para ajudar nesse diagnóstico, como por exemplo, o teste de Kolmogorov-Smirnov ou o teste Qui-quadrado para checar se a amostra de resíduos padronizados vem de uma distribuição $N(0, 1)$.

1.10.2 Coeficiente de Determinação

O Coeficiente de Determinação é uma medida geralmente utilizada para avaliar o quanto boa é uma regressão linear. Esse coeficiente é denominado R^2 e definido por:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} = 1 - \frac{SSE}{SSTO} \tag{1.23}$$

onde SSE é a soma dos resíduos ao quadrado, também chamado de variância residual, e $SSTO$ é a variância total do modelo.

Veja que $SSTO \leq SSE$ uma vez que o nosso modelo realiza estimativas sempre melhores que a média, pois se a média é um caso particular do modelo $\hat{\beta}_1 = 0$ e a escolha de $\hat{\beta}_1$ e $\hat{\beta}_0$ é aquela que minimiza a soma dos resíduos ao quadrado, uma vez que usamos os estimadores de mínimos quadrados. Com isso podemos afirmar que $0 \leq R^2 \leq 1$ e quanto melhor é a regressão comparada com a média amostral mais próximo de 1 fica R^2.

Apesar dessa ser uma medida com alguns problemas, ela é bastante usada na análise de regressão.

Exercícios para Aulas Práticas do Capítulo 1

1. Na aula prática de hoje vamos primeiro fazer as contas a partir das expressões vistas em sala de aula. Depois, a partir do item (1l), vamos comparar a resposta com o resultado fornecida pelos comandos `lm` e `summary` do R.

   A *Toluca Company* fabrica equipamentos de refrigeração e suas peças são produzidas em lotes de tamanhos variados. Quando um programa de melhoria de custo foi implementado a empresa resolveu estudar a relação entre o tamanho do lote produzido e o número de horas de trabalho para a sua produção. Para isso foi recolhida uma amostra de tamanho 25, que está disponível pelo link `CH01TA01.txt` (`https://netfiles.umn.edu/users/nacht001/www/nachtsheim/5th/KutnerData/Chapter%20%201%20Data%20Sets/CH01TA01.txt`). A primeira coluna dessa tabela se refere ao tamanho do lote (x) enquanto a segunda se refere à quantidade de horas trabalhadas (y).

   (a) Salve os dados disponibilizados no link em um arquivo `CH01TA01.txt`.

   (b) Abra o R e usando o comando `read.table` crie um objeto `data.frame` com os dados da amostra.

   (c) Usando o comando `plot` faça o gráfico de dispersão dos pontos (x, y).
   Como o gráfico sugere uma relação linear vamos criar um modelo de regressão linear para descrever a relação entre a variável determinística x = tamanho do lote e a variável aleatória y = horas de trabalho.

   $$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

   (d) Encontre a estimativa para o parâmetro β_1 por mínimos quadrados.

   (e) Encontre a estimativa para o parâmetro β_0 por mínimos quadrados.

   (f) Defina a reta de regressão estimada e faça seus gráfico junto com os pontos já plotados. Para isso use o comando `curve`.

   (g) Encontre uma estimativa para σ^2, a variância dos erros.

   (h) Encontre um intervalo de confiança para β_1 com confiabilidade de 95%.

   (i) Encontre um intervalo de confiança para β_0 com com confiabilidade de 95%.

   (j) Teste se existe uma relação linear entre as variáveis x e y. Use nível de significância de 5%.

   (k) Encontre o p-valor para o teste feito no item acima.

   (l) Usando os comandos `lm` e `summary` identifique nas saídas onde estão os valores de: $\hat{\beta}_1$, $\hat{\beta}_0$, MSE, a estatística de teste e o p-valor do testes t, item (1j).

   (m) Como podemos construir os intervalos de confiança para β_1 e β_0 a partir da saída dos comandos `lm` e `summary`?

   (n) Baseado na saída do comando `summary` podemos ver que o p-valor para o teste t que verifica as hipóteses $H_0 : \beta_0 = 0$ contra $H_1 : \beta_0 \neq 0$ é 0.0259. Usando essa informação, qual a sua conclusão em relação à essas hipóteses? Qual a sua interpretação do resultado encontrado?

2. Vamos continuar usando a o banco de dados do exercício 1.

   (a) Encontre uma estimativa pontual para o número médio de horas trabalhadas em lotes de tamanho 65. Interprete o resultado.

   (b) Encontre um intervalo de confiança com confiabilidade de 90% para $\mathrm{E}[y_h \mid x_h = 65]$. Repita agora para $x_h = 100$. Percebeu que quanto mais longe x_h está de $\bar{X}$ maior é a amplitude do intervalo para a mesma confiabilidade?

   (c) A empresa também está interessada em utilizar o modelo de regressão para prever o tempo de trabalho para um novo lote. Suponha que o tamanho do próximo lote a ser produzido é de $x_h = 100$ unidades. Encontre um intervalo de confiança com confiabilidade de 90% para o número de horas necessário para a produção desse novo lote. Compare o resultado com o intervalo de confiança para $\mathrm{E}[y_h \mid x_h = 100]$ encontrado no exercício anterior.

   (d) Em um mesmo gráfico apresente as seguintes informações:

   - O gráfico de dispersão dos pontos (x, y) da amostra.
   - A reta de regressão estimada.
   - Intervalos de confiança para $\mathrm{E}[y_h \mid x_h]$ com confiabilidade de 90% para diferentes valores de x_h.
   - Intervalos de predição para $y_{novo} \mid x_{novo}$ com confiabilidade de 90% para diferentes valores de x_{novo}.

3. A Charles Plumbing Supplies Company opera 12 armazéns. Na tentativa de reforçar os procedimentos de planejamento e controle um consultor estudou a relação entre o número de unidade de trabalho realizado (x) e o custo com mão de obra (y) para executar esse trabalho nos armazéns durante um período de teste. O resultado desse estudo pode ser encontrado em `CH04TA02.txt` (`https://netfiles.umn.edu/users/nacht001/www/nachtsheim/5th/KutnerData/Chapter%20%204%20Data%20Sets/CH04TA02.txt`), onde a primeira coluna apresenta os valores de x e a segunda os valores de y para cada armazém da empresa.

   O consultor optou por utilizar um Modelo de Regressão Linear passando pela Origem, pois é natural pensar que quando não há produção alguma o gasto médio com mão de obra é nulo.

   Faça os itens a seguir sem usar o comando `lm` a menos que isso seja pedido.

   (a) Primeiro faça o gráfico de dispersão e veja os pontos (x, y) amostrados. Dê a sua opinião sobre a escolha do modelo de regressão linear pela origem.

   (b) Encontre as estimativas para os parâmetros β_1 e σ^2 do modelo de regressão linear pela origem.

   (c) Adicione ao gráfico de dispersão feito no item (3a) a reta de regressão estimada do modelo.

   (d) Encontre um intervalo de confiança para o parâmetro β_1 e interprete o resultado. Use confiabilidade de 95%.

   (e) Encontre um intervalo de confiança para o custo médio com a mão de obra para armazéns que realizam 100 unidades de trabalho.

(f) Suponha que um novo armazém será administrado por essa empresa. Defina um intervalo de predição para o custo com a mão de obra deste armazém se ele realizar 100 unidades de trabalho?

(g) Agora usando o comando `lm` do R encontre os seguintes valores:

- Estimativa pontual para β_1.
- Estimativa intervalar para β_1.
- Estimativa pontual para $\mathrm{E}[y_h \mid x_h = 100]$.
- Estimativa intervalar para $\mathrm{E}[y_h \mid x_h = 100]$.
- Intervalo de predição para $y_{novo} \mid x_{novo} = 100$.

4. Vamos continuar usando a o banco de dados do exercício 1.

(a) Obtenha os resíduos ordinários e faça o gráfico de $x_i \times e_i$ e $\hat{y}_i \times e_i$.

(b) Obtenha os resíduos padronizados e faça o gráfico de $x_i \times e_i^*$ e $\hat{y}_i \times e_i^*$.

(c) A partir dos gráficos acima comente sobre as suposições de linearidade e homocedasticidade do modelo.

(d) Faça agora o `qqnorm` com os resíduos padronizados e comente sobre a suposição de normalidade dos erros.

5. Acredita-se que a massa muscular das pessoas diminui com a idade. Para verificar essa relação 60 mulheres foram selecionadas, de forma aleatória, com idade variando entre 40 e 79 anos. O arquivo `CH01PR27.txt` (`https://netfiles.umn.edu/users/nacht001/www/nachtsheim/5th/KutnerData/Chapter%20%201%20Data%20Sets/CH01PR27.txt`) contém os dados de massa muscular (y - coluna 1) e idade (x - coluna 2) de cada uma dessas mulheres.

(a) Assumindo que o modelo de regressão linear seja adequado determine as estimativas para os parâmetros do modelo e faça o gráfico de dispersão dos dados junto com a reta de regressão estimada.

(b) Conduza o teste que verifica a afirmação de que a massa muscular das mulheres diminui com a idade. Enuncie as hipóteses que serão testadas, a regra de decisão e a conclusão. Qual foi o p-valor deste teste?

(c) Obtenha os resíduos ordinários e os resíduos padronizados.

(d) Faça o(s) gráfico(s) adequado(s) para checar a suposição de linearidade do modelo.

(e) Faça agora o(s) gráfico(s) adequado(s) para checar a suposição de homocedasticidade do modelo, isto é, variância constate. Faça também o teste de *Breusch-Pagan* de forma a reforçar a sua conclusão, para isso use o comando `bptest` do pacote `lmtest`.

(f) Para terminar a análise dos resíduos faça o(s) gráfico(s) adequado(s) para checar a suposição de normalidade dos erros. Faça também o teste qui-quadrado ou Kolmogorov-Smirnov de forma a reforçar a sua conclusão.

(g) Calcule o valor de R^2 deste ajuste.

6. Um químico estudo a concentração de uma solução ao longo do tempo. Para isso ele preparou 15 soluções idênticas que foram separadas em 5 grupos de 3. Cada grupo de 3 soluções teve a concentração medida em um dos instantes 1, 3, 5, 7 e 9. Com isso foi possível construir a amostrada apresentada em `CH03PR15.txt` (`https://netfiles.umn.edu/users/nacht001/www/nachtsheim/5th/KutnerData/Chapter%20%203%20Data%20Sets/CH03PR15.txt`), onde a primeira indica a concentração da solução (y) e a segunda o instante em que a concentração foi medida (x).

   (a) Ajuste o modelo de regressão linear na amostra encontrada e faça o gráfico de dispersão junto com a reta de regressão estimada.

   (b) Obtenha os resíduos ordinários e os resíduos padronizados.

   (c) Faça o(s) gráfico(s) adequado(s) para checar a suposição de linearidade do modelo.

   (d) Faça agora o(s) gráfico(s) adequado(s) para checar a suposição de homocedasticidade do modelo, isto é, variância constate. Faça também o teste de *Breusch-Pagan* de forma a reforçar a sua conclusão, para isso use o comando `bptest` do pacote `lmtest`.

   (e) Para terminar a análise dos resíduos faça o(s) gráfico(s) adequado(s) para checar a suposição de normalidade dos erros. Faça também o teste qui-quadrado ou Kolmogorov-Smirnov de forma a reforçar a sua conclusão.

   (f) Calcule o valor de R^2 deste ajuste.

7. Em uma manufatura deseja-se estudar a relação entre o tamanho dos lotes produzidos e o tempo gasto em sua produção. Para isso 111 lotes de tamanhos diferentes forma produzidos e tempo gasto em sua produção anotado, esses valores podem ser encontrados em `CH03PR18.txt` (`https://netfiles.umn.edu/users/nacht001/www/nachtsheim/5th/KutnerData/Chapter%20%203%20Data%20Sets/CH03PR18.txt`), onde a primeira coluna é o tempo de produção de cada lote (y) e na segunda o tamanho do lote (x).

   (a) Ajuste o modelo de regressão linear nos dados e faça o gráfico de dispersão junto com a reta de regressão estimada. Observando esse primeiro gráfico o que você já pode dizer sobre a linearidade? É adequada?

   (b) Obtenha os resíduos ordinários e os resíduos padronizados.

   (c) Faça o(s) gráfico(s) adequado(s) para checar a suposição de linearidade do modelo.

   (d) Faça agora o(s) gráfico(s) adequado(s) para checar a suposição de homocedasticidade do modelo, isto é, variância constate. Faça também o teste de *Breusch-Pagan* de forma a reforçar a sua conclusão, para isso use o comando `bptest` do pacote `lmtest`.

   (e) Para terminar a análise dos resíduos faça o(s) gráfico(s) adequado(s) para checar a suposição de normalidade dos erros. Faça também o teste qui-quadrado ou Kolmogorov-Smirnov de forma a reforçar a sua conclusão.

   (f) Calcule o valor de R^2 deste ajuste.

Lista de Exercícios do Capítulo 1

1.1. Quando solicitado para definir o modelo de regressão linear simples um aluno escreveu a seguinte equação: $\mathrm{E}[y_i] = \beta_0 + \beta_1 x_i + \varepsilon_i$. Você concorda com o que o aluno escreveu? Justifique.

1.2. Suponha o modelo de regressão linear simples com $\beta_0 = 100$, $\beta_1 = 20$ e $\sigma^2 = 25$. Para esse modelo será observado uma saída y para $x = 5$.

(a) Podemos afirmar a exata probabilidade de y estar entre 195 e 205? Explique.

(b) Se o modelo normal for considerado, agora é possível afirmar a exata probabilidade de y estar entre 195 e 205? Se sim, encontre essa probabilidade.

1.3. Uma substância utilizada na pesquisa médica e biológica é enviado por via aérea aos usuários em embalagens de 1.000 ampolas. Os dados em `CH01PR21.txt` (`https://netfiles.umn.edu/users/nacht001/www/nachtsheim/5th/KutnerData/Chapter%20%201%20Data%20Sets/CH01PR21.txt`) se referem ao número de vezes que a embalagem foi transferida de uma aeronave para uma outra (x - coluna 2) e o número de ampolas quebradas dentro da embalagem (y - coluna 1), verificadas na chegada ao destino. Assuma que o modelo de regressão linear simples seja adequado para relacionar as variáveis x e y.

(a) Encontre a reta de regressão estimada. Faça o gráfico da reta junto com os dados da amostra. O modelo de regressão parece adequado para os dados?

(b) Encontre a estimativa pontual para o número esperado de ampolas quebradas quando $x = 1$ transferência é realizada.

1.4. Acredita-se que o modelo de regressão linear seja apropriado para descrever a relação entre a dureza do plástico (y) e o tempo decorrido desde a finalização do processo de moldagem (x). Para verificar essa relação dezesseis lotes de plástico foram fabricados em diferentes níveis x pré-determinados. De cada lote um produto teste foi selecionado e a sua dureza foi medida. Esses dados encontram-se em `CH01PR22.txt` (`https://netfiles.umn.edu/users/nacht001/www/nachtsheim/5th/KutnerData/Chapter%20%201%20Data%20Sets/CH01PR22.txt`), onde na primeira coluna estão os valores da dureza y, medidos em *Brinell*, e na segunda coluna os valores dos diferentes tempos decorridos desde a finalização do processo de moldagem x, medidos em em horas.

(a) Encontre a reta de regressão estimada. Faça o gráfico da reta junto com os dados da amostra. O modelo de regressão parece adequado para os dados?

(b) Encontre a estimativa pontual para o número esperado de ampolas quebradas quando $x = 1$ transferência é realizada.

(c) Encontre uma estimativa pontual para a mudança na média da dureza do plástico quando quando x é acrescido em 1 hora.

1.5. Qual a implicação na reta de regressão se $\beta_1 = 0$ de forma que o modelo passe a ser $y_i = \beta_0 + \varepsilon_i$? Como seria o gráfico da reta de regressão nesse caso?

1.6. Ainda supondo o modelo de regressão $y_i = \beta_0 + \varepsilon_i$ do exercício 1.5, encontre o estimador por mínimos quadrados para o parâmetro β_0 desse modelo.

1.7. Um aluno que trabalha em um estágio de verão no Departamento de Pesquisas Econômicas de uma grande corporação estuda a relação entre as vendas de um produto (y, em milhões de dólares) e o tamanho da população (x, em milhões de pessoas) em 50 distritos. O modelo de regressão linear normal foi adotado nesse estudo. Primeiro o aluno gostaria de testar se existe uma relação linear entre x e y. Para isso o aluno acessou um simples programa de regressão linear e obteve as seguintes informações com relação aos coeficientes do modelo:

		95 Percent	
Parameter	**Estimated Value**	**Confidence Limits**	
Intercept	7.43119	-1.18518	16.0476
Slope	.755048	.452886	1.05721

(a) Baseado nas informações passadas pelo programa o aluno concluiu que existe uma relação linear entre x e y. Essa conclusão é garantida? Qual o nível de significância adotado?

(b) Alguém questionou o valor negativo para o limite inferior do intervalo de confiança para β_0 uma vez que as vendas em dólar não podem ser negativas mesmo quando a população for nula. Discuta esse comentário com outros alunos da turma.

1.8. Em um teste cujas hipóteses são $H_0 : \beta_1 \leq 0$ contra $H_1 : \beta_1 > 0$ o analista aceitou H_0. Essa conclusão implica em não existir associação linear entre as variáveis x e y? Explique.

1.9. Este problema se refere aos dados do exercício 1.3.

(a) Estime um intervalo de confiança para β_1 com 95% de confiabilidade. Interprete o seu intervalo estimado.

(b) Faça o teste t a fim de verificar se realmente existe uma relação linear entre o número de vezes que a embalagem foi transferida de uma aeronave para uma outra (x) e o número de ampolas quebradas dentro da embalagem (y). Use um nível de significância de 5%. Enuncie as hipóteses que serão testadas, a regra de decisão e a conclusão. Qual foi o p-valor deste teste?

(c) Encontre um intervalo de confiança com 95% de confiabilidade para o parâmetro β_0 e interprete o resultado.

(d) Um consultor sugeriu, baseado em experiências anteriores, que o número médio de ampolas quebradas não passa de 9 quando não é feita transferência alguma. Formalize um teste apropriado para verificar a afirmação do consultor. Use $\alpha = 0.025$. Enuncie as hipóteses que serão testadas, a regra de decisão e a conclusão. Qual foi o p-valor deste teste?

1.10. Este problema se refere aos dados do exercício 1.4.

(a) Encontre uma estimativa intervalar para a mudança na média da dureza do plástico quando quando x é acrescido em 1 hora. Use 99% de confiança. Interprete o resultado.

(b) O fabricante de plástico afirma que a dureza aumenta em média 2 unidades de *Brinell* por hora. Formalize um teste apropriado para verificar a afirmação do fabricante. Use $\alpha = 0.01$. Enuncie as hipóteses que serão testadas, a regra de decisão e a conclusão. Qual foi o p-valor deste teste?

1.11. Seja $k_i = \frac{x_i - \bar{x}}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$. Mostre que $\sum_{i=1}^{n} k_i x_i = 1$.

1.12. Considere o modelo de regressão linear simples $y = \beta_0 + \beta_1 x + \varepsilon$ com $\mathrm{E}(\varepsilon) = 0$, $\mathrm{Var}(\varepsilon) = \sigma^2$ e ε descorrelatados. Mostre que $\mathrm{Cov}(\hat{\beta}_0, \hat{\beta}_1) = -\bar{x}\sigma^2/S_{xx}$.
Dica: escreva $\hat{\beta}_0$ e $\hat{\beta}_1$ como combinação linear de $\{y_i\}$ como fizemos em sala de aula.

1.13. Para cada uma das perguntas a seguir explique se é apropriado usar o intervalo de confiança para a média da variável resposta ou o intervalo de predição para uma nova observação. Aproveite para identificar em cada situação quem é a variável de predição x e quem é a variável resposta y.

(a) Qual será o nível de umidade na estufa amanhã quando a temperatura foi colocada em 31^oC?

(b) Qual o gasto médio de uma família em refeições fora de casa quando seu rendimento disponível para gastos é de \$ 23.500 ?

(c) Qual será o consumo de eletricidade, em kilowatt-hora, no próximo mês pelos comércio e pela industria da cidade, supondo que o índice de atividade empresarial para a área será o mesmo do mês atual?

1.14. Este problema se refere aos dados do exercício 1.3.

(a) Devido a mudanças nas rotas aéreas as embalagens serão transferidas com mais frequências que no passado. Estime o número médio de ampolas quebradas para os seguintes números de transferências entre aeronaves: $x = 2, 4$. Encontre intervalos com 99% de confiança. Interprete os resultados.

(b) Uma próxima embalagem será enviada e a viagem terá 2 transferências. Encontre um intervalo de predição para o número de ampolas quebradas nessa embalagem. Use 99% de confiança. Interprete o intervalo de predição encontrado.

1.15. Este problema se refere aos dados do exercício 1.4.

(a) Encontre um intervalo de confiança para a média da dureza de itens moldados em plástico tais que o tempo decorrido desde a finalização do processo de moldagem seja de 30 horas. Use confiabilidade de 98% e interprete seu resultado.

(b) Encontre um intervalo de confiança para a dureza de um novo item moldado em plástico tal que o tempo decorrido desde a finalização do processo de moldagem seja de 30 horas. Use confiabilidade de 98% e interprete seu resultado.

(c) Discuta a diferença entre os dois intervalos encontrados nos itens acima. Qual a diferença também na interpretação deles?

1.16. Um analista ajustou um modelo de regressão linear e realizou o teste t para verificar se $\beta_1 = 0$ ou $\beta_1 \neq 0$. O p-valor do teste foi 0.033 e o analista então concluiu que $\beta_1 \neq 0$. O nível α usado pelo analista foi maior ou menor que 0.033? Se o nível α fosse 0.01 qual seria a conclusão apropriada?

1.17. Este problema se refere aos dados do exercício 1.3. Faça a análise dos resíduos e verifique se as premissas do Modelo de Regressão Linear parecem estar sendo respeitadas.

1.18. Este problema se refere aos dados do exercício 1.4. Faça a análise dos resíduos e verifique se as premissas do Modelo de Regressão Linear parecem estar sendo respeitadas.

1.19. Um sociólogo aplicou o modelo de regressão linear para relacionar os rendimentos per capita (y) com o número médio de anos na escola (x) em 12 cidades. A tabela no link `CH03PR10.txt` (`https://netfiles.umn.edu/users/nacht001/www/nachtsheim/5th/KutnerData/Chapter%20%203%20Data%20Sets/CH03PR10.txt`) apresenta na primeira coluna os valores ajustados $\hat{y}$ e na segunda coluna os resídios padronizados resultantes dessa modelagem.

(a) Faça o gráfico dos resíduos padronizados versus os valores ajustados. O que o gráfico sugere?

(b) Quantos resíduos padronizados estão fora do intervalo $[-1, 1]$? Aproximadamente quantos resíduos padronizados você espera encontrar fora desse intervalo se o modelo normal for apropriado?

1.20. Uma pesquisa de marketing estuda a venda anual de um produto que foi introduzido no mercado há dez anos. Os dados para a pesquisa estão apresentados no link `CH03PR17.txt` (`https://netfiles.umn.edu/users/nacht001/www/nachtsheim/5th/KutnerData/Chapter%20%203%20Data%20Sets/CH03PR17.txt`), onde a primeira coluna guarda os valores de x, um código que indica o ano de observação, e a segunda coluna guarda os valores de y, as vendas em milhares de unidades.

(a) Faça um gráfico de dispersão. Parece apropriado estabelecer uma relação linear entre x e y?

(b) Encontre a reta de regressão estimada a partir do modelo linear e faça seu gráfico junto com o gráfico de dispersão.

(c) Obtenha os resíduos ordinários e os resíduos padronizados.

(d) Faça o(s) gráfico(s) adequado(s) para checar a suposição de linearidade do modelo.

(e) Faça agora o(s) gráfico(s) adequado(s) para checar a suposição de homocedasticidade do modelo, isto é, variância constate. Faça também o teste de *Breusch-Pagan* de forma a reforçar a sua conclusão, para isso use o comando `bptest` do pacote `lmtest`.

(f) Para terminar a análise dos resíduos faça o(s) gráfico(s) adequado(s) para checar a suposição de normalidade dos erros. Faça também o teste qui-quadrado ou Kolmogorov-Smirnov de forma a reforçar a sua conclusão.

(g) Calcule o valor de R^2 deste ajuste.

Capítulo 2

Regressão Linear Múltipla

Nesse capítulo vamos definir o modelo de regressão linear múltiplo, isto é, quando temos mais de uma variável independente. Por exemplo, até agora a gente tetava explicar a produção de um funcionário y em função da quantidade de horas trabalhadas x. A partir de agora vamos poder introduzir no nosso modelo outras variáveis que talvez sejam relevantes para explicar a produção do funcionário: a idade, a quantidade de anos de experiência, o sexo do funcionário, o turno de trabalho, etc. O objetivo é definir um modelo linear que use todas essas variáveis independentes ao mesmo tempo para explicar a variável resposta y.

2.1 O Modelo de Regressão Linear Múltiplo

Nesta seção vamos definir o Modelo de Regressão Linear Múltiplo, mas antes de definir o caso geral veremos o caso particular do modelo com duas variáveis independentes.

O modelo com 2 variáveis independentes

Nesse caso o a variável resposta y será explicada a partir de duas variáveis independentes x_1 e x_2. O modelo linear supõe que E[y] varia de forma linear com x_1 e x_2, o que é definido pela equação a seguir:

$$y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2) \text{ são v.a. independentes}$$

Veja que agora temos:

- $y_i \sim N(\beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i}, \sigma^2)$;
- β_0 é a média da variável resposta ao nível zero;
- β_1 indica a mudança em E[y] quando x_1 aumenta em uma unidade;
- β_2 indica a mudança em E[y] quando x_2 aumanta em uma unidade;

O modelo geral

Vamos agora definir o modelo geral, isto é, o modelo com $p-1$ variáveis independentes.

Definição 2.1.1 *O* **Modelo de Regressão Linear Múltiplo** *é o modelo que define uma relação estatística linear entre a variável resposta y e* $p-1$ *variáveis independentes:* x_1, x_2, ..., x_{p-1}. *A suposição básica desse modelo é que a média da distribuição de y varia de forma linear com as variáveis* x_1, x_2, ..., x_{p-1}. *Essa relação pode ser estabelecido por:*

$$y_i = \beta_0+\beta_1 x_{i,1}+\beta_2 x_{i,2}+\ldots+\beta_{p-1}x_{i,p-1}+\varepsilon_i, \quad \varepsilon_i \sim N(0,\sigma^2) \text{ são v.a. independentes} \quad (2.1)$$

onde, $x_{i,j}$ *= o valor da j-ésima variável independente na i-ésima observação,* y_i *= o valor da variável resposta na i-ésima observação e* ε_i *o erro aleatório para a i-ésima observação.* β_0, β_1, β_2, ..., β_{p-1} *e* σ^2 *são parâmetros do modelo.*

OBS: p é o número de parâmetros β's. Se $p = 2$ temos o modelo simples. Se $p = 3$ temos o modelo com 2 variáveis independentes.

2.2 Forma Matricial para o Modelo de Regressão Linear Múltiplo

No caso do modelo múltiplo costumamos usar a notação matricial. Antes de apresentar a definição do modelo geral na forma matricial vamos ver como fica na forma matricial o modelo simples e o modelo com duas variáveis independentes.

O modelo simples

No modelo linear simples temos,

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \Rightarrow \begin{cases} y_1 = \beta_0 + \beta_1 x_1 + \varepsilon_1 \\ y_2 = \beta_0 + \beta_1 x_2 + \varepsilon_2 \\ y_3 = \beta_0 + \beta_1 x_3 + \varepsilon_3 \\ \vdots \\ y_n = \beta_0 + \beta_1 x_n + \varepsilon_n \end{cases}$$

Então podemos definir

$$\underline{y} = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{pmatrix}_{n\times 1}, \quad X = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}_{n\times 2}, \quad \underline{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}_{2\times 1} \text{ e } \quad \underline{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \vdots \\ \varepsilon_n \end{pmatrix}_{n\times 1}$$

E escrever

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

Ou seja, o modelo pode ser definido por:

$$\underline{y} = X\underline{\beta} + \underline{\varepsilon}.$$

O modelo com 2 variáveis independentes

No modelo linear com duas variáveis independentes temos,

$$y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \varepsilon_i \Rightarrow \begin{cases} y_1 = \beta_0 + \beta_1 x_{1,1} + \beta_2 x_{1,2} + \varepsilon_1 \\ y_2 = \beta_0 + \beta_1 x_{2,1} + \beta_2 x_{2,2} + \varepsilon_2 \\ y_3 = \beta_0 + \beta_1 x_{3,1} + \beta_2 x_{3,2} + \varepsilon_3 \\ \vdots \\ y_n = \beta_0 + \beta_1 x_{n,1} + \beta_2 x_{n,2} + \varepsilon_n \end{cases}$$

Então podemos definir

$$\underline{y} = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{pmatrix}_{n\times 1}, \quad X = \begin{pmatrix} 1 & x_{1,1} & x_{1,2} \\ 1 & x_{2,1} & x_{2,2} \\ 1 & x_{3,1} & x_{3,2} \\ \vdots & \vdots & \vdots \\ 1 & x_{n,1} & x_{n,2} \end{pmatrix}_{n\times 3}, \quad \underline{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}_{3\times 1} \text{ e } \quad \underline{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \vdots \\ \varepsilon_n \end{pmatrix}_{n\times 1}$$

E escrever

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{1,1} & x_{1,2} \\ 1 & x_{2,1} & x_{2,2} \\ 1 & x_{3,1} & x_{3,2} \\ \vdots & \vdots & \vdots \\ 1 & x_{n,1} & x_{n,2} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

Ou seja, o modelo também pode ser definido por:

$$\underline{y} = X\underline{\beta} + \underline{\varepsilon}.$$

O modelo geral

De forma geral temos,

$$y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \ldots + \beta_{p-1} x_{i,p-1} + \varepsilon_i \Rightarrow \begin{cases} y_1 = \beta_0 + \beta_1 x_{1,1} + \ldots + \beta_{p-1} x_{1,p-1} + \varepsilon_1 \\ y_2 = \beta_0 + \beta_1 x_{2,1} + \ldots + \beta_{p-1} x_{2,p-1} + \varepsilon_2 \\ y_3 = \beta_0 + \beta_1 x_{3,1} + \ldots + \beta_{p-1} x_{3,p-1} + \varepsilon_3 \\ \vdots \\ y_n = \beta_0 + \beta_1 x_{n,1} + \ldots + \beta_{p-1} x_{n,p-1} + \varepsilon_n \end{cases}$$

Então podemos definir

$$\underline{y} = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{pmatrix}_{n\times 1}, \; X = \begin{pmatrix} 1 & x_{1,1} & x_{1,2} & \ldots & x_{1,p-1} \\ 1 & x_{2,1} & x_{2,2} & \ldots & x_{2,p-1} \\ 1 & x_{3,1} & x_{3,2} & \ldots & x_{3,p-1} \\ \vdots & \vdots & \vdots & \ldots & \vdots \\ 1 & x_{n,1} & x_{n,2} & \ldots & x_{n,p-1} \end{pmatrix}_{n\times p}, \; \underline{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{p-1} \end{pmatrix}_{p\times 1} \text{ e } \underline{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \vdots \\ \varepsilon_n \end{pmatrix}_{n\times 1}$$

E escrever

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{1,1} & x_{1,2} & \dots & x_{1,p-1} \\ 1 & x_{2,1} & x_{2,2} & \dots & x_{2,p-1} \\ 1 & x_{3,1} & x_{3,2} & \dots & x_{3,p-1} \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 1 & x_{n,1} & x_{n,2} & \dots & x_{n,p-1} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{p-1} \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

Ou seja, o modelo gral também pode ser definido por:

$$\underline{y} = X\underline{\beta} + \underline{\varepsilon}.$$

Antes de apresentar a nova definição para o Modelo de Regressão Linear, agora na forma matricial, veja que o vetor $\underline{\varepsilon}$ pode ser definido como combinação linear de variáveis aleatórias i.i.d. com distribuição $N(0,1)$, isto é,

$$\underline{\varepsilon} = \sigma^2 \underline{Z}$$

onde $\underline{Z}$ é um vetor de n variáveis aleatórias i.i.d. com distribuição $N(0,1)$. Ou seja, $\underline{\varepsilon}$ é um vetor aleatório com distribuição Normal n-variada com vetor de médias $\underline{0}$ e matriz de covariância $\sigma^2 I$, isto é, $\underline{\varepsilon} \sim N_n\left(\underline{\mu} = \underline{0}, \Sigma = \sigma^2 I\right)$. Para mais detalhes veja Apêndice **??**.

Definição 2.2.1 *O Modelo de Regressão Linear Múltiplo define a seguinte relação entre a variável resposta y e as $p-1$ variáveis independentes x_1, x_2, ..., x_{p-1}:*

$$\underline{y} = X\underline{\beta} + \underline{\varepsilon}, \quad \underline{\varepsilon} \sim N_n\left(\underline{\mu} = \underline{0}, \Sigma = \sigma^2 I\right)$$

onde os vetores $\underline{y}$, $\underline{\beta}$, $\underline{\varepsilon}$ e a matriz X foram definidas acima.

Algumas observações:

- n = número de observações, ou seja, o tamanho da amostra;
- p = número de parâmetros do tipod β;
- X = matriz $n \times p$ de valores constantes (vindos da amostra);
- $\underline{\beta}$ = vetor de dimensão p de parâmetros desconhecidos;
- $\underline{y}$ = vetor aleatório de dimensão n;
- $\underline{\varepsilon}$ = vetor aleatório de dimensão n;

2.3 Estimação dos Coeficientes β's no Modelo Múltiplo

Assim como no modelo simples estamos interessados em encontrar os estimadores para os parâmetros desconhecidos do modelo múltiplo. Com eles sera possível encontrar estimativas para os parâmetros e assim a função de regressão estimada.

2.3.1 Estimadores por Mínimos Quadrados

O estimador para $\underline{\beta}$ por mínimos quadrados é aquele que minimiza a soma dos quadrados dos erros ε_i.

A soma dos quadrados dos erros pode ser definida por:

$$Q = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i,1} - \beta_2 x_{i,2} - \ldots - \beta_{p-1} x_{i,p-1})^2$$

Para facilitar as contas vamos escrever Q na notação matricial.

$$\begin{aligned} Q &= \sum_{i=1}^{n} \varepsilon_i^2 = \underline{\varepsilon}^T \underline{\varepsilon} = \left(\underline{y} - X\underline{\beta}\right)^T \left(\underline{y} - X\underline{\beta}\right) = \left(\underline{y}^T - \underline{\beta}^T X^T\right) \left(\underline{y} - X\underline{\beta}\right) \\ &= \underline{y}^T \underline{y} - \underline{y}^T X \underline{\beta} - \underline{\beta}^T X^T \underline{y} + \underline{\beta}^T X^T X \underline{\beta} = \underline{y}^T \underline{y} - 2\underline{\beta}^T X^T \underline{y} + \underline{\beta}^T X^T X \underline{\beta} \end{aligned}$$

Veja que $\underline{\beta}^T X^T \underline{y}$ é um escalar, isto é, tem dimensão 1×1 e por isso é igual ao seu transposto. Assim concluímos que $\underline{\beta}^T X^T \underline{y} = \underline{y}^T X \underline{\beta}$.

E assim chegamos na expressão:

$$Q(\underline{\beta}) = \underline{y}^T \underline{y} - 2\underline{\beta}^T X^T \underline{y} + \underline{\beta}^T X^T X \underline{\beta}.$$

Se queremos o estimador para $\underline{\beta}$ por mínimos quadrados precisamo encontrar os pontos de mínimo de Q. Para isso vamos derivar Q em relação a $\underline{\beta}$ e igualar a zero. Para facilitar as contas faremos a derivação na notação matricial, para mais detalhes veja Apêndice **??**. A solução será o estimador por mínimos quadrados.

$$\frac{dQ}{d\underline{\beta}} = -2X^T \underline{y} + 2X^T X \underline{\beta} = 0$$

Assim chegamos na solução:

$$\hat{\underline{\beta}} = \left(X^T X\right)^{-1} X^T \underline{y} \tag{2.2}$$

2.3.2 Estimadores por Máxima Verossimilhança

O estimador encontrado a partir do método dos mínimos quadrados (Equação 2.2) coincide com o estimador por máxima verossimilhança. Veremos isso nessa seção.

A função de máxima verossimilhança é definida por:

$$L(\underline{\beta}, \sigma^2) = \prod_{i=1}^{n} f_{y_i}(y_i | \beta_0, \beta_1, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{i,1} - \ldots - \beta_{p-1} x_{i,p-1})} = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}\underline{\varepsilon}^T \underline{\varepsilon}}$$

Pensamos em L somente como função de $\underline{\beta}$ podemos ver que maximizar L é o mesmo que minimizar $\underline{\varepsilon}^T \underline{\varepsilon}$, e assim recaímos no problema de mínimos quadrados.

2.4 Valores Ajustados, Resíduos e a Matriz Hat

Assim como na regressão simples, na regressão múltipla o valor ajustado para a i-ésima observação é definido por:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i,1} + \hat{\beta}_2 x_{i,2} + \ldots + \hat{\beta}_{p-1} x_{i,p-1} \ .$$

E da mesma forma que na regressão simples definimos o i-ésimo resíduo na regressão múltipla por:

$$e_i = y_i - \hat{y}_i \ .$$

Usando a notação matricial podemos definir um vetor com todos os n valores ajustados, $\underline{\hat{y}}$, e o vetor com os n resíduos, $\underline{e}$. Veja que tais vetores podem ser escritos por:

$$\underline{\hat{y}} = X\hat{\beta} \quad \text{e} \quad \underline{e} = \underline{y} - \underline{\hat{y}} = \underline{y} - X\hat{\beta} \ .$$

De acordo com a Equação 2.2, $\underline{\hat{\beta}} = \left(X^T X\right)^{-1} X^T \underline{y}$, logo podemos escrever

$$\begin{aligned} \underline{\hat{y}} &= X\left(X^T X\right)^{-1} X^T \underline{y} = H\underline{y} \\ \underline{e} &= \underline{y} - X\left(X^T X\right)^{-1} X^T \underline{y} = \left(I - X\left(X^T X\right)^{-1} X^T\right)\underline{y} = (I-H)\underline{y} \end{aligned}$$

onde $H = X\left(X^T X\right)^{-1} X^T$ é chamada de matriz *Hat*. A Proposição 2.4.1 a seguir apresenta algumas propriedades da matriz *Hat*.

Proposição 2.4.1 *Seja* $H = X\left(X^T X\right)^{-1} X^T$. *Então:*

(i) H *é uma matriz quadrada* $n \times n$.

(ii) H *é uma matriz simétrica.*

(iii) H *é uma matriz idempotente, isto é,* $HH = H$.

(iv) $tr(H) = p$.

Demonstração:

(i) Para demonstrar esse item basta lembrar que X é uma matriz $n \times p$, logo $X^T X$ tem dimensão $p \times p$ e assim a dimensão de $X\left(X^T X\right)^{-1} X^T$ é $n \times n$.
(ii) Para mostrar que uma matriz é simétrica basta mostrar que $H = H^T$. Vamos então encontrar H^T. $H^T = \left(X\left(X^T X\right)^{-1} X^T\right)^T = X\left(X^T X\right)^{-1} X^T$.
(iii) $HH = X\left(X^T X\right)^{-1} X^T X\left(X^T X\right)^{-1} X^T = X\left(X^T X\right)^{-1} X^T = H$. Logo, H é uma matriz idempotente.
(iv) Antes da demonstração desse último item vamos recordar algumas propriedades do traço de uma matriz.

- $tr(M)$ = soma dos elementos da diagonal principal de M.
- Se M é uma matriz quadrada, então $tr(M) = tr(M^T)$.

- Sejam A e B duas matrizes de mesma dimensão e $\alpha \in \mathbb{R}$ um escalar. Então $tr(\alpha A + B) = \alpha tr(A) + tr(B)$.
- Sejam A e B duas matrizes de dimensões $n \times m$ e $m \times n$, ou seja, existe AB e BA. Então, $tr(AB) = tr(BA)$.

Vamos agora encontrar traço da matriz H.

$tr(H) = tr(X\left(X^TX\right)^{-1}X^T) = tr(X^TX\left(X^TX\right)^{-1})$, pois $tr(AB) = tr(BA)$. Continuando, $tr(H) = tr(X^TX\left(X^TX\right)^{-1}) = tr(I_p) = p$. Logo, $tr(H) = p$.

□

2.5 Distribuição Amostral de $\hat{\beta}$

Nessa seção vamos demonstrar a Proposição 2.5.1 a seguir, que define a distribuição amostral do estimador $\underline{\hat{\beta}}$.

Proposição 2.5.1 *Considere o Modelo de Regressão Linear Múltiplo e $\underline{\hat{\beta}} = (X^TX)^{-1}X^T\underline{y}$ o estimador de mínimos quadrados para o vetor de parâmetros β. Então,*

(i) $\underline{\hat{\beta}}$ é vetor aleatório com distribuição Normal p-variada.

(iii) $E[\underline{\hat{\beta}}] = \underline{\beta}$

(iv) $Var\left(\underline{\hat{\beta}}\right) = \sigma^2(X^TX)^{-1}$

Ou seja, $\underline{\hat{\beta}} \sim N_p\left(\underline{\mu} = \underline{\beta}, \Sigma = \sigma^2(X^TX)^{-1}\right)$.

Demonstração:

(i) Já vimos que $\underline{\varepsilon} \sim N_n\left(\underline{\mu} = \underline{0}, \Sigma = \sigma^2 I\right)$ (Definição 2.2.1). Como $\underline{y} = X\underline{\beta} + \underline{\varepsilon}$ e combinação linear de normal multivariada também tem distribuição normal multivariada (Apêndice **??**), então $\underline{y} \sim N_n\left(\underline{\mu} = X\beta, \Sigma = \sigma^2 I\right)$. Para terminar, $\underline{\hat{\beta}} = (X^TX)^{-1}X^T\underline{y}$, ou seja, $\underline{\hat{\beta}}$ é combinação linear de $\underline{y}$, que é normal multivariada. Logo, $\underline{\hat{\beta}}$ também é normal multivariada. Como $\underline{\hat{\beta}}$ é vetor aleatório com p variáveis aleatórias, $\underline{\hat{\beta}}$ é normal p-multivariada. □

(ii) Antes de fazermos as contas para encontrar $\mathrm{E}[\underline{\hat{\beta}}]$ veja que podemos escrever

$$\begin{aligned}
\underline{\hat{\beta}} &= \left(X^TX\right)^{-1}X^T\underline{y} \\
&= \left(X^TX\right)^{-1}X^T\left(X\underline{\beta} + \underline{\varepsilon}\right) \\
&= \left(X^TX\right)^{-1}X^TX\underline{\beta} + \left(X^TX\right)^{-1}X^T\underline{\varepsilon} \\
&= \underline{\beta} + \left(X^TX\right)^{-1}X^T\underline{\varepsilon}
\end{aligned}$$

Logo, $\mathrm{E}\left[\underline{\hat{\beta}}\right] = \mathrm{E}\left[\underline{\beta} + \left(X^TX\right)^{-1}X^T\underline{\varepsilon}\right] = \underline{\beta} + \left(X^TX\right)^{-1}X^T\,\mathrm{E}\left[\underline{\varepsilon}\right] = \underline{\beta}$, pois $\mathrm{E}\left[\underline{\varepsilon}\right] = \underline{0}$. Assim concluímos que $\underline{\hat{\beta}}$ é um estimador não-tendencioso para o vetor de parâmetros $\underline{\beta}$.

□

(iii) Vamos agora ao cálculo da variância de $\hat{\underline{\beta}}$. Por definição, $\text{Var}\left(\hat{\underline{\beta}}\right) = \text{E}\left[\left(\hat{\beta}-\beta\right)\left(\hat{\beta}-\beta\right)^T\right]$. Veja que $\hat{\beta}-\beta = \left(X^TX\right)^{-1}X^T\underline{\varepsilon}$, logo

$$\begin{aligned}
\text{Var}\left(\hat{\underline{\beta}}\right) &= \text{E}\left[\left(\hat{\beta}-\beta\right)\left(\hat{\beta}-\beta\right)^T\right] = \text{E}\left[\left(\left(X^TX\right)^{-1}X^T\underline{\varepsilon}\right)\left(\left(X^TX\right)^{-1}X^T\underline{\varepsilon}\right)^T\right] \\
&= \text{E}\left[\left(X^TX\right)^{-1}X^T\underline{\varepsilon}\underline{\varepsilon}^TX\left(X^TX\right)^{-1}\right] = \left(X^TX\right)^{-1}X^T\,\text{E}\left[\underline{\varepsilon}\underline{\varepsilon}^T\right]X\left(X^TX\right)^{-1} \\
&= \left(X^TX\right)^{-1}X^T\sigma^2IX\left(X^TX\right)^{-1} = \sigma^2\left(X^TX\right)^{-1}X^TX\left(X^TX\right)^{-1} \\
&= \sigma^2\left(X^TX\right)^{-1}
\end{aligned}$$

Finalmente chegamos em $\text{Var}\left(\hat{\underline{\beta}}\right) = \sigma^2\left(X^TX\right)^{-1}$.

□

2.6 Estimador para σ^2

Nessa seção queremos mostrar que $MSE = \frac{\sum_{i=1}^n e_i}{n-p}$ é um estimado não tendencioso para σ^2. Para isso vamos mostrar que $\text{E}\left[\sum_{i=1}^n e_i\right] = \text{E}\left[\underline{e}^T\underline{e}\right] = \sigma^2(n-p)$. O primeiro passo será mostrar que o vetor de resíduos $\underline{e}$ pode ser escrito como combinação linear de $\underline{\varepsilon}$.

$$\begin{aligned}
\underline{e} &= \underline{y} - X\hat{\beta} = X\beta + \underline{\varepsilon} - X\left(\beta + \left(X^TX\right)^{-1}X^T\underline{\varepsilon}\right) \\
&= X\beta + \underline{\varepsilon} - X\beta - X\left(X^TX\right)^{-1}X^T\underline{\varepsilon} = \underline{\varepsilon} - X\left(X^TX\right)^{-1}X^T\underline{\varepsilon} \\
&= (I-H)\underline{\varepsilon}
\end{aligned}$$

Assim concluímos que $\underline{e} = (I-H)\underline{\varepsilon}$, ou seja, o vetor de resíduos é combinação linear dos erros. Veja algumas propriedades da matriz $(I+H)$.

Proposição 2.6.1 *Seja H uma matriz $n \times n$ simétrica e idempotente e I a matriz identidade $n \times n$. Então $I-H$ é uma matriz $n \times n$ simétrica e idempotente.*

Demonstração:

Para mostrar que $I-H$ é simétrica basta mostrar que $I-H = (I-H)^T$. Vamos às contas. $(I-H)^T = I^T - H^T = I - H$, ou seja, $I-H$ é simétrica.

Vamos agora ver que $I-H$ é idempotente. $(I-H)(I-H) = II - IH - HI + HH = I - H - H + H = I - H$.

□

Mas o que queremos é encontrar $\text{E}\left[\underline{e}^T\underline{e}\right]$. Vamos fazer essa conta substituindo e por $(I-H)\varepsilon$ e usar as propriedades de $I-H$ demonstradas na Proposição 2.6.1.

$$\begin{aligned}
\text{E}\left[\underline{e}^T\underline{e}\right] &= \text{E}\left[\left((I-H)\varepsilon\right)^T\left((I-H)\varepsilon\right)\right] \\
&= \text{E}\left[\varepsilon^T(I-H)^T(I-H)\varepsilon\right] \\
&= \text{E}\left[\varepsilon^T(I-H)(I-H)\varepsilon\right] \\
&= \text{E}\left[\varepsilon^T(I-H)\varepsilon\right]
\end{aligned}$$

Para seguirmos com as contas veja que $\varepsilon^T(I-H)\varepsilon$ tem dimensão 1×1, logo podemos dizer que $\varepsilon^T(I-H)\varepsilon = tr(\varepsilon^T(I-H)\varepsilon)$. Além disso, usando as propriedades do traço, podemos afirmar que $tr(\varepsilon^T(I-H)\varepsilon) = tr((I-H)\varepsilon\varepsilon^T)$. E para terminar as observações e voltamos as contas, como o traço é um operador linear podemos comutar ele com o valor esperado, ou seja, $\mathrm{E}[tr(.)] = tr(\mathrm{E}[.])$.

$$\begin{aligned}
\mathrm{E}\left[\underline{e}^T\underline{e}\right] &= \mathrm{E}\left[\varepsilon^T(I-H)\varepsilon\right] = \mathrm{E}\left[tr\left(\varepsilon^T(I-H)\varepsilon\right)\right] \\
&= \mathrm{E}\left[tr\left((I-H)\varepsilon\varepsilon^T\right)\right] = tr\left(\mathrm{E}\left[(I-H)\varepsilon\varepsilon^T\right]\right) \\
&= tr\left((I-H)\,\mathrm{E}\left[\varepsilon\varepsilon^T\right]\right) = tr\left((I-H)\sigma^2 I\right) \\
&= \sigma^2 tr\,(I-H) = \sigma^2\left(tr\,(I) - tr\,(H)\right) \\
&= \sigma^2\,(n-p)
\end{aligned}$$

Então,

$$MSE = \frac{\sum_{i=1}^{n} e_i}{n-p} = \frac{\underline{e}^T\underline{e}}{n-p}$$

é um estimador não tendencioso para σ^2.

Para terminar essa seção vamos apresentar a generalização do teorema visto para o modelo simples.

Teorema 2.6.2 *Suponha o Modelo de Regressão Linear. Então,*

(i) $(n-p)MSE/\sigma^2 \sim \chi^2_{n-p}$

(ii) MSE *é v.a. independente de* $\hat{\beta}_k$.

A demonstração desse teorema também sera omitida.

2.7 Inferências para cada β_k

Já vimos que $\underline{\hat{\beta}}$ é um vetor aleatório com distribuição normal p-variada com vetor de média $\underline{\beta}$ e matriz de covariância $\sigma^2(X^TX)^{-1}$, ou seja, $\underline{\hat{\beta}} \sim N_p(\underline{\mu} = \underline{\beta}, \Sigma = \sigma^2(X^TX)^{-1})$. Ou seja, cada $\hat{\beta}_k \sim N(\beta_k, \sigma^2 C_{k+1,k+1})$, onde C é a matriz $(X^TX)^{-1}$ e $C_{k+1,k+1}$ o valor na sua posição $(k+1, k+1)$.

Vimos também que MSE é um estimador não tendencioso para σ^2 e que $MSE(n-p)/\sigma^2 \sim \chi^2_{n-p}$. O que queremos nessa seção é usar essas informações para construir intervalos de confiança e testes de hipóteses para os parâmetros β_k.

2.7.1 Intervalo de Confiança para cada β_k

Para encontrar intervalos de confiança para β_k precisamos primeiro de uma quantidade pivotal para esse parâmetros, que será definida a partir do resultado da Proposição 2.7.1 a seguir.

Proposição 2.7.1

$$\frac{\hat{\beta}_k - \beta_k}{\sqrt{MSE\ C_{k+1,k+1}}} \sim t_{n-p}$$

Demonstração:

Esta demonstração é igual a muitas outras que já foram feitas ou deixadas como exercício.

Sabemos que $\hat{\beta}_k \sim N(\beta_k, \sigma^2 C_{k+1,k+1})$, logo, a partir da padronização da normal podemos afirmar que $Z = (\hat{\beta}_k - \beta_k)/\sqrt{\sigma^2 C_{k+1,k+1}} \sim N(0,1)$.

Também sabemos que $Q = MSE(n-p)/\sigma^2 \sim \chi^2_{n-p}$.

Como $\frac{Z}{\sqrt{Q/(n-p)}} \sim t_{n-p}$, pela definição da variável aleatória t, podemos afirmar que:

$$\frac{Z}{\sqrt{Q/(n-p)}} \sim t_{n-p} \Rightarrow \frac{(\hat{\beta}_k - \beta_k)/\sqrt{\sigma^2 C_{k+1,k+1}}}{\sqrt{\left(MSE(n-p)/\sigma^2\right)/(n-p)}} \sim t_{n-p} \Rightarrow$$

$$\frac{\hat{\beta}_k - \beta_k}{\sqrt{\sigma^2 C_{k+1,k+1}}\sqrt{MSE(n-p)/\left(\sigma^2(n-p)\right)}} \sim t_{n-p} \Rightarrow \frac{\hat{\beta}_k - \beta_k}{\sqrt{\sigma^2 C_{k+1,k+1} MSE/\sigma^2}} \sim t_{n-p}$$

$$\frac{\hat{\beta}_k - \beta_k}{\sqrt{\sigma^2 C_{k+1,k+1}}\sqrt{MSE(n-p)/\left(\sigma^2(n-p)\right)}} \sim t_{n-p} \Rightarrow \frac{\hat{\beta}_k - \beta_k}{\sqrt{\sigma^2 C_{k+1,k+1} MSE/\sigma^2}} \sim t_{n-p}$$

Logo,

$$\frac{\hat{\beta}_k - \beta_k}{\sqrt{C_{k+1,k+1} MSE}} \sim t_{n-p} \ .$$

□

A Proposição 2.7.1 acima nos dá uma quantidade pivotal para o parâmetro β_k. Com ela podemos construir um intervalo de confiança para esse parâmetro, que está definido na Equação 2.3 as seguir.

$$\hat{\beta}_k \pm t_{1-\frac{\alpha}{2},n-p}\sqrt{MSE\ C_{k+1,k+1}} \tag{2.3}$$

2.7.2 Teste de Hipótese para cada β_k

Se quisermos testar as hipóteses

$$H_0 : \beta_k = 0 \quad H_1 : \beta_k \neq 0$$

podemos usar a estatística de teste

$$t^* = \frac{\hat{\beta}_k}{\sqrt{C_{k+1,k+1} MSE}}$$

que sob H_0 tem distribuição t_{n-p}. Com essa estatística de teste a regra de decisão será:

- Se $|t^*| > t_{1-\frac{\alpha}{2},n-p}$ rejeitamos H_0, ou seja, aceitamos $\beta_k \neq 0$.
- Se $|t^*| < t_{1-\frac{\alpha}{2},n-p}$ aceitamos H_0, ou seja, aceitamos $\beta_k = 0$.

Semelhante ao que foi feito nas seções 1.4.3 e 1.5.3 podemos também criar estatísticas de testes e regras de decisão para verificar as hipóteses: $H_0 : \beta_k \leq 0$ contra $H_1 : \beta_k > 0$ ou $H_0 : \beta_k = \beta^*$ contra $H_1 : \beta_k \neq \beta^*$ ou $H_0 : \beta_k \leq \beta^*$ contra $H_1 : \beta_k > \beta^*$.

2.8 Intervalo de Confiança para a Média da Variável Resposta

Nessa seção queremos construir um intervalo de confiança para a média da variável resposta dado um certo nível $\underline{x_h}$. Na regressão múltipla um nível h era definido simplesmente pelo valor da única variável independente do modelo, já na regressão múltipla o nível h será definido pelo valor de todas as $p-1$ variáveis independentes do modelo. Dessa forma vamos representar o nível h pelo vetor de valores $\underline{x_h}^T = (1, x_{h,1}, x_{h,2}, \ldots, x_{h,p-1})$. Veja que $\underline{x_h}^T$ é semelhante às linhas da matriz X.

Então, se queremos encontrar um intervalo de confiança para a média da variável resposta em um certo nível $\underline{x_h}$, queremos um intervalo de confiança para o parâmetro

$$\mathrm{E}[y_h] = \mu_h = \underline{x_h}^T \underline{\beta}.$$

A quantidade pivotal para esse parâmetro será definida a partir da distribuição amostral do valor da variável resposta ajustada ao nível $\underline{x_h}$.

Proposição 2.8.1 *Considere o Modelo de Regressão Linear Múltiplo, um certo nível* $\underline{x_h}^T = (1, x_{h,1}, x_{h,2}, \ldots, x_{h,p-1})$ *e o valor ajustado* $\hat{y}_h = \hat{y}|x_h = \underline{x_h}^T \underline{\hat{\beta}}$. *Então,*

(i) $\hat{y}_h$ *é variável aleatória Normal.*

(iii) $E[\hat{y}_h] = \mu_h = \underline{x_h}^T \underline{\beta}$.

(iv) $Var\,(\hat{y}_h) = \sigma^2 \underline{x_h}^T (X^T X)^{-1} \underline{x_h}$

Demonstração:

(i) Veja que $\hat{y}_h$ é combinação linear dos estimadores $\hat{\beta}_k$. Já vimos que cada $\hat{\beta}_k$ é variável aleatória normal, logo $\hat{y}_h$ também é variável aleatória normal.

(ii) $\mathrm{E}[\hat{y}_h] = \mathrm{E}[\underline{x_h}^T \underline{\hat{\beta}}] = \underline{x_h}^T \mathrm{E}[\underline{\hat{\beta}}] = \underline{x_h}^T \underline{\beta}$.

(iii)

$$\begin{aligned}
\mathrm{Var}\,(\hat{y}_h) &= \mathrm{Var}\left(\underline{x_h}^T \underline{\hat{\beta}}\right) = \mathrm{E}\left[\left(\underline{x_h}^T \underline{\hat{\beta}} - \underline{x_h}^T \underline{\beta}\right)\left(\underline{x_h}^T \underline{\hat{\beta}} - \underline{x_h}^T \underline{\beta}\right)^T\right] \\
&= \mathrm{E}\left[\left(\underline{x_h}^T \left(\underline{\hat{\beta}} - \underline{\beta}\right)\right)\left(\underline{x_h}^T \left(\underline{\hat{\beta}} - \underline{\beta}\right)\right)^T\right] = \mathrm{E}\left[\underline{x_h}^T \left(\underline{\hat{\beta}} - \underline{\beta}\right)\left(\underline{\hat{\beta}} - \underline{\beta}\right)^T \underline{x_h}\right] \\
&= \underline{x_h}^T \mathrm{E}\left[\left(\underline{\hat{\beta}} - \underline{\beta}\right)\left(\underline{\hat{\beta}} - \underline{\beta}\right)^T\right] \underline{x_h} = \underline{x_h}^T \mathrm{Var}\left(\underline{\hat{\beta}}\right) \underline{x_h} = \underline{x_h}^T \sigma^2 (X^T X)^{-1} \underline{x_h} \\
&= \sigma^2 \underline{x_h}^T (X^T X)^{-1} \underline{x_h}
\end{aligned}$$

□

Com isso concluímos que $\hat{y}_h \sim N(\mu_h, \sigma^2 \underline{x_h}^T (X^T X)^{-1} \underline{x_h})$. Já sabemos que $MSE(n-p)/\sigma^2 \sim \chi^2_{n-p}$ (Teorema 2.6.2). Podemos então demonstrar a Proposição 2.8.2. Na verdade deixo sua demonstração como exercício, uma vez que já fizemos muitas semelhantes.

Proposição 2.8.2

$$\frac{\hat{y}_h - \mu_h}{\sqrt{MSE\underline{x_h}^T(X^TX)^{-1}\underline{x_h}}} \sim t_{n-p}$$

A Proposição 2.8.2 acima nos dá uma quantidade pivotal para o parâmetro μ_h. Com ela podemos construir um intervalo de confiança com confiabilidade de $1-\alpha$ para esse parâmetro, que está definido na Equação 2.4 as seguir.

$$\hat{y}_h \pm t_{1-\frac{\alpha}{2},n-p}\sqrt{MSE\underline{x_h}^T(X^TX)^{-1}\underline{x_h}} \tag{2.4}$$

2.9 Intervalo de Predição para uma Nova Observação

Nessa seção queremos construir um intervalo de predição para a variável resposta dado um certo nível $\underline{x_h}$. Ou seja, queremos encontrar um intervalo de confiança para y_h dado um certo nível $\underline{x_h}$. As contas feitas aqui serão semelhantes aquelas feitas no caso da Regressão Simples.

Seja $\underline{x_h}$ um nível qualquer. A variável resposta para esse nível é definida por:

$$y_h = \underline{x_h}^T\underline{\beta} + \varepsilon_h \sim N(\underline{x_h}^T\underline{\beta}, \sigma^2)$$

e o erro ε_h é independente de todos os erros ε_i, $1 \leq i \leq n$.

Já o valor ajustado da variável resposta no nível $\underline{x_h}$ é definida por:

$$\hat{y}_h = \underline{x_h}^T\underline{\hat{\beta}} \sim N(\underline{x_h}^T\underline{\beta}, \sigma^2\underline{x_h}^T(X^TX)^{-1}\underline{x_h})$$

como mostrado na Proposição 2.8.1.

Para construir o intervalo de predição vamos usar a variável aleatória $y_h - \hat{y}_h$. Por isso é importante conhecer a sua distribuição, como mostra a Proposição 2.9.1 a seguir.

Proposição 2.9.1 *Considere o Modelo de Regressão Linear Múltiplo e um certo nível* $\underline{x_h}^T = (1, x_{h,1}, x_{h,2}, \ldots, x_{h,p-1})$. *Então,*

(i) $y_h - \hat{y}_h$ *é variável aleatória Normal.*

(iii) $E[y_h - \hat{y}_h] = 0$.

(iv) $Var\,(y_h - \hat{y}_h) = \sigma^2\left(1 + \underline{x_h}^T(X^TX)^{-1}\underline{x_h}\right)$

Demonstração:

(i) Como y_h e $\hat{y}_h$ são variável aleatória normais e qualquer transformação linear de 'variáveis aleatórias normais também é uma variável aleatória normal, podemos afirmar que $y_h - \hat{y}_h$ é variável aleatória Normal.

(ii) $E[y_h - \hat{y}_h] = \mathrm{E}[y_h] - \mathrm{E}[\hat{y}_h] = \underline{x_h}^T\beta - \underline{x_h}^T\beta = 0$.

(iii) Na demonstração da Proposição 2.5.1 vimos que $\underline{\hat{\beta}} = \underline{\beta} + \left(X^TX\right)^{-1}X^T\underline{\varepsilon}$, ou seja, $\underline{\hat{\beta}}$ pode ser escrito como função somente das variáveis aleatórias ε_i, $1 \leq i \leq n$. Consequentemente $\hat{y}_h$ também pode ser escrito como função somente das variáveis aleatórias

ε_i, $1 \leq i \leq n$. Como y_h é função somente da variável aleatória ε_h e esta é independente de cada ε_i podemos concluir que $\hat{y}_h$ e y_h são variáveis aleatórias independentes. Logo,

$$\text{Var}(y_h - \hat{y}_h) = \text{Var}(y_h) + \text{Var}(\hat{y}_h) = \sigma^2 + \sigma^2 \underline{x_h}^T (X^T X)^{-1} \underline{x_h} = \sigma^2 (1 + \underline{x_h}^T (X^T X)^{-1} \underline{x_h})$$

□

Com isso concluímos que $y_h - \hat{y}_h \sim N(0, \sigma^2(1 + \underline{x_h}^T (X^T X)^{-1} \underline{x_h}))$. Já sabemos que $MSE(n-p)/\sigma^2 \sim \chi^2_{n-p}$ (Teorema 2.6.2). Podemos então demonstrar a Proposição 2.9.2. Na verdade deixo sua demonstração como exercício, uma vez que já fizemos muitas semelhantes.

Proposição 2.9.2

$$\frac{y_h - \hat{y}_h}{\sqrt{MSE(1 + \underline{x_h}^T (X^T X)^{-1} \underline{x_h})}} \sim t_{n-p}$$

A Proposição 2.9.2 acima nos possibilita construir um intervalo de confiança com confiabilidade de $1 - \alpha$ para y_h, ou seja, um intervalo de predição. Este intervalo está definido na Equação 2.5 as seguir.

$$\hat{y}_h \pm t_{1-\frac{\alpha}{2}, n-p} \sqrt{MSE(1 + \underline{x_h}^T (X^T X)^{-1} \underline{x_h})} \tag{2.5}$$

2.10 Extrapolação na Regressão Múltipla

Ao realizar uma previsão considerando um nível que não está na amostra original temos que tomar um cuidado extra. Esse alerta serve não só para os intervalos de predição como também para os intervalos de confiança para a média da variável resposta, sempre que o nível $\underline{x_h}$ analisado não fizer parte da amostra original. Acontece que se o nível $\underline{x_h}$ for muito diferente dos valores observados $\underline{x_i}$ pode ser que essa previsão seja muito ruim. Isso acontece pois o modelo procura se ajustar bem aos dados originais e ele pode funcionar muito mal para observações muito diferente.

Por exemplo considere o exercício 5 das Aulas Práticas do Capítulo 1. Se estamos interessados em analisar a diminuição da massa muscular em função da idade e para isso foram coletados dados de mulheres entre 40 e 79 anos, o modelo ajustado deve funcionar bem para prever a massa muscular de uma mulher com 50 anos ou a massa muscular média de mulheres com 50 anos. Mas não podemos usar o modelo ajustado para prever a massa muscular de um homem com 50 anos ou de uma mulher com 20 anos. Para realizar essas previsões seria preciso levantar uma nova amostra com homens e mulheres com idades variando entre 10 e 80 anos e a partir dessa nova amostra ajustar um novo modelo de regressão linear.

Mas como saber o quanto próximo $\underline{x_h}$ está de $\underline{x_i}$? Para isso vamos usar a medida h. Para os pontos da amostra a medida h da observação de nível $\underline{x_i}$ é exatamente o valor da posição $h_{i,i}$ da Matriz Hat. Esse valor é função da distância Euclidiana do ponto $\underline{x_i}$ ao centroide de de todos os pontos da amostra e quanto maior o seu valor mais distante o ponto $\underline{x_i}$ está do centroide da amostra. Então pontos da amostra muito próximos do centroide terão $h_{i,i}$ pequeno e os pontos mais distantes terão $h_{i,i}$ grande. Se definirmos $h_{max} = \max_{1 \leq i \leq n} h_{i,i}$ o ponto na amostra $\underline{x_k}$ tal que $h_{k,k} = h_{max}$ é o mais distante do centroide.

Para uma nova observação a medida h para um certo nível $\underline{x_h}$ é definida por:

$$\underline{x_h}^T (X^T X)^{-1} \underline{x_h} \tag{2.6}$$

Veja que se $\underline{x_h} = \underline{x_i}$ temos exatamente o valor de $h_{i,i}$.

Vamos usar o seguinte critério para definir se o modelo ajustado pode ser usado para prever um novo nível: se a medida h desse novo nível for menor que h_{max}, isto é,

$$\underline{x_h}^T (X^T X)^{-1} \underline{x_h} < h_{max},$$

vamos considerar a previsão para esse novo nível. Caso contrário a previsão não será considerada.

2.11 Análise dos Resíduos na Regressão Múltipla

Assim como no Modelo Simples, no Modelo Múltiplo a análise dos resíduos será feita considerando os resíduos padronizados, definidos por:

$$e_i^* = \frac{e_i}{\sqrt{MSE}} = \frac{y_i - \hat{y}_i}{\sqrt{MSE}} \tag{2.7}$$

Ou, na notação matricial,

$$\underline{e}^* = \frac{\underline{e}}{\sqrt{MSE}} = \frac{\underline{y} - \underline{\hat{y}}}{\sqrt{MSE}} \tag{2.8}$$

Vejamos como serão os diagnósticos no caso da Regressão Múltipla.

Não-Linearidade

O diagnóstico de não-linearidade na regressão linear múltipla será feito em relação a cada variável preditiva x_k. Então se queremos verificar a suposição de que a média de y varia de forma linear com x_k devemos plotar em um gráfico os valores de $x_{i,k}$ versus os resíduos padronizados (e_i^*). O padrão esperado para esse gráfico (quando há linearidade) é o mesmo da análise de resíduos no modelo simples, veja Figura 1.5. Se o gráfico não tiver o padrão esperado, isto é, ele se parece com a Figura 1.7, será diagnosticado a não-linearidade de E[y] com a variável x_k.

Também é possível detectar a não-linearidade a partir do gráfico dos valores ajustados $\hat{y}_i$ versus os resíduos padronizados (e_i^*). O único problema de usar esse gráfico para diagnosticar a não-linearidade é que não saberemos ao certo qual a variável x_k que não tem relação linear com y.

Heterocedasticidade

O diagnóstico de heterocedasticidade na regressão linear múltipla será feito da mesma forma que na regressão simples, a única diferença é que aqui temos que fazer o gráfico dos resíduos padronizados (ou do seu módulo, ou do seu quadrado) sempre versus $\hat{y}_i$. O padrão esperado para esse gráfico (quando não há heterocedasticidade) é o mesmo da análise de resíduos no modelo simples, veja Figura 1.5. Se o gráfico não tiver o padrão esperado, isto é, ele se parece com a Figura 1.9, a heterocedasticidade será diagnosticada.

Não-Normalidade

A não-normalidade, assim como no caso da Regressão Simples, será diagnosticada a partir do gráfico `qqplot` dos resíduos padronizados. Seu padrão esperado é que estejam em torno da reta identidade, veja Figura 1.11. Se o padrão esperado não for encontrado, isto é, o gráfico se parece com um dos gráficos da Figura 1.12, a não-normalidade dos erros será diagnosticada.

Coeficiente de Determinação

O Coeficiente de Determinação é definido na Regressão Múltipla da mesma forma que ele foi definido para a Regressão simples. Sua interpretação também é a mesma já comentada na Seção 1.10.2.

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} = 1 - \frac{\underline{e}^T \underline{e}}{\sum_{i=1}^{n} (y_i - \bar{y})^2} = 1 - \frac{SSE}{SSTO} \tag{2.9}$$

Exercícios para Aulas Práticas do Capítulo 2

1. Um estúdio fotográfico, especializado em retratos de infantis, tem sede em 21 cidades e está interessado em expandir para outras cidades. Diante desse objetivo os representantes do estúdio resolveram analisar se as vendas em uma cidade (y) pode ser prevista a partir do número de habitantes menores de 16 anos (x_1) e a renda per capita (x_2) na cidade em questão. Para realizar essa análise foram coletados, para cada uma das 21 cidades em que o estúdio já tem sede, os valores das variáveis em questão. Tais dados estão disponíveis em `CH06FI05.txt` (`https://netfiles.umn.edu/users/nacht001/www/nachtsheim/5th/KutnerData/Chapter%20%206%20Data%20Sets/CH06FI05.txt`). Uma breve descrição dos dados:

   - Na primeira coluna estão os valores da variável x_1, número de habitantes menores de 16 anos. Esta variável está expressa em milhares de pessoas.
   - Na segunda coluna estão os valores da variável x_2, renda per capita. Esta variável está expressa em milhares de dólares.
   - Na terceira coluna estão os valores da variável y, total de vendas. Esta variável está expressa em milhares de dólares.

   (a) Encontre as estimativas para os parâmetros β_0, β_1 e β_2.

   (b) Encontre o vetor de valores ajustados $\hat{\mathbf{y}}$.

   (c) Encontre o vetor de resíduos $\underline{\mathbf{e}}$.

   (d) Encontre a estimativa para σ^2.

   (e) Encontre o IC de 95% para os parâmetros β_0, β_1 e β_2.

   (f) O que você pode dizer sobre a existência de uma relação linear entre y e x_1?

   (g) O que você pode dizer sobre a existência de uma relação linear entre y e x_2?

   (h) Seria razoável acreditar que não há vendas quando o número de habitantes e renda per capita da cidade são ambos zero. Verifique se essa afirmação é sustentada pela amostra recolhida e pelo modelo linear ajustado.

   (i) Uma das cidades que o estúdio pretende construir uma nova sede tem 65.400 habitantes com menos de 16 anos e renda per capita de US$17.600,00. Os representantes gostariam de ter uma estimativa intervalar para o valor médio a ser vendido por esta nova sede. O que você responderia a eles?

   (j) Considerando a mesma cidade o item anterior, informe uma estimativa intervalar para as vendas nesta nova sede.

   (k) Faça o gráfico de dispersão das variáveis x_1 e x_2. Em seguida identifique no gráfico os pontos com maior e menor valor da medida h.

   (l) Agora encontre o valor da medida h para o nível considerado nos itens 1i e 1j. Em seguida adicione esse ponto no gráfico de dispersão do exercício anterior. Comente.

   (m) Faça gráficos dos resíduos padronizados versus $\hat{y}$ e versus cada variável preditiva considerada no modelo. Faça também o `qqplot` dos resíduos padronizados. Interprete os gráficos e resuma suas conclusões sobre diagnósticos de violação das suposições do modelo.

2. Uma imobiliária analisa algumas características sobre imóveis comerciais para aluguel a fim de obter informações quantitativas que auxiliem nas tomadas de decisão da empresa. As características analisadas são: a idade do imóvel em anos (x_1), as despesas operacionais em dólar (x_2), a taxa de desocupação (x_3) e a área total em pés quadrados (x_4). A variável de interesse é a taxa de aluguel (y). Os dados para análise estão disponíveis em `CH06PR18.txt` (`https://netfiles.umn.edu/users/nacht001/www/nachtsheim/5th/KutnerData/Chapter%20%206%20Data%20Sets/CH06PR18.txt`), onde a primeira coluna informa os valores de y, a segunda os valores de x_1, a terceira os valores de x_2, a quarta os valores de x_3 e a última os valores de x_4.

   (a) Ajuste os dados à uma regressão linear múltipla, considerando as quatro variáveis preditoras, e apresente a função de regressão estimada.

   (b) Encontre o vetor de resíduos padronizados e faça um box plot com estes valores a fim de verificar se a amostra é aparentemente simétrica.

   (c) Faça os gráficos apropriados para investigar se a suposição de linearidade está sendo respeitada, em relação a cada variável preditiva.

   (d) Faça os gráficos apropriados para investigar se a suposição de que os erros têm variância constante está sendo respeitada. Faça também o teste de *Breusch-Pagan* para auxiliar na sua decisão.

   (e) Faça os gráficos apropriados para investigar se a suposição de normalidade dos erros está sendo respeitada.

   (f) Calcule o R^2 desse ajuste.

   (g) Encontre intervalos de confiança com 95% para cada parâmetros β. Comente os resultados.

   (h) O que você pode dizer sobre a existência de uma relação linear entre y e x_k, para $k = 1, 2, 3, 4$?

   (i) Considere o modelo de regressão linear múltiplo seja apropriado para ajustar esses dados. Três imóveis novos chegaram na imobiliária sem a taxa de aluguel.

	1	2	3
x_1:	4,0	6,0	12,0
x_2:	10,0	11,5	12,5
x_3:	0,10	0,00	0,32
x_4:	80.000	120.000	340.000

   A fim de analisar a viabilidade de incluir tais imóveis no porrifólio da empresa os analistas pediram que fossem estimados intervalos para os valores das taxas de aluguel de cada um dos três imóveis. Você pode dizer que esses intervalos são confiáveis? Encontre tais intervalos e comente os resultados.

   (j) Para os mesmos três imóveis do exercício anterior encontre intervalos de confiança para a taxa média de aluguel para imóveis com tais características. Você pode dizer que esses intervalos são confiáveis?

Lista de Exercícios do Capítulo 2

2.1. Um estudo de pequena escala em uma industria alimentícia relacionou o grau de satisfação do consumidor com uma certa marca (y) com o teor de umidade (x_1) e a doçura (x_2) do produto em questão. Os dados referentes a esse estudo estão codificados por critério de confidenciabilidade e podem ser encontrados em `CH06PR05.txt` (`https://netfiles.umn.edu/users/nacht001/www/nachtsheim/5th/KutnerData/Chapter%20%206%20Data%20Sets/CH06PR05.txt`), onde a primeira coluna guarda os valores de y, a segunda os valores de x_1 e a terceira os valores de x_2.

(a) Ajuste o modelo de regressão linear para os dados acima e apresente a função de regressão estimada. Qual a interpretação para a estimativa de β_1?

(b) Defina e execute os testes apropriados para verificar se existe relação linear entre y e cada variável independente do modelo. Enuncie as hipóteses, a regra de decisão, o p-valor e a conclusão para cada teste. Interprete o resultado.

(c) Encontre intervalos de confiança com 90% de confiabilidade para cada parâmetro β considerado no modelo. Interprete os resultados.

(d) Encontre um intervalo de confiança para $\mathrm{E}[y_h]$ quando $x_{h,1} = 5$ e $x_{h,2} = 4$. Interprete o resultado.

(e) Encontre um intervalo de predição para y_h quando $x_{h,1} = 5$ e $x_{h,2} = 4$. Interprete o resultado.

(f) Verifique o valor da medida h para o nível considerado nos dois itens acima. Você diria que é razoável fazer predições com o modelo ajustado para esse nível?

(g) Faça gráficos dos resíduos padronizados versus $\hat{y}$ e versus cada variável preditiva considerada no modelo. Faça também o `qqplot` dos resíduos padronizados. Interprete os gráficos e resuma suas conclusões sobre diagnósticos de violação das suposições do modelo.

2.2. A administração de um hospital deseja estudar a relação entre a satisfação dos paciente (y), a sua idade (x_1), a gravidade da sua doença (x_2) e seu nível de ansiedade (x_3). Para realizar esse estudo 46 pacientes foram entrevistados e o resultado dessas entrevistas podem ser encontrados em `CH06PR15.txt` (`https://netfiles.umn.edu/users/nacht001/www/nachtsheim/5th/KutnerData/Chapter%20%206%20Data%20Sets/CH06PR15.txt`), onde a primeira coluna fornece o índice que indica a satisfação do paciente (y), a segunda coluna fornece a idade do paciente medida em anos (x_1), a terceira coluna fornece um índice que indica a gravidade da sua doença (x_2) e a quarta coluna fornece um índice que indica o seu nível de ansiedade (x_3). Os índices para as variáveis y, x_2 e x_3 são tais que quanto maior seu valor mais satisfeito está o paciente, mais grave é a sua doença e mais ansioso ele está.

(a) Ajuste o modelo de regressão linear para os dados acima e apresente a função de regressão estimada. Qual a interpretação para a estimativa de β_2?

(b) Defina e execute os testes apropriados para verificar se existe relação linear entre y e cada variável independente do modelo. Enuncie as hipóteses, a regra de decisão, o p-valor e a conclusão para cada teste. Interprete o resultado.

(c) Encontre intervalos de confiança com 90% de confiabilidade para cada parâmetro β considerado no modelo. Interprete os resultados.

(d) Encontre um intervalo de confiança com 90% de confiabilidade para a satisfação média de pacientes com 35 anos de idade, índice de gravidade da doença de 45 e índice de ansiedade de 22. Interprete o resultado.

(e) Encontre um intervalo de predição com 90% de confiabilidade para a satisfação de um novo paciente com 35 anos de idade, índice de gravidade da doença de 45 e índice de ansiedade de 22. Interprete o resultado.

(f) Verifique o valor da medida h para o nível considerado nos dois itens acima. Você diria que é razoável fazer predições com o modelo ajustado para esse nível?

(g) Faça gráficos dos resíduos padronizados versus $\hat{y}$ e versus cada variável preditiva considerada no modelo. Faça também o `qqplot` dos resíduos padronizados. Interprete os gráficos e resuma suas conclusões sobre diagnósticos de violação das suposições do modelo.

2.3. Os seguintes dados se referem a um estudo de regressão linear de pequena escala.

i:	**1**	**2**	**3**	**4**	**5**	**6**
$x_{i,1}$:	7	4	16	3	21	8
$x_{i,2}$:	33	41	7	49	5	31
y_i:	42	33	75	28	91	55

Assuma que o modelo linear seja apropriado para ajustar os dados. Usando apenas as contas com matrizes (isto é, não use o computador, é para fazer na mão) determine:

(a) O vetor de estimativas $\underline{\hat{\beta}}$

(b) O vetor de resíduos $\underline{e}$

(c) As matrizes (X^TX), $(X^TX)^{-1}$ e H

(d) A estimativa para a variância de $\underline{\hat{\beta}}$

(e) A estimativa para o valor ajustado $\hat{y}_h$ quando $x_{h,1} = 10$ e $x_{h,2} = 30$

(f) A estimativa para a variância de $\hat{y}_h$ quando $x_{h,1} = 10$ e $x_{h,2} = 30$

2.4. Considere o modelo de regressão múltipla

$$y_i = \beta_1 x_{i,1} + \beta_2 x_{i,2} + \varepsilon_i, \quad i = 1, 2, \ldots, n$$

onde $\mathrm{E}[\varepsilon_i] = 0$, $Var(\varepsilon_i) = \sigma^2$ e $\mathrm{E}[\varepsilon_i \varepsilon_j] = 0$ para $i \neq j$.

(a) Escreva o problema na forma matricial, ou seja, defina a matriz $\mathbf{X}$ e os vetores $\underline{\beta}$, $\underline{y}$ e $\underline{\varepsilon}$ tais que $\underline{y} = \mathbf{X}\underline{\beta} + \underline{\varepsilon}$.

(b) Encontre os estimadores por mínimos quadrados para β_1 e β_2.

(c) Assumindo agora o modelo normal, isto é, assumindo que $\varepsilon \sim N_p(\underline{0}, \sigma^2 I)$, encontre a função de máxima verossimilhança e em seguida os estimadores de máxima verossimilhança para β_1 e β_2. Estes foram os mesmo estimadores encontrados pelo método de mínimos quadrados?

2.5. Considere o Modelo de Regressão Linear Múltiplo com $p-1$ variáveis independentes definido em sala de aula (Definição 2.2.1). Mostre que $\underline{\hat{y}} \sim N_n\left(X\underline{\beta}, \sigma^2 H\right)$.

2.6. Considere o Modelo de Regressão Linear Múltiplo com $p-1$ variáveis independentes definido em sala de aula (Definição 2.2.1). Mostre que $\sum_{i=1}^{n} \mathrm{Var}(\hat{y}_i) = \sigma^2 p$.

Capítulo 3

Alguns Tópicos em Regressão Linear Múltipla

3.1 ANOVA no Modelo de Regressão Múltipla

Nessa seção veremos a definição da Tabela ANOVA e os testes de hipóteses que podemos definir a partir dela.

3.1.1 Decomposição dos Desvios

Antes de apresentar a Tabela ANOVA para o Modelo de Regressão Linear vejamos algumas definições e relações.

É fácil mostrar que qualquer que seja a regressão linear é sempre verdade que

$$(y_i - \bar{y}) = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i) \;\; \forall i = 1, 2, \ldots, n$$

O que já não é tão simples de mostrar é que

$$\underbrace{\sum_{i=1}^{n} (y_i - \bar{y})^2}_{SSTO} = \underbrace{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}_{SSR} + \underbrace{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}_{SSE} \tag{3.1}$$

Onde casa siglas representa:

$SSTO$	=	*Total Sum of Square*	=	Variação Total,
SSR	=	*Regression Sum of Square*	=	Variação Explicada
SSE	=	*Error Sum of Square*	=	Variação não Explicada

Veja a seguir a demonstração da Equação 3.1.

$$\begin{aligned}
\underbrace{\sum_{i=1}^{n} (y_i - \bar{y})^2}_{SSTO} &= \sum_{i=1}^{n} \left[(y_i - \hat{y}_i) + (\hat{y}_i - \bar{y}) \right]^2 \\
&= \underbrace{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}_{SSE} + \underbrace{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}_{SSR} + \underbrace{\sum_{i=1}^{n} 2 (y_i - \hat{y}_i)(\hat{y}_i - \bar{y})}_{\star}
\end{aligned}$$

Então precisamos apenas mostrar que $\star = 0$. Veja que,

$$\begin{aligned}
\star &= \sum_{i=1}^{n} 2\left(y_i - \hat{y}_i\right)\left(\hat{y}_i - \bar{y}\right) = 2\left[\sum_{i=1}^{n} \hat{y}_i\left(y_i - \hat{y}_i\right) - \sum_{i=1}^{n} \bar{y}\left(y_i - \hat{y}_i\right)\right] \\
&= 2\left[\sum_{i=1}^{n} \hat{y}_i e_i - \bar{y}\sum_{i=1}^{n} e_i\right] = 2\left[\underline{\hat{y}}^T\underline{e} - \bar{y}\cancel{\sum_{i=1}^{n} e_i}\right] = 2\left[\left(X\underline{\beta}\right)^T(I-H)\underline{y}\right] \\
&= 2\left[\underline{\beta}^T X^T(I-H)\underline{y}\right] = 2\left[\underline{\beta}^T X^T\underline{y} - \underline{\beta}^T X^T X(X^TX)^{-1}X^T\underline{y}\right] \\
&= 2\left[\underline{\beta}^T X^T\underline{y} - \underline{\beta}^T X^T\underline{y}\right] = 0
\end{aligned}$$

3.1.2 Decomposição do SSR

Vamos definir $SSR(x_1)$ a variação explicada do modelo que contem apenas a variável preditiva x_1, $SSR(x_1, x_2)$ a variação explicada do modelo que contem as variáveis preditivas x_1 e x_2, e assim por diante.

Veja que a inclusão de uma variáveis no modelo de regressão linear diminui o valor de SSE (no pior dos casos fica igual, nunca maior) e em nada altera o valor de $SSTO$. Dessa forma podemos concluir que o valor de SSR aumenta com a inclusão de uma variável no modelo, uma vez que a igualdade $SSTO = SSR + SSE$ vale para qualquer modelos de regressão linear.

Dessa maneira podemos definir

$$SSR(x_2|x_1) = SSR(x_1, x_2) - SSR(x_1)$$

como o aumento do SSR com a inclusão da variável x_2 no modelo que já tinha a variável x_1. Podemos estender essa definição para o caso de três variáveis:

$$SSR(x_3|x_1, x_2) = SSR(x_1, x_2, x_3) - SSR(x_1, x_2)$$

será o aumento do SSR com a inclusão de x_3 no modelo com as variáveis x_1 e x_2.

Veja que podemos escrever as mesmas equações de outra forma:

$$\begin{aligned}
SSR(x_1, x_2) &= SSR(x_1) + SSR(x_2|x_1) \\
SSR(x_1, x_2, x_3) &= SSR(x_1, x_2) + SSR(x_3|x_1, x_2) = SSR(x_1) + SSR(x_2|x_1) + SSR(x_3|x_1, x_2)
\end{aligned}$$

3.1.3 A Tabela ANOVA

Tradicionalmente a Tabela ANOVA é definida nos livros por:

Fonte de Variação	Soma dos Quadrados	Graus de Liberdade	Média
Regressão	SSR	p-1	MSR = SSR/(p-1)
Erro	SSE	n-p	MSE = SSE/(n-p)
Total	SSTO	n-1	

Tabela 3.1: Tabela ANOVA definida nos livros

No R essa tabela é gerada a partir do comando `anova` e a primeira linha é desmembrada na soma incremental do SSR. Suponha um modelo com três variáveis preditivas x_1, x_2 e x_3 definido no R pelo comando `ml <- lm(y  x1 + x2 + x3)`. Ao digitar `anova(ml)` o R fornece os seguintes valores apresentados na Tabela 3.2.

Fonte de Variação	Soma dos Quadrados	Graus de Liberdade	Média
x_1	SSR(x_1)	1	MSR(x_1) = SSR(x_1)/1
$x_2\|x_1$	SSR($x_2\|x_1$)	1	MSR($x_2\|x_1$) = SSR($x_2\|x_1$)/1
$x_3\|x_1, x_2$	SSR($x_3\|x_1, x_2$)	1	MSR($x_3\|x_1, x_2$) = SSR($x_3\|x_1, x_2$)/1
Erro	SSE	n-p	MSE = SSE/(n-p)
Total	SSTO	n-1	

Tabela 3.2: Tabela ANOVA fornecida pelo R

3.1.4 Teste de Significância da Regressão (Teste F)

Em Regressão Linear a Tabela ANOVA é usada geralmente para definir o Teste da Significância da Regressão. Já vimos o teste t para o modelo múltiplo, onde verificávamos se cada $\beta_k = 0$, um de cada vez. O que vamos fazer agora é verificar se todos os β_k's são nulos simultaneamente, a menos do β_0. Ou seja, queremos verificar com esse teste se algumas das variáveis preditivas contribuem para a regressão linear.

Esse novo teste é definido pelas hipóteses:

$$\begin{aligned} H_0&: \quad \beta_1 = \beta_2 = \ldots = \beta_{p-1} = 0 \\ H_1&: \quad \text{algum } \beta_k \neq 0,\ k = 1, 2, \ldots, p-1 \end{aligned}$$

Antes de construir a estatística de teste usada vamos fazer algumas observações. Veja que sob H_0, isto é, supondo $\beta_1 = \beta_2 = \ldots = \beta_{p-1} = 0$, temos $\hat{y}_i \approx \bar{y}$ e por isso esperamos ter $SSR \approx 0 \Rightarrow MSR \approx 0$. Consequentemente, ainda sob H_0, esperamos ter $SSE \approx SSTO > 0 \Rightarrow MSE > 0$. A partir desse raciocínio podemos construir a ideia do teste, que será aceitar H_0 se $MSR \approx 0$ e $MSR << MSE$, ou seja, se MSR/MSE for bem pequeno. Caso contrário vamos rejeitar H_0.

Para formalizar a regra de decisão do teste é preciso definir o que é "bem pequeno" para a razão MSR/MSE. Para isso vamos precisar saber que:

$$(p-1)MSR/\sigma^2 \sim \chi^2_{p-1}\ , \quad \text{sob } H_0$$

e que MSE e MSR são variáveis independentes, veja a demonstração desse resultado no Apêndice C3 de [Montgomery et al., 2012]. Além disso já vimos que $(n-p)MSE/\sigma^2 \sim \chi^2_{n-p}$. Com esses resultados é fácil encontrar a distribuição, sob H_0, da estatística de teste F definida por:

$$F = \frac{MSR}{MSE} \sim F_{p-1,n-p}\ , \quad \text{sob } H_0$$

Então a regra de decisão será definida por:

- $F \geq F_{1-\alpha,p-1,n-p}$ rejeitamos H_0;
- $F < F_{1-\alpha,p-1,n-p}$ não rejeitamos H_0.

Veja que o teste F definido acima é um teste unilateral, mesmo as hipóteses sendo do tipo "=" e "≠". Isso segue do fato de que $E[MSR] = \sigma^2 + cte$, com $cte = 0$ sob H_0 e $cte > 0$ sob H_1. Ou seja,

$$E[F|H_0] < E[F|H_1]$$

e por isso vamos rejeitar H_0 somente se F assumir valores relativamente grandes e nunca se ela assumir valores relativamente pequenos.

O valor da estatística F e o p-valor para esse teste são fornecidos pelo R quando chamamos o comando `summary`. Eles aparecem na última linha no seguinte formato: `F-statistic: 99.1 on 2 and 18 DF, p-value: 1.921e-10`

Para esse exemplo a estatística F assumiu o valor 99.1, os graus de liberdade dela são 2 e 18 (isto é, temos $p = 3$ e $n = 21$), e o p-valor do teste F é 1.921×10^{-10}.

3.1.5 Teste Geral - para um subconjunto dos β_k's

Além do Teste da Significância da Regressão podemos usar a Tabela ANOVA também para criar um teste que verifica se um subconjunto dos β's são nulos, o que significa verificar se um grupo de variáveis preditivas contribuem para o modelo conjuntamente.

Sem perda de generalidade vamos supor que queremos verificar se as variáveis x_q, x_{q+1}, ..., x_{p-1} contribuem para a regressão simultaneamente. Ou seja, queremos verificar as hipóteses:

$$\begin{array}{ll} H_0: & \beta_q = \beta_{q+1} = \ldots = \beta_{p-1} = 0 \\ H_1: & \text{algum } \beta_k \neq 0,\ k = q, q+1, \ldots, p-1 \end{array}$$

O que estamos fazendo ao testar essas hipóteses é comparar dois modelo: o primeiro, vamos chamá-lo de modelo completo (F - Full), inclui todas as variáveis preditivas e o segundo, vamos chamá-lo de reduzido (R - Reduced), incluí as variáveis preditivas $x_1, x_2, \ldots, x_{q-1}$, ou seja, excluí as variáveis que estão sendo testadas.

A ideia desse teste é comparar os valores de $SSE(F)$ e $SSE(R)$. Já vimos que $SSE(F) \leq SSE(R)$ sempre. Mas sob H_1, isto é, supondo as variáveis x_q, x_{q+1}, ..., x_{p-1} importantes para a regressão, esperamos que $SSE(F) <<< SSE(R)$ e sob H_0 esperamos $SSE(F) \approx SSE(R)$.

Para encontrar a distribuição nula da estatística de teste precisamos saber que

$$(p-q)\left(\frac{SSE(R) - SSE(F)}{p-q}\right)/\sigma^2 \sim \chi^2_{p-q}\ , \quad \text{sob } H_0.$$

Logo a estatística de teste usada para verificar essas hipóteses será:

$$F = \frac{(SSE(R) - SSE(F))\,/p-q}{MSE} \sim F_{p-q,n-p}\ , \quad \text{sob } H_0$$

e a regra de decisão definida por:

- $F \geq F_{1-\alpha,p-q,n-p}$ rejeitamos H_0;
- $F < F_{1-\alpha,p-q,n-p}$ não rejeitamos H_0.

Veja que $SSE(R) - SSE(F) = SSR(F) - SSR(R) = SSR(x_q, \ldots, x_{p-1}|x_1, \ldots, x_{q-1})$. Este valor pode ser retirado da Tabela ANOVA 3.2 somando as $p-q$ últimas linhas referentes ao SSR. Logo a estatística F para esse teste pode ser retirada facilmente da Tabela ANOVA.

Para fazer uma analogia com a estatística usada no Teste da Significância da Regressão alguns livros apresentam essa estatística de teste no seguinte formato:

$$F = \frac{MSR(x_q, x_{q+1}, \ldots, x_{p-1}|x_1, \ldots, x_{q-1})}{MSE} \sim F_{p-q,n-p} \ \text{ sob } H_0$$

3.1.6 Comparação entre os Testes

Já vimos três testes: o Teste t, o Teste F e o Teste Geral. Primeiro veja que se o teste Geral for feito para $q = 1$ teremos exatamente o teste F. Ou seja, o teste F é um caso particular do Teste Geral. Gostaria agora de fazer algumas observações e comparações entre esses três testes.

Toda vez que um desses três testes é feito estamos na verdade comparando dois modelos diferentes e testando qual dos dois é mais adequado para os dados. Para exemplificar essa afirmação vamos supor o modelo com três variáveis preditivas definido por:

$$y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,3} + \varepsilon_i \quad (F)$$

Esse modelo será chamado de completo e representado pela letra F do inglês *full*.

Em cada um dos testes descritos a seguir iremos compara o modelo completo (F) com algum outro modelo, chamado de reduzido e representado pela letra R. O modelo reduzido sempre será definido supondo H_0 verdadeiro. Vejamos cada caso.

Primeiro suponha o teste t definido pelas hipóteses

$$\begin{array}{ll} H_0: & \beta_2 = 0 \\ H_1: & \beta_2 \neq 0 \end{array}$$

Nesse caso o modelo reduzido será definido por

$$y_i = \beta_0 + \beta_1 x_{i,1} + \beta_3 x_{i,3} + \varepsilon_i \quad (R_1)$$

e as hipóteses acima podem ser substituídas por:

$$\begin{array}{ll} H_0: & \text{O modelo } R_1 \text{ é mais adequado} \\ H_1: & \text{O modelo } F \text{ é mais adequado} \end{array}$$

Considere agora o teste F (Teste da Significância) definido pelas hipóteses

$$\begin{array}{ll} H_0: & \beta_1 = \beta_2 = \beta_3 = 0 \\ H_1: & \text{pelo menos um } \beta \neq 0 \end{array}$$

Nesse caso o modelo reduzido será definido por

$$y_i = \beta_0 + \varepsilon_i \quad (R_2)$$

e as hipóteses acima podem ser substituídas por:

$$\begin{array}{ll} H_0: & \text{O modelo } R_2 \text{ é mais adequado} \\ H_1: & \text{O modelo } F \text{ é mais adequado} \end{array}$$

Para terminar suponha o teste Geral definido pelas hipóteses

$$\begin{array}{ll} H_0: & \beta_2 = \beta_3 = 0 \\ H_1: & \beta_2 \neq 0 \text{ ou } \beta_3 \neq 0 \end{array}$$

Nesse caso o modelo reduzido será definido por

$$y_i = \beta_0 + \beta_1 x_{i,1} + \varepsilon_i \quad (R_3)$$

e as hipóteses acima podem ser substituídas por:

$$\begin{array}{ll} H_0: & \text{O modelo } R_3 \text{ é mais adequado} \\ H_1: & \text{O modelo } F \text{ é mais adequado} \end{array}$$

Pensando dessa forma fica mais fácil entender a conclusão de cada teste e saber a hora certa de usar cada um deles.

3.2 Inclusão de Variáveis Qualitativas

Até agora já comentamos que é possível incluir variáveis preditivas qualitativas como sexo ou classe social no Modelo de Regressão Linear, mas ainda não foi visto nenhum exemplo desse tipo de variável. Nessa seção veremos como essas variáveis serão incluídas no modelo e as interpretações que esse tipo de variável possibilita.

3.2.1 Variáveis Qualitativas com 2 Classes

Primeiro vejamos o caso mais simples, onde a variável qualitativa possui apenas duas classes. Por exemplo, suponha que o interesse seja explicar o índice de satisfação dos funcionários de uma empresa (y) a partir do lucro dessa empresa (x_1). Para realizar esse estudo foram coletadas informações referentes a empresas tanto do Rio quanto de São Paulo. Se ajustarmos os dados para o modelo simples

$$y_i = \beta_0 + \beta_1 x_{i,1} + \varepsilon_i$$

estamos assumindo que o índice médio de satisfação dos funcionários de uma empresa depende apenas do lucro dessa empresa e não da sua cidade. Mas será que, em média, o índice de satisfação de empresas com mesmo lucro, sendo uma no Rio e a outra em SP, é o mesmo? Talvez a variável *cidade* seja uma variável importante e devesse ser incorporada ao modelo.

Para incorporar uma variável qualitativa no modelo de regressão linear vamos utilizar variáveis indicadoras, também chamadas de *dummy* ou binárias. Nesse caso, como a variável qualitativa *cidade* tem apenas duas classes será criada apenas uma variável indicadora, definida por:

$$x_2 = \begin{cases} 1 & , \text{ se RJ} \\ 0 & , \text{ se SP.} \end{cases}$$

O modelo com essa nova variável passa a ser definido por:

$$y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \varepsilon_i \tag{3.2}$$

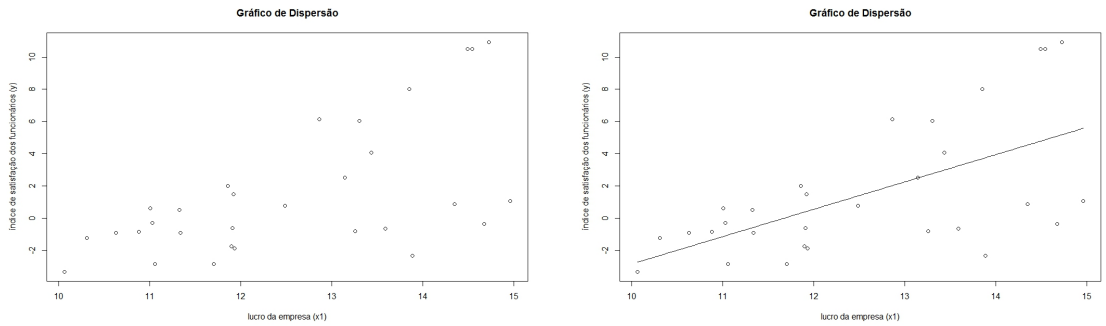
Vejamos alguns comentários importantes:

- Não devemos criar uma variável indicadoras para cada classe. Se fizermos isso a matriz $(X^T X)$ não será inversível e por isso não teremos estimadores para β por mínimos quadrados. Vamos criar apenas uma nova variável indicadora, que representa as duas classes ao mesmo tempo.

- Veja que a matriz X para o modelo com x_1 e x_2 terá a última coluna com entradas iguais a 0's ou 1's.

Interpretação dos Parâmetros

Para interpretar os parâmetros desse modelo vamos continuar com o exemplo citado acima, onde y indica um índice de satisfação dos funcionários de uma empresa, x_1 o lucro dessa empresa (medido em milhões de reais) e x_2 a variável que indica se a empresa é do Rio de Janeiro ou de São Paulo. Suponha que 30 empresas tenham sido avaliadas e os valores de y e x_1 para essas empresas sejam os apresentados na Figura 3.1(a).



(a) Gráfico de Dispersão (b) Reta de Regressão Estimada

Figura 3.1: Ajuste sem incluir a variável x_2 que define a cidade

Se ajustarmos um modelo de regressão linear simples considerando apenas a variável x_1, a função de regressão será $E[y_i] = \beta_0 + \beta_1 x_{i,1}$ e os parâmetros têm as interpretações já conhecidas: β_0 representa a média do índice de satisfação de funcionários em empresas com lucro zero e β_1 a quantidade que esse índice cresce (ou decresce) quando o lucro da empresa cresce em 1 milhão de reais. Nesse caso a nossa reta de regressão estimada seria como na Figura 3.1(b).

Depois de uma observação mais detalhada percebemos que a variável *cidade* parece importante na descrição do índice de satisfação, veja o gráfico de dispersão da Figura 3.2(a), onde as empresas do Rio de Janeiro são representadas pelas bolinhas cheias e as empresas de São Paulo pelas bolinhas vazias. Baseado nesse gráfico decidimos então incorporar ao modelo a variável *cidade* e isso será feito a partir da criação de uma variável indicadora x_2, como definida anteriormente. Então o modelo de regressão linear adotado é aquele definido na Equação 3.2 e a função de regressão será $E[y_i] = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2}$.

Para melhor interpretar os parâmetros desse modelo vamos separá-lo em dois casos: primeiro considerando $x_2 = 0$, empresas de São Paulo; e depois $x_2 = 1$, empresas no Rio de Janeiro. A função de regressão para as empresas de São Paulo é definida por

$$E[y_i] = \beta_0 + \beta_1 x_{i,1},$$

uma vez que nesse caso $x_2 = 0$. Então β_0 representa a média do índice de satisfação dos funcionários de empresas localizadas na cidade de São Paulo com lucro zero e β_1 representa o acréscimo (ou decréscimo) na média do índice de satisfação dos funcionário quanto o lucro da empresa de São Paulo aumenta em 1 milhão de reais.

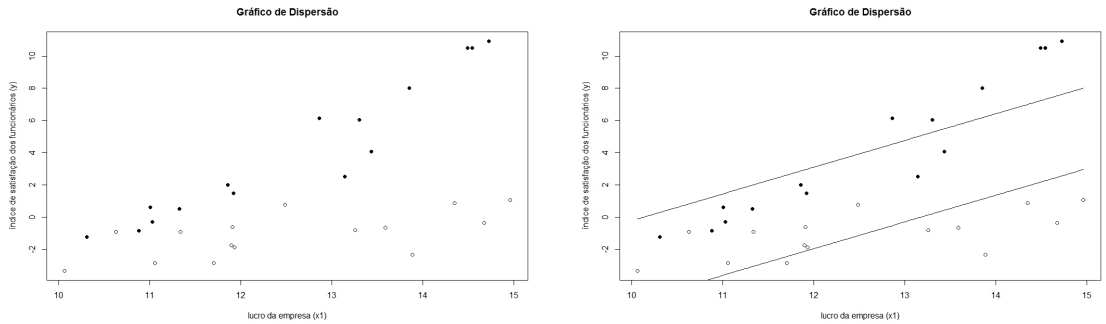
Vejamos agora para as empresas do Rio de Janeiro. Para essas empresas a função de regressão será definida por:

$$E[y_i] = \beta_0 + \beta_1 x_{i,1} + \beta_2 = (\beta_0 + \beta_2) + \beta_1 x_{i,1},$$

uma vez que nesse caso $x_2 = 1$. Veja que para as empresas do Rio de Janeiro a média do índice de satisfação dos funcionários quando a empresa tem lucro zero é representado por $\beta_0 + \beta_2$ e, assim como em São Paulo, β_1 representa o acréscimo (ou decréscimo) na média do índice de satisfação dos funcionários quanto o lucro da empresa aumenta em 1 milhão de reais.

Para esse exemplo a função de regressão estimada será representada por duas retas, uma para as empresas o Rio de Janeiro e outra para as de São Paulo. Ambas encontram-se na Figura 3.2(b). Veja que a partir desse gráfico é possível afirmar que $\hat{\beta}_2 > 0$, uma vez

que a reta referente às empresas do Rio de Janeiro encontra-se acima da reta referente às empresas de São Paulo.



(a) Gráfico de Dispersão (b) Reta de Regressão Estimada

Figura 3.2: Ajuste incluindo a variável x_2 que define a cidade

Assim concluímos que o modelo definido pela Equação 3.2 considera um comportamento diferente entre as empresas do Rio de Janeiro e São Paulo, mas essa diferença é apenas no índice médio quando o lucro é zero. Nesse modelo a taxa na mudança do índice de satisfação dos funcionários (β_1) é a mesma para as empresas nas duas cidades.

Modelo com o Termo Cruzado

Continuando ainda com o exemplo em que y indica um índice de satisfação dos funcionários de uma empresa, x_1 o lucro e x_2 a cidade, podemos perceber, observando a Figura 3.2(a), que a taxa de crescimento é diferente para as empresas do Rio de Janeiro e de São Paulo. Aparentemente o aumento de 1 milhão de reais no lucro das empresas gera um aumento maior no índice de satisfação dos funcionários do Rio de Janeiro do que em São Paulo.

Para incorporar essa diferença no modelo de regressão linear podemos optar pela inclusão do termo cruzado x_1x_2 e definir um novo modelo com mais uma variável preditiva:

$$y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,1} x_{i,2} + \varepsilon_i. \tag{3.3}$$

A interpretação de cada parâmetro para esse novo modelo será feita novamente separando os casos $x_2 = 0$ e $x_2 = 1$. Considerando o modelo definido na Equação 3.3, a função de regressão para as empresas de São Paulo continuará sendo

$$E[y_i] = \beta_0 + \beta_1 x_{i,1},$$

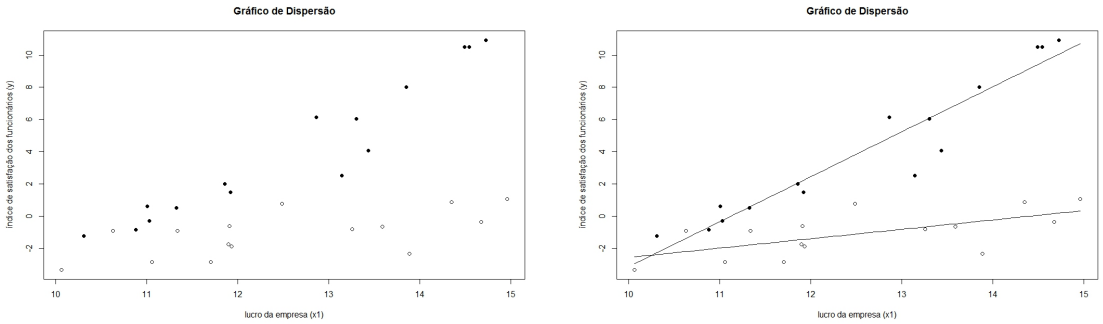
já a função de regressão para as empresas do Rio de Janeiro será

$$E[y_i] = \beta_0 + \beta_1 x_{i,1} + \beta_2 + \beta_3 x_{i,1} = (\beta_0 + \beta_2) + (\beta_1 + \beta_3)\, x_{i,1}.$$

Veja que agora as funções de regressão para cada uma das duas cidades se diferem não somente no coeficiente linear como também no coeficiente angular. Isso significa que para esse novo modelo o índice de satisfação médio dos funcionários em empresas com lucro zero são diferentes para empresas de diferentes cidades: β_0 para as empresas de São Paulo e $\beta_0+\beta_2$ para as empresas do Rio de Janeiro. Além disso o novo modelo também considera que o aumento de 1 milhão de reais no lucro das empresas gera diferentes mudanças na

média do índice de satisfação dos funcionários para empresas de diferentes cidades: β_1 para as empresas de São Paulo e $\beta_1 + \beta_3$ para as empresas do Rio de Janeiro.

A função de regressão estimada continua sendo representada por duas retas, mas agora estas retas não são paralelas como no caso anterior. A Figura 3.3(b) apresenta o ajuste dos dados para o modelo da Equação 3.3.

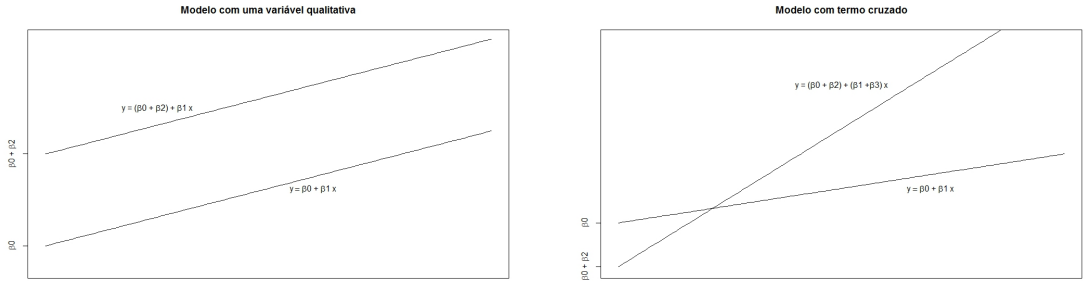


(a) Gráfico de Dispersão (b) Reta de Regressão Estimada

Figura 3.3: Ajuste incluindo a variável x_2 e o termo cruzado x_1x_2.

A partir do gráfico da Figura 3.3(b) é possível concluir que o ajuste dos dados para o modelo apresentado na Equação 3.3 gerou $\hat{\beta}_2 < 0$, pois quando $x = 0$ a reta referente às empresas do Rio de Janeiro encontra-se abaixo da reta de São Paulo. Além disso também podemos afirmar que para esse ajuste $\hat{\beta}_3 > 0$, uma vez que a reta referente às empresas do Rio de Janeiro é mais inclinada que a reta de São Paulo.

A Figura 3.4 resume a diferença entre os dois modelos: o apresentado na Equação 3.2, sem o termo cruzado, e o apresentado na Equação 3.3, com o termo cruzado.



(a) Modelo sem o termo cruzado (b) Modelo com o termo cruzado

Figura 3.4: Comparação entro os modelos com e sem o termo cruzado.

A escolha de qual dos dois modelos usar depende se queremos ou não diferenciar o coeficiente angular das duas retas de regressão. Em geral iniciamos com o modelo completo (Equação 3.3) e depois realizamos o teste t para testar se $H_0 : \beta_3 = 0$ contra $H_1 : \beta_3 \neq 0$. Se a conclusão do teste for rejeitar H_0 ficamos com o modelo completo. Se a conclusão for aceitar H_0 ficamos com o modelo simplificado (Equação 3.2). Veremos esse procedimento com mais detalhes na Seção 3.4 deste capítulo.

3.2.2 Variáveis Qualitativas com mais de 2 Classes

Para introduzir as variáveis qualitativas com mais de 2 classes no modelo de regressão linear também serão usadas variáveis indicadoras. O único cuidado que temos que ter é que o número de variáveis indicadoras usadas será sempre uma a menos do número de classes da variável qualitativa em questão.

Considere ainda o exemplo em que y representa o índice de satisfação dos funcionários e x_1 o lucro das empresas. Se agora em vez de empresas do Rio de Janeiro e de São Paulo a amostra conter também informações de empresas de Minas Gerais a variável *cidade* será uma variável qualitativa com três classes e para que ela seja introduzida no modelo será preciso criar 2 variáveis indicadoras:

$$x_2 = \begin{cases} 1 & , \text{ se RJ} \\ 0 & , \text{ caso contrário.} \end{cases} \qquad x_3 = \begin{cases} 1 & , \text{ se SP} \\ 0 & , \text{ caso contrário.} \end{cases}$$

O modelo de regressão linear será definido por:

$$y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,3} + \varepsilon_i. \tag{3.4}$$

Interpretação dos Parâmetros

A interpretação dos parâmetros será feita de forma semelhante ao caso em que a variável qualitativa possui apenas 2 classes. Nesse caso teremos que considerar todas as codificações para x_2 e x_3 que definem a cidade, de acordo com a Tabela 3.3 abaixo.

Cidade	x_2	x_3
RJ	1	0
SP	0	1
MG	0	0

Tabela 3.3: Codificação de x_2 e x_3 para cada cidade.

Então, para as empresas do Rio de Janeiro, codificadas por $x_2 = 1$ e $x_3 = 0$, a função de regressão será definida por

$$E[y_i] = \beta_0 + \beta_1 x_{i,1} + \beta_2 = (\beta_0 + \beta_2) + \beta_1 x_{i,1}.$$

Para as empresas de São Paulo, codificadas por $x_2 = 0$ e $x_3 = 1$, a função de regressão será definida por

$$E[y_i] = \beta_0 + \beta_1 x_{i,1} + \beta_3 = (\beta_0 + \beta_3) + \beta_1 x_{i,1}.$$

Por fim, para as empresas de Minas Gerais, codificadas por $x_2 = 0$ e $x_3 = 0$, a função de regressão será definida por

$$E[y_i] = \beta_0 + \beta_1 x_{i,1}.$$

Dessa maneira podemos concluir que, de acordo com o modelo linear definido na Equação 3.4, diferentes cidades têm diferente média do índice de satisfação dos funcionários considerando empresas com lucro zero. Para o exemplo, a média do índice de satisfação dos funcionários de empresas com lucro zero em Minas Gerais é de β_0, a média do índice de satisfação dos funcionários de empresas com lucro zero no Rio de Janeiro é $\beta_0 + \beta_2$ e a média do índice de satisfação dos funcionários de empresas com lucro zero em São Paulo é $\beta_0 + \beta_3$.

Por outro lado esse modelo não diferencia a taxa de crescimento da média do índice de satisfação dos funcionários em função do lucro da empresa para diferentes cidades.

Modelo com o Termo Cruzado

Assim como no caso das variáveis qualitativas com duas classes, se queremos diferenciar a taxa de crescimento para as diferentes cidades é preciso introduzir no modelo de regressão linear além das variáveis indicadoras também os termos cruzados. Nesse caso o modelo de regressão linear com os termos cruzados será definido por:

$$y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,3} + \beta_4 x_{i,1} x_{i,2} + \beta_5 x_{i,1} x_{i,3} + \varepsilon_i. \tag{3.5}$$

Veja que quanto mais classes, mais variáveis indicadoras, mais termos cruzados, mais complexo fica o modelo.

Se quisermos encontrar as funções de regressão para cada cidade será preciso novamente separar em casos. Para as empresas do Rio de Janeiro, codificadas por $x_2 = 1$ e $x_3 = 0$, a função de regressão será definida por

$$E[y_i] = \beta_0 + \beta_1 x_{i,1} + \beta_2 + \beta_4 x_{i,1} = (\beta_0 + \beta_2) + (\beta_1 + \beta_4)\, x_{i,1}.$$

Para as empresas de São Paulo, codificadas por $x_2 = 0$ e $x_3 = 1$, a função de regressão será definida por

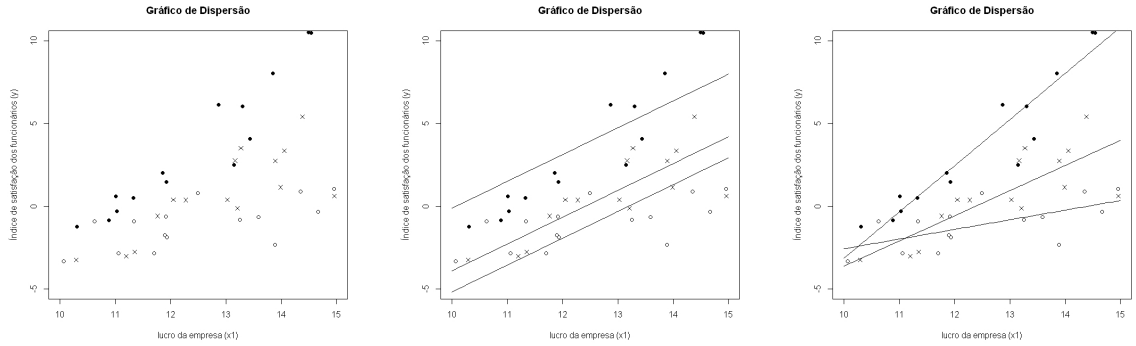
$$E[y_i] = \beta_0 + \beta_1 x_{i,1} + \beta_3 + \beta_5 x_{i,1} = (\beta_0 + \beta_3) + (\beta_1 + \beta_5)\, x_{i,1}.$$

Para as empresas de Minas Gerais, codificadas por $x_2 = 0$ e $x_3 = 0$, a função de regressão será definida por

$$E[y_i] = \beta_0 + \beta_1 x_{i,1}.$$

Dessa forma as funções de regressão para cada cidade serão diferentes não só no coeficiente linear como também no angular.

Na Figura 3.5 abaixo estão três gráficos. O primeiro, 3.5(a), apresenta o gráfico de dispersão para o exemplo incluindo as empresas de Minas Gerais, representadas pelo símbolo $\times$. O segundo, 3.5(b), estão as três retas de regressão estimadas considerando o modelo sem os termos cruzados, apresentado na Equação 3.4. O terceiro, 3.5(c), estão as três retas de regressão estimadas considerando o modelo com os termos cruzados, apresentado na Equação3.5.



(a) Gráfico de Dispersão (b) Modelo sem o termo cruzado (c) Modelo com o termo cruzado

Figura 3.5: Exemplo de variável qualitativa com 3 classes.

Veja que mesmo sem conhecer os valores das estimativas é possível fazer algumas afirmações. Primeiro considerando a Figura 3.5(b), referente ao modelo sem os termos

cruzados (Equação 3.4), veja que nesse caso $\hat{\beta}_2 > 0$ e $\hat{\beta}_3 < 0$. Podemos fazer essas afirmações pois a reta referente a cidade do Rio de Janeiro está acima da reta de Minas Gerais e a reta de São Paulo está abaixo da de Minas Gerais.

Agora considere a Figura 3.5(c), referente ao modelo com os termos cruzados (Equação 3.5). Veja que nesse caso $\hat{\beta}_2 > 0$, $\hat{\beta}_3 > 0$, $\hat{\beta}_4 > 0$ e $\hat{\beta}_5 < 0$. Nesse exemplo Minas Gerais é a cidade de referência pois ela foi a cidade codificada por $x_2 = 0$ e $x_3 = 0$.

3.2.3 Modelo com várias variáveis qualitativas

Não só podemos incluir um variável qualitativa com várias classes como também podemos incluir várias variáveis qualitativas. Nesse caso para cada variável qualitativa serão usadas variáveis indicadoras para representá-la. O número de variáveis indicadoras será sempre um a menos do número de classes.

Continuando ainda no exemplo trabalhado nessa seção, suponha que além das cidades queremos introduzir no modelo o tamanho da empresa, que será representado por: micro, pequena, média e grande. Vamos considerar o modelo sem os termos cruzados para simplificar. A variável `cidade` será codificada pelas variáveis indicadoras x_2 e x_3 definidas na Tabela 3.3. Já para a variável `tamanho da empresa` vamos criar outras 3 variáveis indicadoras:

$$x_4 = \begin{cases} 1 & , \text{ se micro} \\ 0 & , \text{ caso contrário.} \end{cases} \quad x_5 = \begin{cases} 1 & , \text{ se pequena} \\ 0 & , \text{ caso contrário.} \end{cases} \quad x_6 = \begin{cases} 1 & , \text{ se média} \\ 0 & , \text{ caso contrário.} \end{cases}$$

A tabela de codificação dessa nova variável qualitativa é definida por:

Cidade	x_4	x_5	x_6
Micro	1	0	0
Pequena	0	1	0
Média	0	0	1
Grande	0	0	0

Tabela 3.4: Codificação de x_4, x_5 e x_6 para a variável `tamanho da empresa`.

O modelo de regressão linear considerando tanto a variável `cidade` quanto a variável `tamanho da empresa` é definido por:

$$y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,3} + \beta_4 x_{i,4} + \beta_5 x_{i,5} + \beta_6 x_{i,6} + \varepsilon_i. \tag{3.6}$$

A interpretação dos parâmetros será feita considerando cada combinação de possibilidades das variáveis x_2, x_3, x_4, x_5 e x_6. Por exemplo, se temos $x_2 = x_3 = x_4 = x_5 = x_6 = 0$ estamos considerando as Grandes empresas de São Paulo, para essas empresas a função de regressão do modelo que representa a relação entre a satisfação dos funcionários e o lucro das empresas é definido por:

$$E[y_i] = \beta_0 + \beta_1 x_{i,1}.$$

Se quisermos saber a função de regressão para as empresas de médio porte do Rio de Janeiro temos que considerar $x_2 = 1$, $x_3 = 0$, $x_4 = 0$, $x_5 = 0$ e $x_6 = 1$. Para esse caso a função de regressão será:

$$E[y] = \beta_0 + \beta_1 x_{i,1} + \beta_2 + \beta_6 = (\beta_0 + \beta_2 + \beta_6) + \beta_1 x_{i,1}$$

Para terminar, em relação ao modelo 3.6, podemos ainda excluir a variável quantitativa x_1 e ficar apenas com variáveis qualitativas. Nesse caso o modelo é chamado de "Modelos de Análise de Variância".

3.3 Multicolinearidade

Nessa seção vamos começar a discutir quais variáveis preditivas devem ser incluídas no modelo e quais devem ser deixadas de fora. Um dos critérios para escolher algumas entre muitas variáveis preditivas é a análise de multicolinearidade. Depois de feita essa análise, e talvez descartadas algumas variáveis preditivas, seguimos para a seleção do modelo, que será discutida na Seção 3.4 a seguir.

A multicolinearidade existe quando uma ou mais variáveis preditivas são bem correlacionadas entre si. Como consequência disso alguns problemas podem ocorrer, como veremos a seguir. Por isso é de grande de importância checar se existe multicolinearidade nos dados. Caso existe, algumas variáveis preditivas devem ser eliminadas (antes da seleção do modelo) a fim de acabar com a multicolinearidade.

3.3.1 Os Problemas

Quando a correlação entre todos os pares de variáveis preditivas for muito pequena, isto é, próxima de zero, estamos em uma situação favorável: as estimativas para $\underline{\beta}$ pouco vão mudar com a inclusão ou eliminação de variáveis preditivas e o desvio padrão dos estimadores não será grande. Caso contrário, quando a correlação de duas variáveis preditivas for grande, isto é, próxima de 1, teremos alguns sérios problemas.

Se a correlação entre duas variáveis preditivas for grande a matriz X^TX será "quase não inversível", isto significa que esta matriz tem um autovalor muito próximo de zero e a sua inversa $(X^TX)^{-1}$ tem um autovalor muito grande (os autovalores de $(X^TX)^{-1}$ são os inversos dos autovalores de X^TX). O problema disso é que, mesmo conseguindo inverter X^TX, a estimativa para o parâmetro $\underline{\beta}$, dada por $\underline{\hat{\beta}} = (X^TX)^{-1}X^T\underline{y}$, será pouco precisa. Isto é, pequenas mudanças nos valores observados de $\underline{y}$ alteram muito $\underline{\hat{\beta}}$, o que não é nada desejado. Além disso, alguns elementos da diagonal principal de $(X^TX)^{-1}$ serão bem grandes e com isso aumenta a variância dos estimadores de mínimos quadrados $\underline{\hat{\beta}}$, lembre-se que $Var(\beta_k) = \sigma^2 C_{k+1,k+1}$.

Conclusão: temos que evitar trabalhar com variáveis preditivas bem correlacionadas pois isso pode resultar em estimadores pouco precisos e incertezas nas inferências. Como consequência, imprecisão nas estimativas pode gerar interpretações erradas.

3.3.2 Como Diagnosticar

Existem algumas formas de identificar a existência de multicolinearidade nos dados, veremos aqui duas. A primeira delas trata-se simplesmente de encontrar a matriz de correlação r_{XX} entre as variáveis preditivas e procurar se existe alguma entrada, fora da diagonal principal, maior que 0.7 ou 0.8. Caso exista identificamos que existem variáveis com grande correlação e inclusive sabemos quem são elas.

A outra alternativa é usar o *Fator de Inflação de Variância* da variável x_k denominado VIF_k, sigla em inglês de *Variance Inflaction Factor*. Este fator é definido pelos elementos da diagonal principal de $C^\star = (X^{\star T}X^\star)^{-1}$ onde $X^\star$ é a matriz do modelo cujas variáveis preditivas e variável resposta são definidas por:

$$x^\star_{i,k} = \frac{x_{i,k} - \bar{x}_k}{\sqrt{\sum_{i=1}^n (x_{i,k} - \bar{x}_k)^2}} \qquad \text{e} \qquad y^\star_i = \frac{y_- \bar{y}}{\sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}.$$

O modelo de regressão linear definido com as variáveis $x^{\star}_{i,k}$ e $y^{\star}$ é chamado de modelo de regressão padronizado e para esse modelo a matriz $X^{\star}$ é uma matriz $p-1 \times p-1$ cujas colunas são formadas pelas variáveis preditivas $x^{\star}_{i,k}$ e não temos a primeira coluna com todas as entradas iguais a 1. Para mais detalhes sobre o modelo padronizado veja Seção 7.5 de [Kutner et al., 2005].

Quando estamos trabalhando com o modelo de regressão padronizado temos $(X^{\star T}X^{\star}) = r_{XX}$, matriz de correlação das variáveis preditivas do modelo. Então uma maneira mais fácil de encontrar os fatores VIF_k é inverter r_{XX} em vez de precisar ajustar o modelo de regressão padronizado. Dessa forma podemos definir:

$$VIF_k = C^{\star}_{k,k} \quad \text{onde} \quad C^{\star} = r^{-1}_{XX}\ ,\ k = 1, 2, \ldots, p-1.$$

Podemos mostrar que:

$$VIF_k = C^{\star}_{k,k} = \frac{1}{1 - R^2_k}\ , \quad k = 1, \ldots, p-1 \tag{3.7}$$

onde R^2_k é o coeficiente de determinação da regressão que tenta explicar x_k em função das outras $p-2$ variáveis preditivas. Para uma demonstração da Equação 3.7 veja [Kutner et al., 2005] ou [Montgomery et al., 2012].

Analisando a Equação 3.7 podemos perceber que se alguma variável for bem correlacionada com outras o valor de R^2_k será próximo de 1 e o VIF_k muito grande. Caso contrário, todos os R^2_k serão próximos de zero e VIF_k próximos de 1. Dessa forma, se um ou mais valores de VIF_k forem grandes acreditamos na existência de multicolinearidade. Na prática [Montgomery et al., 2012] sugere que se algum VIF_k for maior que que 5 há indícios de que os coeficientes do modelo de regressão não estão bem estimados por causa da multicolinearidade.

3.3.3 Como Tratar

Uma vez identificada a multicolinearidade o que devemos fazer? Antes de iniciar a seleção do modelo devemos eliminar algumas variáveis para acabar com o problema de multicolinearidade. A matriz de correlação e os VIF_k's serão úteis para identificar quais as variáveis que devem ser analisadas.

Escolha uma entre as variáveis que apresentaram valores altos de VIF_k e/ou grande correlação com outras variáveis para sair do modelo. Essa escolha pode ser feita de várias maneiras, por exemplo, você pode escolher a variável com maior VIF_k ou escolher aquela variável com maior p-valor para o teste t, quando ajustado o modelo simples para cada variável que apresentou grande correlação. Depois de eliminada essa variável repita a análise de multicolinearidade e verifique se ainda há multicolinearidade. Caso positivo, escolha mais uma variável preditiva para sair. O processo só deve ser encerrado quando não houver mais indicação da existência de multicolinearidade, isto é, quando todas as correlações forem menores que 0.7 e todos os VIF_k's forem menores que 5.

3.4 Seleção do Modelo

Após realizar a análise de multicolinearidade devemos partir para a seleção do modelo, que consiste em selecionar quais variáveis preditivas farão parte do nosso modelo de regressão linear. A ideia principal da seleção do modelo é que as variáveis que agregam

informação para o modelo devem ser incluídas e as variáveis que agregam pouca ou nenhuma informação devem ser excluídas. Veremos dois métodos de seleção do modelo, o método da comparação entre todos os possíveis modelos e o método da seleção passo-a-passo.

3.4.1 Comparação entre todos os modelos possíveis

Uma maneira para fazer a seleção de quais variáveis deve ou não entrar no modelo de regressão linear é a comparação entre todos os possíveis modelos a partir da comparação das medidas de bom ajustamento de cada modelo. Algumas dessas medidas serão discutidas a seguir, como por exemplo, o Coeficiente de Determinação Ajustado e a Informação de Akaike.

Todos os modelos possíveis são definidos por todas as possíveis combinações das variáveis preditivas. Por exemplo, se temos três variáveis preditivas, x_1, x_2 e x_3, todos os modelos possíveis com essas variáveis são:

$$\begin{aligned}
y_i &= \beta_0 + \beta_1 x_{i,1} + \varepsilon_i \\
y_i &= \beta_0 + \beta_2 x_{i,2} + \varepsilon_i \\
y_i &= \beta_0 + \beta_3 x_{i,3} + \varepsilon_i \\
y_i &= \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \varepsilon_i \\
y_i &= \beta_0 + \beta_1 x_{i,1} + \beta_3 x_{i,3} + \varepsilon_i \\
y_i &= \beta_0 + \beta_2 x_{i,2} + \beta_3 x_{i,3} + \varepsilon_i \\
y_i &= \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,3} + \varepsilon_i
\end{aligned}$$

O principal problema desse método é que se tivermos muitas variáveis preditivas o número de modelos possíveis pode ser inviável de se comparar. Se temos N variáveis preditivas o número de diferentes modelos possíveis será $2^N - 1$. Para o exemplo acima como temos $N = 3$ variáveis preditivas o número de possíveis modelos é $2^3 - 1 = 7$. Se tivermos 4 variáveis preditivas o número de diferentes modelos será $2^4 - 1 = 15$ e se $N = 5$ teremos $2^5 - 1 = 31$. O crescimento é exponencial e rápido chegamos em um número muito grande de possíveis modelos para um número não tão grande de variáveis preditivas.

Mesmo com esse problema a comparação entre todos os possíveis modelos é um critério muito comum para seleção das variáveis. Nesse caso podemos usar como critério de comparação algumas medidas de comparação, como por exemplo, o Coeficiente de Determinação Ajustado e a Informação de Akaike. Cada uma dessas medidas será definida agora.

Uma recomendação para melhor visualizar a comparação dessas medidas de comparação é fazer o gráfico da medida em questão versus o parâmetro p de cada modelo. Esse gráfico vai permitir encontrar para cada valor do parâmetro p, ou seja, para cada número de variáveis preditivas o melhor modelo de acordo com a medida.

Coeficiente de Determinação Ajustado

Já definimos o coeficiente de determinação R^2 por

$$R^2 = 1 - \frac{SSE}{SSTO}.$$

O problema de usar esse critério como comparação entre todos os possíveis modelos é que sempre termos o maior R^2 no modelo com mais variáveis preditivas, uma vez que SSE sempre diminui com a inclusão de variáveis e $SSTO$ é o mesmo para todos os possíveis modelos. Então esse não é um critério comparação.

Motivado por esse problema criou-se o Coeficiente de Determinação Ajustado, R^2, que leva em consideração não só o bom ajuste do modelo dentro da amostra como também o número de variáveis preditivas no modelo. Esse novo coeficiente é definido por

$$R_a^2 = 1 - \frac{(n-1)SSE}{(n-p)SSTO} \tag{3.8}$$

Veja que a inclusão de uma nova variável reduz SSE mas reduz também $n-p$. Dessa forma para que o valor de R_a^2 cresça com a inclusão de uma nova variável é preciso que o SSE reduza significativamente a ponto da razão $SSE/(n-p)$ reduzir com a inclusão dessa nova variável.

Dessa forma R_a^2 pode ser um critério de comparação entre todos os possíveis modelos e o modelo escolhido sera aquele com maior R_a^2.

Critérios da Informação de Akaike

No Critério da Informação de Akaike o melhor modelo entre todos os possíveis modelos é escolhido a partir da medida AIC definida por:

$$AIC = n\ln(SSE) - n\ln(n) + 2p \tag{3.9}$$

Veja que quanto menor for SSE menor será o valor de AIC. Além disso quanto menor for p, isto é, quanto menos variáveis preditivas tiver o modelo, menor será o valor de AIC. Dessa forma queremos modelos com medida AIC pequena, uma vez que queremos modelos com poucas variáveis preditivas e com SSE pequeno.

A partir do critério da Informação de Akaike o modelo escolhido será aquele com menor valor de AIC.

Critérios da Informação Bayesiana

No Critério da Informação Bayesiana o melhor modelo entre todos os possíveis modelos é escolhido a partir da medida BIC definida por:

$$BIC = n\ln(SSE) - n\ln(n) + p\ln(n) \tag{3.10}$$

Veja que assim como a medida AIC, quanto menor for SSE menor será o valor de BIC e quanto menor for p menor será o valor de BIC. Dessa forma também queremos modelos com medida BIC pequenas.

A partir do critério da Informação Bayesiana o modelo escolhido será aquele com menor valor de AIC.

3.4.2 Métodos de seleção passo-a-passo

O método da seleção passo-a-passo é recomendado para realizar a seleção do modelo quando o número de modelos possíveis é muito grande e a comparação entre todos eles passa a ser inviável. Podemos escolher entre dois métodos: inclusão progressiva de variáveis ou eliminação progressiva de variáveis. O primeiro começa com o modelo com

nenhuma variável preditiva e vai incluindo as variáveis, uma a uma, começando pela mais explicativa. O segundo começa com o modelo completo, com todas as variáveis preditivas, e vai eliminando as variáveis, uma a uma, começando pela menos explicativa.

Método da Inclusão Progressiva

Vamos definir o passo-a-passo desse método. Para isso suponha que temos N variáveis preditivas para fazer a seleção do modelo.

Passo 1) Ajuste os N modelos lineares simples para cada uma das N variáveis preditivas. Para cada modelo ajustado determine o p-valor do teste t, que defini se a variável em questão deve ou não ser incluída no modelo.

Passo 2) Se todos os p-valores forem maiores que α, nenhuma variável será incluída no modelo e FIM do algoritmo. Caso contrário, se algum p-valor for menor que α, inclua no modelo a variável preditiva referente ao menor p-valor. Essa é a variável mais explicativa para a variável resposta em questão. Vamos chamá-la de $x_{(1)}$.

Passo 3) Ajuste agora todos os $N-1$ modelos lineares com duas variáveis preditivas, sendo uma delas $x_{(1)}$. Para cada modelo ajustado determine o p-valor do teste t para a variável do modelo diferente de $x_{(1)}$, que defini se essa a variável deve ou não ser incluída no modelo.

Passo 4) Se todos os p-valores forem maiores que α, nenhuma variável nova será incluída no modelo, o modelo final será $y = \beta_0 + \beta_1 x_{(1)} + \varepsilon$ e FIM do algoritmo. Caso contrário, se algum p-valor for menor que α, inclua no modelo a variável preditiva referente ao menor p-valor. Vamos chamá-la de $x_{(2)}$.

Passo 5) Ajuste agora todos os $N-2$ modelos lineares com três variáveis preditivas, sendo duas delas $x_{(1)}$ e $x_{(2)}$. Para cada modelo ajustado determine o p-valor do teste t para a variável do modelo diferente de $x_{(1)}$ ou $x_{(2)}$, que defini se essa a variável deve ou não ser incluída no modelo.

Passo 6) Se todos os p-valores forem maiores que α, nenhuma variável nova será incluída no modelo, o modelo final será $y = \beta_0 + \beta_1 x_{(1)} + \beta_2 x_{(2)} + \varepsilon$ e FIM do algoritmo. Caso contrário, se algum p-valor for menor que α, inclua no modelo a variável preditiva referente ao menor p-valor. Vamos chamá-la de $x_{(3)}$. ...

O algoritmo continua até que todas as variáveis sejam incluídas no modelo ou até que todos os p-valores para a inclusão de uma nova variável sejam todos maiores que α.

Um recomendação para melhorar ainda mais esse algoritmo é, logo após a inclusão de uma nova variável no modelo, realizar um teste t com todas as variáveis já incluídas, considerando inclusive essa última que acabou se entrar, para verificar se alguma das variáveis já incluída deve ser retirada.

Método da Eliminação Progressiva

Vamos definir o passo-a-passo desse método. Para isso suponha novamente que temos N variáveis preditivas para fazer a seleção do modelo.

Passo 1) Ajuste o modelo linear completo com todas as N variáveis preditivas. Determine o p-valor do teste t para cada uma das N variáveis preditivas do modelo.

Passo 2) Se todos os p-valores forem menores que α, nenhuma variável será eliminada do modelo e FIM do algoritmo. Caso contrário, se algum p-valor for maior que α, elimine do modelo a variável preditiva referente ao maior p-valor. Vamos chamá-la de $x_{(1)}$.

Passo 3) Ajuste agora o modelo com todas as $N-1$ variáveis preditivas que restaram, isto é, todas as N variáveis iniciais menos $x_{(1)}$. Para o modelo ajustado determine o p-valor do teste t para cada uma das $N-1$ variáveis preditivas do modelo.

Passo 4) Se todos os p-valores forem menores que α, mais nenhuma variável será eliminada do modelo e FIM do algoritmo. Caso contrário, se algum p-valor for maior que α, elimine do modelo a variável preditiva referente ao maior p-valor. Vamos chamá-la de $x_{(2)}$.

Passo 5) Ajuste agora o modelo com todas as variáveis preditivas que restaram. Para o modelo ajustado determine o p-valor do teste t para cada uma das $N-2$ variáveis preditivas.

Passo 6) Se todos os p-valores forem menores que α, mais nenhuma variável será eliminada do modelo e FIM do algoritmo. Caso contrário, se algum p-valor for maior que α, elimine do modelo a variável preditiva referente ao maior p-valor. ...

O algoritmo continua até que todas as variáveis sejam eliminadas do modelo ou até que os p-valores do teste t sejam todos menores que α.

3.5 Resíduos e Pontos Influentes

Nessa última seção deste capítulo serão apresentadas mais algumas definições de resíduos em modelos lineares. Em seguida veremos a definição de pontos influentes e a sua importância na análise de regressão.

3.5.1 Resíduos Padronizados, Studentizados e Deletados

Já vimos na Equação 2.7 da Seção 2.11 a definição abaixo para o i-ésimo resíduo padronizado na regressão múltipla.

$$e_i^* = \frac{e_i}{\sqrt{MSE}} = \frac{y_i - \hat{y}_i}{\sqrt{MSE}}$$

Veremos agora a definição de mais dois outros resíduos.

Resíduo Studentizado

O resíduo padronizado e_i^* veio como uma tentativa de padronização do resíduo ordinário e_i, onde este último foi dividido por uma estimativa do seu desvio padrão considerando a aproximação $Var(e_i) = \sigma^2$. Já foi comentado, para o modelo simples na Seção 1.10, que $Var(e_i) \neq \sigma^2$ e por isso essa padronização considera aproximações para grandes amostras. Veremos agora a demonstração desse resultado.

Na forma matricial podemos escrever:

$$\underline{e} = (I - H)\underline{y} = (I - H)\underline{\varepsilon} \;\Rightarrow\; \underline{e} \sim N_n(\underline{\mu}, \Sigma)$$

onde,

$$\begin{aligned} \underline{\mu} &= E\left[\underline{e}\right] = E\left[(I - H)\underline{\varepsilon}\right] = (I - H)E\left[\underline{\varepsilon}\right] = 0 \\ \Sigma &= Var(\underline{e}) = Var\left((I - H)\underline{y}\right) = (I - H)^T Var(\underline{y})(I - H) \\ &= (I - H)\sigma^2 I(I - H) = \sigma^2(I - H)(I - H) = \sigma^2(I - H). \end{aligned}$$

Ou seja, $Var(e_i) = \sigma^2(1 - h_{i,i})$, onde $h_{i,i}$ é o elemento da posição (i, i) da matriz H. Por isso a padronização mais adequada é

$$r_i = \frac{e_i}{\sqrt{MSE(1 - h_{i,i})}} = \frac{y_i - \hat{y}_i}{\sqrt{MSE(1 - h_{i,i})}} \tag{3.11}$$

e r_i é chamado de Resíduo Studentizado.

Uma propriedade que ainda não foi citada é que $0 \leq h_{i,i} \leq 1 \;\forall\; i$. Ela é importante para garantirmos que $Var(e_i) \geq 0$. Veremos agora sua demonstração.

Proposição 3.5.1 *Seja $H = X(X^TX)^{-1}X^T$ a matriz Hat e $h_{i,i}$ o i-ésimo elemento da diagonal principal. Então,*

(i) $h_{i,i} = \sum_{j=1}^{n} h_{i,j}^2 = h_{i,i}^2 + \sum_{j \neq i} h_{i,j}^2$

(ii) $0 \leq h_{i,i} \leq 1$

Demonstração:

(i) Como H é idempotente temos que $H = H^2 = HH$. Logo a posição (i, i) da matriz H, $h_{i,i}$, é igual a posição (i, i) da matriz H^2, que é o produto interno entre o i-ésimo vetor linha da matriz H com o seu i-ésimo vetor coluna. Isto é:

$$h_{i,i} = \langle (h_{i,1}, h_{i,2}, \ldots, h_{i,n}), (h_{1,i}, h_{2,i}, \ldots, h_{n,i}) \rangle.$$

Como a matriz H é simétrica temos $h_{i,j} = h_{j,i}$ e por isso podemos reescrever a equação acima da seguinte forma:

$$h_{i,i} = \langle (h_{i,1}, h_{i,2}, \ldots, h_{i,n}), (h_{i,1}, h_{i,2}, \ldots, h_{i,n}) \rangle = \sum_{j=1}^{n} h_{i,j}^2 = h_{i,i}^2 + \sum_{j \neq i} h_{i,j}^2$$

e assim demonstra-se a afirmação (i).

(ii) Primeiro veja que $h_{i,i} > 0$ uma vez que $h_{i,i} = \sum_{j=1}^{n} h_{i,j}^2 > 0$. Veremos agora a outra desigualdade.

Veja que a partir da equação apresentada em (i) podemos escrever

$$h_{i,i} - h_{i,i}^2 = \sum_{j \neq i} h_{i,j}^2 \geq 0.$$

Suponha por absurdo que $1 < h_{i,i}$, nesse caso $h_{i,i} < h_{i,i}^2 \;\Rightarrow\; h_{i,i} - h_{i,i}^2 < 0$, chegando assim em uma contradição. Logo $h_{i,i} \leq 1$.

□

Resíduos Deletados

Um outra maneira de medir o resíduo de uma observação é a partir de diferença entre o valor observado y_i e o valor ajustado pelo modelo criado a partir da amostra original sem a observação i, vamos representar esse valor ajustado por $\hat{y}_{i(i)}$. A partir dessa ideia defini-se o *Resíduo Deletado* por:

$$d_i = y_i - \hat{y}_{i(i)} \tag{3.12}$$

A boa notícia é que não será preciso ajustar n modelos de regressão linear, um para cada observação que seria excluída, para encontrar todos os n resíduos deletados. A partir de transformações algébricas podemos chegar na seguinte equivalência:

$$d_i = \frac{e_i}{1 - h_{i,i}} \tag{3.13}$$

onde e_i é o resíduo ordinal da observação i e $h_{i,i}$ o i-ésimo elemento da matriz *Hat*, ambos definidos para o modelos ajustados com a amostra original.

Repare que sempre $d_i > e_i$ uma vez que $1 - h_{i,i} < 1$. Se temos $d_i \approx e_i$ significa que a eliminação da i-ésima observação da amostra praticamente não altere o ajuste, isto é o que esperamos de um modelo bem ajustado. Caso isso não aconteça, isto é, caso $d_i > e_i$, significa que a retirada da i-ésima observação altera as estimativas, o que indica que esse ponto tem alguma influência no modelo.

Como $d_i \approx e_i$ quando $h_{i,i} \approx 0$, $d_i > e_i$ quando $h_{i,i} > 0$ e quanto maior o valor de $h_{i,i}$ maior será a diferença entre d_i e e_i podemos concluir que a medida h indica não só os pontos que estão perto do centroide da amostra como também indica os pontos que se excluídos geram mudanças significativas no ajuste do modelo.

3.5.2 Pontos Influentes

Em Modelos Lineares uma observação é dita influente quanto a sua eliminação da amostra altera de forma significativa o modelo ajustado. Isso significa que a sua eliminação altera as estimativas para os parâmetros e, consequentemente, os resíduos, as previsões e até as interpretações.

A ideia geral é identificar esse(s) ponto(s) influente(s) e, depois de uma análise nos dados, decidir se ele pode ou não ser excluído da amostra. Isso porque buscamos um modelo que se ajuste bem na maioria da amostra e se um único ponto está prejudicando esse ajuste e a sua eliminação (se viável) pode melhorar a qualidade do ajuste nos demais pontos.

Existem diversas medidas para identificar os pontos influentes, veremos aqui a mais usada delas, a *Distância de Cook*, que é definida por:

$$D_i = \frac{\sum_{j=1}^{n} \left(\hat{y}_j - \hat{y}_{j(i)}\right)^2}{p\ MSE} \tag{3.14}$$

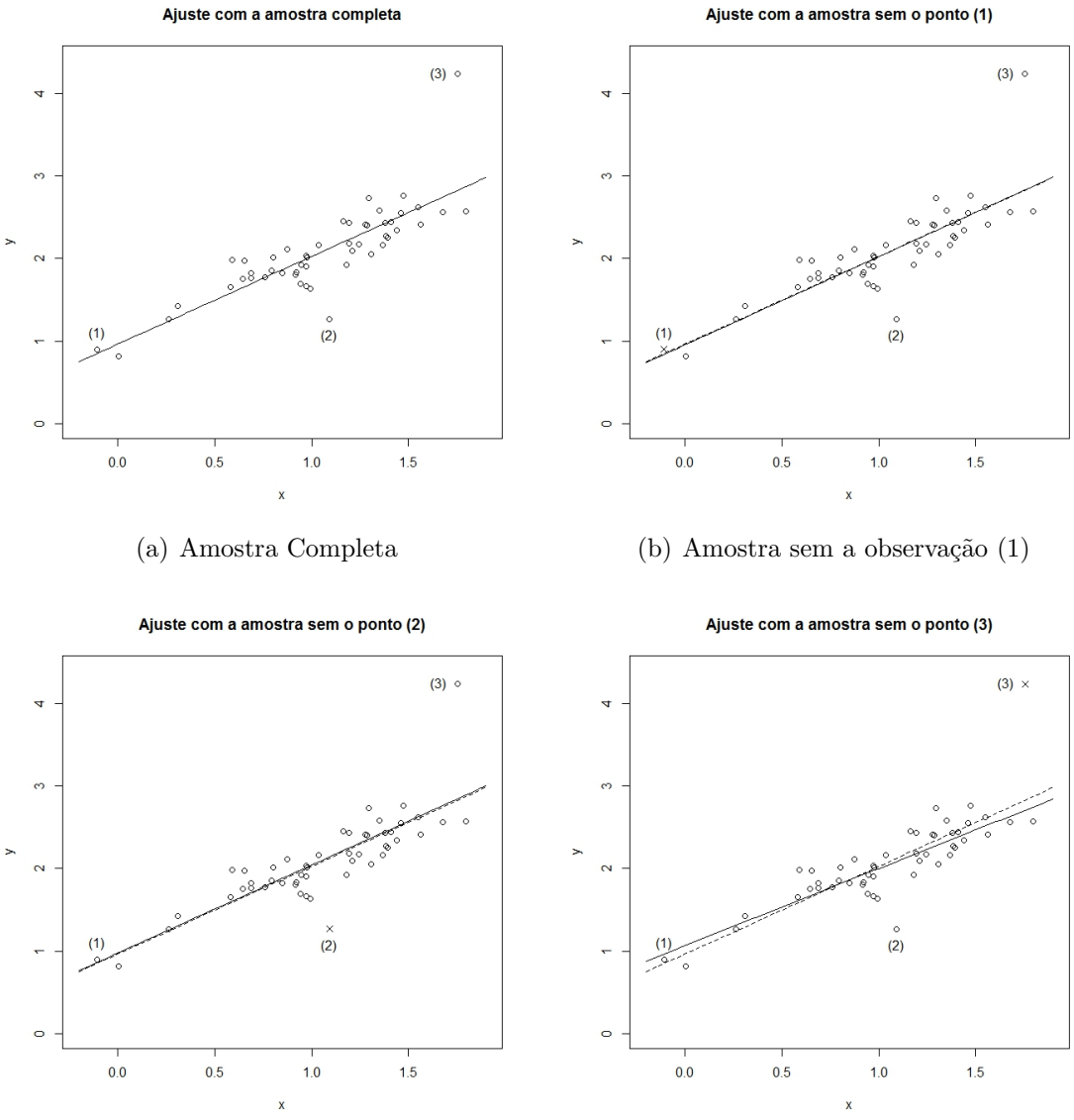
onde $\hat{y}_{j(i)}$ é o valor ajustado para a j-ésima observação considerando a amostra sem a observação i.

Aqui também não será preciso ajustar o modelo com e sem a observação i para encontrar todas as n Distancias de Cook. A partir de manipulações algébricas é possível chegar na seguinte expressão equivalente:

$$D_i = \frac{e_i^2}{p\ MSE}\left[\frac{h_{i,i}}{(1 - h_{i,i})^2}\right] = \frac{1}{p\ MSE}\ h_{i,i}\ d_i^2 \tag{3.15}$$

Veja na Equação 3.15 que o valor de D_i leva em consideração o resíduo deletado d_i e a medida h para a observação i, $h_{i,i}$. Valores grandes de d_i e de $h_{i,i}$ geram valores grandes de D_i, e quanto maior o valor de D_i maior será a influência da observação i no modelo ajustado. Dessa forma a Distância de Cook identifica como pontos influentes aqueles com valores grande para d_i e moderados para $h_{i,i}$, valores grandes para $h_{i,i}$ e moderados para d_i ou valores grandes tanto para d_i quanto $h_{i,i}$.

Como já comentado na Seção 2.10, a medida h indica os pontos distantes do centroide definido pelas variáveis preditivas, estes são pontos discrepantes em $\underline{x}$. Já valores grandes para o resíduo ordinal e_i indicam pontos discrepantes em y, estes são também os valores com grande resíduo deletado d_i.



(a) Amostra Completa

(b) Amostra sem a observação (1)

(c) Amostra sem a observação (2)

(d) Amostra sem a observação (3)

Figura 3.6: Pontos Influnetes

Veja na Figura 3.6(b) que o ponto (1) tem medida h grande e resíduo deletedo (e ordinal) bem pequeno. Já o ponto (2) tem medida h pequena e resíduo deletado (e ordinal) grande (Figura 3.6(c)). Por fim o ponto (3) tem tanto a medida h quanto o

resíduo deletado grande (Figura 3.6(d)).

As Figuras 3.6(b), 3.6(c) e 3.6(c) também apresentam a reta ajustada considerando a amostra completa (em pontilhado) e a reta ajustada retirando um dos pontos da amostra (reta contínua). Dessa forma podemos perceber que se os pontos (1) e (2) fossem retirados da amostra pouco mudaria a reta ajustada, inclusive ela está a'te difícil de ver pois está muito próxima da reta contínua. Isso já não acontece com o ponto (3), que quando retirado da amostra gera uma mudança significativa na reta ajustada. Podemos chegar nessa conclusão a partir dos gráficos da Figura 3.6(a) e também comparando os valores das estimativas para os coeficientes com a amostra completa e retirando cada um dos três pontos como mostra a Tabela 3.5.

Modelo	$\hat{\beta}_0$	$\hat{\beta}_1$
Amostra completa	0.9646864	1.0616810
Amostra sem (1)	0.9552241	1.0695892
Amostra sem (2)	0.9776282	1.0659651
Amostra sem (3)	1.0678464	0.9343609

Tabela 3.5: Tabela com estimativas para os parâmetros β_0 e β_1 com e sem cada um dos pontos (1), (2) e (3)

Pela Tabela 3.5 podemos ver que o ponto (3) exerce maior influência sobre as estimativas dos parâmetros β_0 e β_1.

Quais desses três pontos você acredita que representam pontos influentes de acordo com a Distância de Cook? A partir da análise feita na Figura 3.6(a) e na Tabela 3.5 acima já temos uma ideia da resposta, mas vejamos o valor da Distância de Cook para cada um desses três pontos na Tabela 3.6 abaixo, para comprovar nossa desconfiança.

Observação	Distância de Cook
(1)	0.003243441
(2)	0.08355043
(3)	1.000206

Tabela 3.6: Tabela com as Distâncias de Cook para as observações (1), (2) e (3)

Mas o quanto grande tem que ser o valor de D_i para que a observação i deva ser identificada como um ponto influente? Uma análise mais detalhada sugere considerar $D_i \sim F_{p,n-p}$, logo valores significativamente grande para D_i seriam aqueles maiores que o $1-\alpha$ quantil da $F_{p,n-p}$. Mas algumas referências recomendam simplesmente considerar pontos influentes aqueles que apresentarem Distância de Cook maior que 1. Essa última proposta pode ser a estratégia adotada, uma vez que é mais simples e mais conservadora que a anterior, pois considerando a distribuição F os valores de corte seriam na maioria das vezes maiores que 1.

3.6 Medidas Corretivas para Não-Linearidade

Nesta seção veremos como continuar usando o Modelo de Regressão Linear mesmo quando uma variável preditiva x_k não aparentar ter uma relação linear com a média da variável resposta y. Isto é, queremos é verificar a suposição de linearidade e saber o que fazer caso essa suposição não seja aceita.

Para isso primeiro ajuste o modelo simples considerando apenas a variável x_k e verifique, a partir do teste t, se podemos considerar adequado usar a variável preditiva x_k para explicar a variável resposta y. Lembre-se, se o p-valor for grande aceitamos a hipótese $H_0 : \beta_k = 0$ e por isso não iremos seguir adiante, pois o modelo linear não é adequado para relacionar as variáveis x_k e y. Caso contrário seguiremos com o modelo simples

$$y = \beta_0 + \beta_k x_k + \varepsilon.$$

3.6.1 Diagnóstico

O diagnóstico de não-linearidade já foi apresentado em 1.10.1, vamos continuar com esse mesmo procedimento: para cada variável x_k ajustamos o modelo linear simples e em seguida fazemos a análise dos resíduos a partir do gráfico de $x_{i,k}$ versus o r_i, i-ésima Resíduos Studentizados 3.11. Se os pontos do gráfico estiverem aleatoriamente distribuídos em torno do zero vamos aceitar a suposição de linearidade. Caso os pontos apresentem um dos padrões da Figura 1.7 vamos rejeitar a suposição de linearidade. Neste último caso temos o diagnóstico de não-linearidade para a variável x_k.

3.6.2 Medidas Corretivas - Transformação em x_k

Suponha que já tenha sido feito o diagnóstico de não-linearidade para a variável preditiva x_k. Isso significa que apesar dessa variável explicar a variável resposta y, pois estamos assumindo que o p-valor do teste t foi pequeno, a sua relação com $E[y]$ não é linear. Provavelmente isso ocorre pois os pontos (x_k, y) não estão em torno de uma reta e sim em torno de um curva, como mostra os gráficos da Figura 3.7.

Nesse caso a relação entre x_k e $E[y]$ não é linear, mas pode ser que seja linear a relação entre x'_k e $E[y]$, onde x'_k é uma transformação da variável x_k. Por exemplo, suponha que a real (porém desconhecida) relação entre x_k e y seja

$$y = \beta_0 + \beta_k x_k^2 + \varepsilon.$$

Nesse caso o gráfico de dispersão entre x_k e y será parecido com o gráfico da Figura 3.7(a) (se $\beta_k > 0$) ou com o gráfico da Figura 3.7(c) (se $\beta_k < 0$). Apesar de x_k e $E[y]$ não se relacionarem de forma linear a relação entre $x'_k = x_k^2$ e $E[y]$ é linear uma vez que podemos escrever

$$y = \beta_0 + \beta_k x_k^2 + \varepsilon \;\Rightarrow y = \beta_0 + \beta_k x'_k + \varepsilon.$$

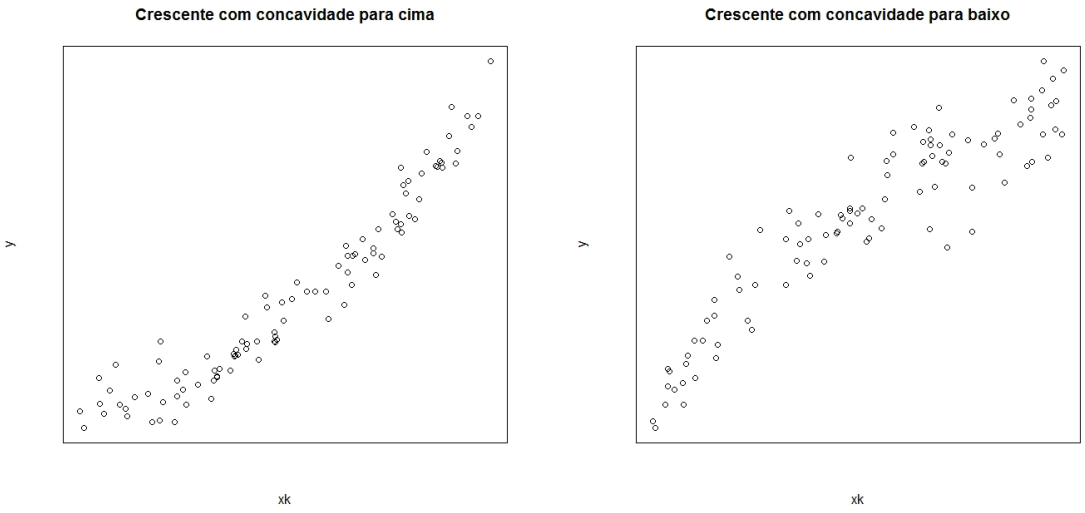
Nesse caso em vez de ajustar o modelo de regressão linear com a variável x_k faça o ajuste com $x'_k = x_k^2$.

Vejamos outro exemplo. Se a relação entre x_k e y for

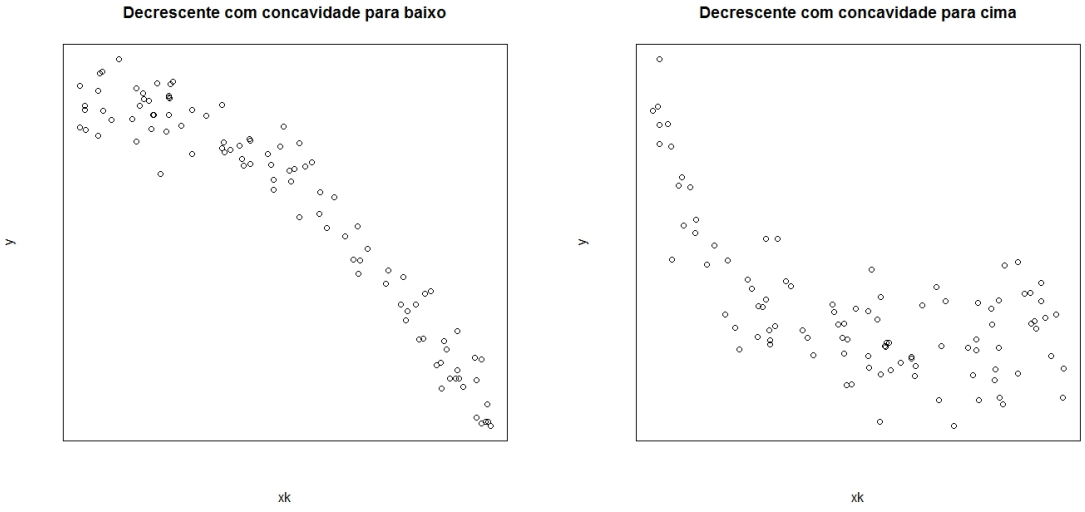
$$y = \beta_0 + \beta_k \ln(x_k) + \varepsilon$$

o gráfico de dispersão entre x_k e y será parecido com o gráfico da Figura 3.7(b) (se $\beta_k > 0$) ou com o gráfico da Figura 3.7(d) (se $\beta_k < 0$). Nesse caso novamente x_k e $E[y]$ não se relacionarem de forma linear, mas $x'_k = \ln(x_k)$ e $E[y]$ sim. Então o modelo linear deve ser ajustado com $x'_k = \ln(x_k)$ em vez de x_k:

$$y = \beta_0 + \beta_k \ln(x_k) + \varepsilon \;\Rightarrow y = \beta_0 + \beta_k x'_k + \varepsilon.$$



(a) x^2 e^x (b) $log(x)$ $\sqrt{x}$ $1/x$ e^{-x}



(c) x^2 e^x (d) $log(x)$ $\sqrt{x}$ $1/x$ e^{-x}

Figura 3.7: Gráfico de dispersão de x_k e y para diagnóstico de não-linearidade.

As vezes é difícil escolher de primeira a transformação mais adequada, apesar dos gráficos de dispersão da Figura 3.7 darem uma dica. Por isso muitas vezes acabamos tentando várias transformações e ficando com aquela que melhor resolver o problema de não-linearidade. E para saber qual melhor resolveu o problema vamos rodar o modelo simples com cada variável transformada e novamente fazer a análise dos resíduos. Vamos ficar com a transformação que mais minimizar o problema de não-linearidade, ou seja, a transformação para a qual o gráfico dos resíduos ficar o mais aleatório em torno de zero.

3.6.3 Algumas Observações

- O procedimento de diagnosticar e tratar do problema de não-linearidade deve ser feito para cada variável quantitativa x_k e, para cada diagnóstico de não-linearidade uma transformação deve ser escolhida.

- As transformações podem ser: $x'_k = x_k^2$, $x'_k = e^{x_k}$, $x'_k = \ln(x_k)$, $x'_k = e^{-x_k}$, $x'_k = \sqrt{x_k}$ ou $x'_k = 1/x_k$. Também podemos ser adicionadas contantes de deslocamento para melhorar ainda mais a correção do problema: $x'_k = (x_k + cte)^2$, $x'_k = e^{x_k+cte}$, $x'_k = \ln(x_k + cte)$, $x'_k = e^{-(x_k+cte)}$, $x'_k = \sqrt{x_k + cte}$ ou $x'_k = 1/(x_k + cte)$. Alguns cuidados na escolha da constante cte:
  - Não escolha uma constante que faça com que alguma $x_{i,k} + cte$ fique fora do domínio da função. Por exemplo, se escolhermos a transformação $x'_k = \ln(x_k + cte)$ a constante cte escolhida tem que ser tal que $x_{i,k} + cte > 0$ para todo ponto da amostra i.
  - Não escolha uma constante que faça com que os pontos $(x'_{i,k}, y_i)$ deixem de ter um comportamento crescentes ou decrescentes.
- Se houver poucas variáveis quantitativas faça o procedimento de diagnóstico e medidas corretivas antes da seleção do modelo e realize a seleção do modelo já com as variáveis transformadas. Caso haja muitas variáveis preditivas quantitativas, para simplificar, primeiro faça a seleção do modelo e só realize esse procedimento de diagnóstico e medidas corretivas nas variáveis preditivas quantitativas que ficarem no modelo final.
- Depois que o modelo foi ajustado considerando a variável x_k transformada cuidado com a expressão da função de regressão. Considerando a variável x'_k a função de regressão será linear, mas considerando a variável original não. Suponha que a transformação tenha sido $x'_k = \ln(x_k)$, então

$$E[y] = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x'_k + \ldots = \beta_0 + \beta_1 x_1 + \ldots + \beta_k \ln(x_k) + \ldots$$

- Cuidado também com a interpretação dos parâmetros β referentes à variáveis transformadas. Eles agora representam a mudança na média da variável resposta quando x'_k muda uma unidade e não para x_k. Por esse motivo quando uma transformação é feita na variável x_k em geral perde-se a interpretação do parâmetro β_k.
- Veja que dessa forma estamos usando o ajuste linear para variáveis que se relacionam de forma não-linear.

3.6.4 Modelo de Regressão Polinomial com uma Variável

Quando os pontos $(x_{i,k}, y_i)$ apresentarem um comportamento não só não-linear mas também aparentemente polinomial podemos optar por usar o Modelo de Regressão Polinomial para para encontrar a curva que melhor se ajusta aos pontos e assim definir um bom modelo de predição para essa amostra. O Modelo de Regressão Polinomial de grau g com uma variável preditiva é definido por:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \ldots + \beta_g x_i^g + \varepsilon_i. \tag{3.16}$$

Em geral costuma-se usar o grau pequeno, $g = 2$ ou 3. Também é muito comum usar o seguinte modelo alternativo:

$$y_i = \beta_0 + \beta_1 (x_i - \bar{x}) + \beta_2 (x_i - \bar{x})^2 + \beta_3 (x_i - \bar{x})^3 + \ldots + \beta_g (x_i - \bar{x})^g + \varepsilon_i.$$

onde $\bar{x}$ é a média amostral da variável preditiva x.

Quando queremos encontrar a curva polinomial de grau g que melhor se ajusta aos pontos estamos buscando as estimativas dos parâmetros β para o modelo polinomial. Uma alternativa bem simples de resolver esse problema é transformar esse modelo polinomial em uma variável para um modelo linear múltiplo (em várias variáveis) e então usar os estimadores de mínimos quadrados do modelo linear para estimar os β's do modelo polinomial.

Isso pode ser feito simplesmente a partir de transformações na variável preditiva x. Veja como. Primeiro faça as transformações

$$x'_{i,1} = x_i,\ x'_{i,2} = x_i^2,\ x_{i,3} = x_i^3, \ldots$$

Assim o Modelo Polinomial da Equação 3.16 pode ser reescrito por:

$$y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,3} + \ldots + \beta_g x_{i,g} + \varepsilon_i,$$

ou seja, por um modelo linear múltiplo. E assim podemos usar o que já aprendemos para encontrar as estimativas dos parâmetros do modelo e então fazer previsões e análises.

Algumas Observações

- Os testes de hipóteses do modelo linear múltiplo podem ser usados para verificar se algum monômio x^k deve ou não ser utilizado no modelo.
- Podemos optar por usar o Modelo de Regressão Polinomial mesmo quando o teste t para o modelo linear simples com a variável x tenha sido grande. Nesse caso concluímos que o ajuste linear não é adequado, mas talvez o ajuste polinomial seja.
- Nesse caso a função de regressão não será mais uma reta e sim uma curva de regressão.
$$E[y] = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \ldots + \beta_g x^g,$$
- A decisão de usar ou não o Modelo de Regressão Polinomial pode ser tomada a partir de uma análise do gráfico de dispersão das variáveis x e y.
- Podemo generalizar o Modelo de Regressão Polinomial para mais de uma variável, mas nesse caso devemos incluir os termos cruzados e o modelos erá bem maior. Por exemplo, a expressão para duas variáveis com grau 2 é definida por:
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta 22 x_2^2 + \beta_{12} x_1 x_2 + \varepsilon_i.$$

3.7 Medidas Corretivas para Heterocedasticidade

Nesta seção veremos que medidas tomar quando for diagnosticada a heterocedasticidade, isto é, a variância não constante dos erros.

3.7.1 Diagnóstico

O diagnóstico de Heterocedasticidade será feito como na Seção 2.11. Depois de feita a seleção das variáveis preditivas ajuste o modelo completo e encontre os resíduos (Ordinário e_i ou Studentizados r_i). Em seguida analise os gráficos: (i) $\hat{y}_i$ versus e_i; (ii) $\hat{y}_i$ versus $|e_i|$.

Se os pontos nos gráficos apresentarem um padrão constante, aceitamos a suposição de homocedasticidade. Se algum dos padrões das Figuras 1.9 ou 1.10 for encontrado não aceitamos a suposição de variância constante e diagnosticamos a heterocedasticidade.

Podemos também usar o Teste de *Breusch-Pagan* para ajudar no diagnóstico de heterocedasticidade. Lembrando que a hipótese nula deste teste é H_0 : variância constante, ou seja, p-valor grande para o teste indica variância constante e p-valor pequeno rejeitamos a hipótese de variância não constante, isto é, diagnosticamos a heterocedasticidade.

3.7.2 Medidas Corretivas - Mínimos Quadrados Ponderados

Quando diagnosticada a heterocedasticidade vamos tomar a seguinte medida corretiva: a estimativa dos parâmetros será feita de forma a não considerar mais variância constante. Ou seja, vamos encontrar o estimador de mínimos quadrados para o modelo definido por:

$$y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \ldots + \varepsilon_i \ , \quad \varepsilon_i \sim N(0, \sigma_i^2) \quad \text{independentes.} \tag{3.17}$$

Nesse caso o vetor de erros $\underline{\varepsilon}$ tem distribuição $N_n(\underline{0}, \Sigma)$ com

$$\Sigma = \begin{pmatrix} \sigma_1^2 & 0 & 0 & \ldots & 0 \\ 0 & \sigma_2^2 & 0 & \ldots & 0 \\ 0 & 0 & \sigma_3^2 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & \sigma_n^2 \end{pmatrix} \tag{3.18}$$

Primeiro será encontrado o novo estimador para $\underline{\beta}$ supondo que as variâncias σ_i^2 são diferentes para cada i e conhecidas. É claro que esse é um cenário irreal, na prática não conhecemos a matriz Σ, mas esse caso mais simples vai ser importante para chegamos na solução do caso real, quando as variâncias são diferentes e desconhecidas.

Variâncias Conhecidas

Vamos agora encontrar os estimadores de máxima verossimilhança para $\underline{\beta}$ da Equação 3.17, supondo que Σ é conhecida, ou seja, σ_i^2 serão constantes para a função de verossimilhança. Veja que nesse caso temos:

$$L(\underline{\beta}) = \prod_{i=1}^{n} f_{y_i}(\underline{\beta}|\underline{y}, X) = \prod_{i=1}^{n} \frac{1}{(2\pi\sigma_i^2)^{1/2}} \exp\left[-\frac{1}{2\sigma_i^2}(y_i - \beta_0 - \beta_1 x_{i,1} - \ldots - \beta_{p-1} x_{i,p-1})^2\right]$$

Vamos definir os pesos $w_i = \frac{1}{\sigma_i^2}$ e trocar σ_i^2 por $\frac{1}{w_i}$ na expressão acima. Assim temos:

$$L(\underline{\beta}) = \left[\prod_{i=1}^{n} \left(\frac{w_i}{2\pi}\right)^{1/2}\right] \exp\left[-\frac{1}{2}\sum_{i=1}^{n} w_i(y_i - \beta_0 - \beta_1 x_{i,1} - \ldots - \beta_{p-1} x_{i,p-1})^2\right]$$

Veja que estamos supondo $\prod_{i=1}^{n} \left(\frac{w_i}{2\pi}\right)^{1/2}$ constante conhecida. Então encontrar o $\underline{\beta}$ que minimiza L é o mesmo que encontrar o $\underline{\beta}$ que maximiza

$$Q = \sum_{i=1}^{n} w_i(y_i - \beta_0 - \beta_1 x_{i,1} - \ldots - \beta_{p-1} x_{i,p-1})^2$$

Veja que a Q também pode ser escrito como a soma dos quadrados dos resíduos do modelo definido por

$$W\underline{y} = WX\underline{\beta} + \underline{\varepsilon} \qquad \text{com} \qquad W = \begin{pmatrix} w_1 & 0 & 0 & \dots & 0 \\ 0 & w_2 & 0 & \dots & 0 \\ 0 & 0 & w_3 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & w_n \end{pmatrix}$$

Então a solução do problema de minimização será a solução do sistema

$$W\underline{y} = WX\underline{\beta} \;\Rightarrow\; X^TW\underline{y} = X^TWX\underline{\beta}$$

Dessa forma, o estimador para $\underline{\beta}$ por mínimos quadrados supondo variância não constante e conhecida é definido por

$$\underline{\hat{\beta}_w} = (X^TWX)^{-1}X^TW\underline{y}. \tag{3.19}$$

Além disso, supondo variância não constante $\underline{y} \sim N_n(X\underline{\beta}, \Sigma)$, com Σ definida em 3.18. Podemos então encontrar a distribuição de $\underline{\beta}$. Primeiro veja que $\underline{\beta}$ continua sendo um vetor aleatório com distribuição Normal p-Variada, uma vez que ele é combinação linear de $\underline{y}$. Vamos agora encontrar o seu vetor de média e a sua matriz de variância.

$$\begin{aligned} E\left[\underline{\hat{\beta}_w}\right] &= E\left[(X^TWX)^{-1}X^TW\underline{y}\right] \\ &= (X^TWX)^{-1}X^TWE\left[\underline{y}\right] \\ &= (X^TWX)^{-1}X^TWX\underline{\beta} = \underline{\beta} \end{aligned}$$

Para encontrar a matriz de variância antes veja que $W\Sigma = I$, que $W^T = W$ e que X^TWX é simétrica, logo $(X^TWX)^{-1}$ também é. Vamos as contas.

$$\begin{aligned} Var\left(\underline{\hat{\beta}_w}\right) &= Var\left((X^TWX)^{-1}X^TW\underline{y}\right) \\ &= (X^TWX)^{-1}X^TWVar\left(\underline{y}\right)WX(X^TWX)^{-1} \\ &= (X^TWX)^{-1}X^TW\Sigma WX(X^TWX)^{-1} \\ &= (X^TWX)^{-1}X^TWX(X^TWX)^{-1} = (X^TWX)^{-1} \end{aligned}$$

Então chegamos no resultado,

$$\underline{\hat{\beta}_w} \sim N_p(\underline{\beta}, (X^TWX)^{-1})$$

Veja que se voltarmos para o caso da variância constante temos $W = \frac{1}{\sigma^2}I$ e o estimado para $\underline{\beta_w}$ será o mesmo definido em 2.

Variâncias Desconhecidas

Se as variâncias σ_i^2 fossem conhecidas os pesos seriam definidos por $w_i = 1/\sigma_i^2$. Quando isso não acontecer vamos definir os pesos por $w_i = 1/\hat{s}_i^2$, onde $\hat{s}_i^2$ será uma estimativa para σ_i^2. Veja agora como essas estimativas serão encontradas.

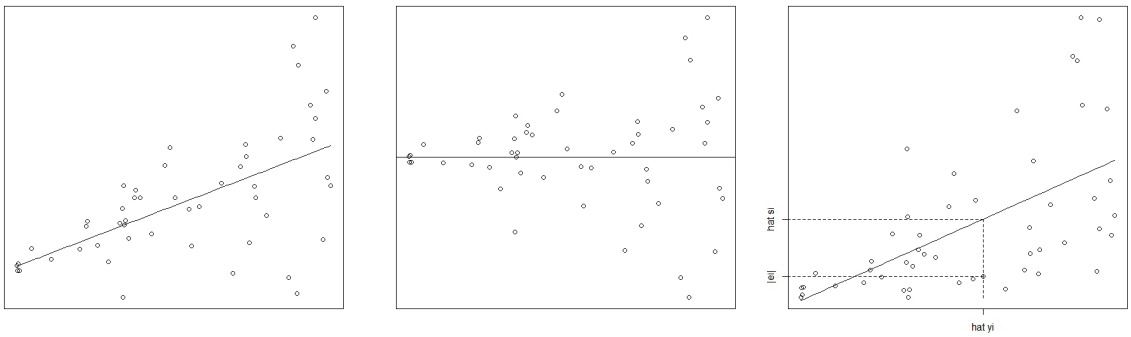
Supondo variância não constante temos $\varepsilon_i \sim N(0, \sigma_i^2)$ e por isso podemos escrever

$$\mathrm{E}[\varepsilon_i^2] = \sigma_i^2 \quad \text{e} \quad \mathrm{E}[|\varepsilon_i|] = \sigma_i.$$

Logo encontrar estimativas para σ_i^2 é o mesmo que encontrar estimativas para $\mathrm{E}[\varepsilon_i^2]$ e encontrar estimativas para σ_i é o mesmo que encontrar estimativas para $\mathrm{E}[|\varepsilon_i|]$. Veja que é razoável considerar que uma boa estimativa para $\mathrm{E}[e_i^2]$ também é uma boa estimativa para $\mathrm{E}[\varepsilon_i^2]$ e consequentemente uma boa estimativa para σ_i^2. Da mesma forma, uma boa estimativa para $\mathrm{E}[|e_i|]$ também será uma boa estimativa para $\mathrm{E}[|\varepsilon_i|]$ e consequentemente uma boa estimativa para σ_i. O que vamos fazer encontrar estimativas para $\mathrm{E}[|e_i|]$ e defini-las como $\hat{s}_i$.

Suponha que ao fazer o gráfico de $\hat{y}_i$ versus e_i (ou $x_{i,k}$ versus e_i) encontramos um formato de megafone. Nesse caso podemos fazer os gráficos de $\hat{y}_i$ versus $|e_i|$ (ou $x_{i,k}$ versus $|e_i|$) e perceber um padrão crescente. Se um modelo de regressão linear for ajustado a esses pontos teremos a equação da reta ajustada por esse modelo e assim uma forma de prever $\mathrm{E}[|e_i|]$ em função de $\hat{y}_i$ (ou de $x_{i,k}$). Vamos definir $\hat{s}_i$ como sendo o valor ajustado por essa reta de regressão e este será a estimativa para σ_i. A estimativa para σ_i^2 será $\hat{s}_i{}^2$ e o peso w_i definido por $w_i = 1/\hat{s}_i{}^2$.

A Figura ilustra como encontrar $\hat{s}_i$ a partir do modelo de regressão linear definido pelos pontos $(\hat{y}_i, |e_i|)$.



(a) $x_i \times y_i$ (b) $\hat{y}_i \times e_i$ (c) $\hat{y}_i \times |e_i|$

Figura 3.8: Como estimar $\hat{s}_i$ a partir da relação entre $\hat{y}_i$ e $|e_i|$

Quando $\underline{\beta}$ for estimado por mínimos quadrados ponderados teremos que fazer algumas modificações nas inferências do modelo. Todas elas serão baseadas nas novas distribuições amostrais:

$$\underline{\beta_w} \sim N_p(\underline{\beta}, (X^T W X)^{-1}) \tag{3.20}$$

$$\hat{y}_h \sim N(\underline{x_h}^T \underline{\beta_w}, \underline{x_h}^T (X^T W X)^{-1} \underline{x_h}) \tag{3.21}$$

Então o intervalo de confiança para cada β_k será

$$\hat{\beta}_{kw} \pm t_{1-\frac{\alpha}{2}, n-p} \sqrt{C_{k+1,k+1}}$$

onde $C_{k+1,k+1}$ é o elemento da posição diagonal de $(X^T W X)^{-1}$ e $\hat{\beta}_{kw}$ é a k-ésima posição do vetor de estimadores $\underline{\hat{\beta}_w}$ por mínimos quadrados ponderados.

Já o intervalo de confiança para a média da variável resposta dado um nível x_h será definido por:

$$\hat{y}_h \pm t_{1-\frac{\alpha}{2}, n-p} \sqrt{\underline{x_h}^T (X^T W X)^{-1} \underline{x_h}}$$

onde o valor $\hat{y}_h = \underline{x_h}^T \underline{\hat{\beta}_w}$ é o valor ajustado para o nível x_h considerando o estimador por mínimos quadrados.

Além disso os resíduos padronizados e a matriz *Hat* também mudam. A nova matriz *Hat* será definida por

$$H = X(X^T W X)^{-1} X^T W$$

e dessa forma continuamos podendo escrever

$$\underline{e} = (I - H)\underline{y} = (I - H)\underline{\varepsilon} \;\Rightarrow\; \underline{e} \sim N_n(\underline{0}, \Sigma(I - H))$$

onde Σ está definida em 3.18, mas também poderia ser definida por $\Sigma = W^{-1}$.

Resumindo, aqui está um passo-a-passo referente aos procedimentos quando detectada heterocedasticidade.

passo 1) Ajuste o modelo de regressão linear para a variável resposta y e as variáveis preditivas x_k's e encontre o estimador $\underline{\hat{\beta}}$ por mínimos quadrados ordinários (não ponderados). Defina os resíduos ordinários $\underline{e}$ e os valores ajustados $\hat{y}$ desse ajuste.

passo 2) Ajuste o modelo de regressão linear simples considerando como variável resposta $|e|$ e variável preditiva $\hat{y}$ (ou algum x_k). Os valores ajustados para esse modelo será denominado $\underline{\hat{s}}$.

passo 3) Defina os pesos $w_i = 1/\hat{s}_i^2$ e faça um novo o ajuste do modelo de regressão linear para a variável resposta y e as variáveis preditivas x_k's, considerando agora o estimador $\underline{\hat{\beta}_w}$ por mínimos quadrados ponderados.

Se as estimativas por mínimos quadrados ponderados ($\underline{\hat{\beta}_w}$) for bem diferente das estimativas por mínimos quadrados ordinários ($\underline{\hat{\beta}}$) o processo pode ser repetido. Para isso encontre novas estimativas para os pesos w_i, agora usando os resíduos do último modelo ajustado (já com as estimativas por mínimos quadrados ponderados). Geralmente uma ou duas iterações são suficientes para estabilizar o modelo. Esse processo iterativo é chamado de Método Iterativo de Mínimos Quadrados Ponderados.

A função `lm` do R tem a opção de trabalhar com mínimos quadrados ponderados. Para isso basta informar o vetor de pesos w_i.

Exercícios para Aulas Práticas do Capítulo 3

1. Uma empresa opera duas linhas de produção de sabonetes. Para cada linha de produção a relação entre a velocidade da linha e a quantidade de resíduos produzidos foi analisada a partir de um modelo linear. Os dados referentes à essa análise estão disponíveis em `CH08TA05.txt` (`https://netfiles.umn.edu/users/nacht001/www/nachtsheim/5th/KutnerData/Chapter%20%208%20Data%20Sets/CH08TA05.txt`). Nesse arquivo a primeira coluna contém os valores para os resíduos produzidos (y), a segunda coluna contém os valores das velocidades das linhas de produção (x_1) e a terceira coluna indica se a produção é referente à linha 1 ou 2 (x_2).

   (a) Primeiro faça o gráfico de dispersão das variáveis x_1 e y. Em seguida ajuste esses dados ao modelo de regressão linear simples

   $$y_i = \beta_0 + \beta_1 x_{i,1}.$$

   Junto com o gráfico de dispersão desenhe a reta de regressão estimada.

   (b) Faça agora o gráfico de dispersão das variáveis x_1 e y de forma que os pontos referentes à cada linha de produção sejam diferenciados.

   (c) Agora ajuste os dados ao modelo de regressão linear definido por:

   $$y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \varepsilon_i.$$

   Em seguida desenhe a reta de regressão estimada para cada linha de produção junto com o gráfico de dispersão do item (1b).

   (d) Considerando o modelo do item (1c), qual seria o teste apropriado para saber se as duas linhas de produção se comportam de forma semelhante? Determine o valor da estatística de teste e do p-valor. Em seguida tome uma decisão.

   (e) Agora ajuste os dados ao modelo de regressão linear definido por:

   $$y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,1} x_{i,2} + \varepsilon_i.$$

   Em seguida desenhe a função de regressão estimada para cada linha de produção junto com o gráfico de dispersão do item (1b).

   (f) Considerando agora o modelo do item (1e), qual seria o teste apropriado para saber se as duas linhas de produção se comportam de forma semelhante? Determine o valor da estatística de teste e do p-valor. Em seguida tome uma decisão.

   (g) Como podemos verificar se a taxa de crescimento da produção média de resíduos em função da velocidade é a mesma nas duas linhas de produção? Defina a hipótese a ser testada. Qual o teste apropriado para verificar essa hipótese? Determine o valor da estatística de teste e do p-valor. Em seguida tome uma decisão.

   (h) Como você escolheria sobre a inclusão ou não do termo cruzado? Isto é, como você escolheria entre o modelo do item (1c) ou o modelo do item (1e)?

2. Nesse exercício vamos usar os dados apresentada no arquivo `CH09TA01.txt` (`https://netfiles.umn.edu/users/nacht001/www/nachtsheim/5th/KutnerData/Chapter%`

20%209%20Data%20Sets/CH09TA01.txt). Esses dados foram recolhidos por uma unidade cirúrgica de um hospital a fim de investigar o tempo de vida de pacientes que forma submetidos a algum tipo de cirurgia vital. As colunas do banco de dados são referentes às seguintes variáveis.

coluna 1) nível de coagulação do sangue (x_1)

coluna 2) índice do prognóstico (x_2)

coluna 3) teste de enzima (x_3)

coluna 4) teste de função hepática (x_4)

coluna 5) idade, em anos (x_5)

coluna 6) sexo (0 = homem, 1 = mulher) (x_6)

coluna 7 e 8) variáveis indicadoras para identificar o uso de álcool dos pacientes (00 = nenhum, 10 = moderado, 01 = grande) (x_7 e x_8)

coluna 9) tempo de vida (y_1)

coluna 10) índice que mede o tempo de vida: $y = \ln(y_1)$ (y)

Estamos interessados em ajustar o modelo linear para explicar a variável y em função das variáveis $x_1, \ldots, x_8$.

(a) Usando o comando `pairs` veja o gráfico de dispersão de cada par de variáveis preditivas. Quais variáveis parecem bem correlacionadas a partir do gráfico?

(b) Verifique se há multicolinearidade nos dados. Caso haja, tome as medidas recomendadas para tratar desse problema.

(c) Faça a seleção de variáveis a partir da comparação entre todos os possíveis modelos. Para facilitar considere apenas as variáveis preditivas x_1, x_2, x_3 e x_5. Utilize as três medidas comparativas vistas em sala de aula: R^2_a, AIC e BIC.

(d) Faça a seleção de variáveis (agora com todas) a partir do Método da Inclusão Progressiva.

(e) Faça a seleção de variáveis (agora com todas) a partir do Método da Eliminação Progressiva.

(f) Interprete as estimativas pontuais e intervalares para os parâmetro do modelo final. Qual a variável mais explicativa para o índice de tempo de vida?

(g) Discuta sobre a inclusão somente da variável x_8 e a não inclusão da variável x_7. O que isso significa? Interprete nesse caso a estimativa do parâmetro β referente a variável x_8.

(h) Analise se o termo cruzado para a variável x_8 deve ser incluída no modelo.

(i) Encontre os Resíduos Studentizados e faça o gráfico dos Resíduos Studentizados versus cada variável preditiva x_k e versus o valor ajustado $\hat{y}$ e comente sobre as suposições de linearidade e homocedasticidade.

(j) Encontre as Distância de Cook e faça o gráfico de D_i versus o índice i. Comente sobre a existência ou não de pontos influentes.

3. Os dados do arquivo `CH07TA01.txt` (`https://netfiles.umn.edu/users/nacht001/www/nachtsheim/5th/KutnerData/Chapter%20%207%20Data%20Sets/CH07TA01.txt`) apresentam informações de 20 pacientes mulheres, saudáveis e com idade entre 25-34 anos de num consultório de nutrição. A primeira coluna informa a medida da prega cutânea tricipital em milímetros (x_1), a segunda coluna a medida da circunferência da coxa em centímetros (x_2), a terceira coluna a medida da circunferência do braço em centímetros (x_3) e a quarta coluna o percentual de gordura (y). Para medir o percentual de gordura os pacientes passaram por um procedimento complicado e caro que exige a imersão das pessoas na água.

   Estamos interessados em ajustar um modelo linear para explicar o percentual de gordura y em função das medidas x_1, x_2 e x_3. Se isso for possível teremos uma alternativa simples e barata para encontrar uma estimativa para o percentual de gordura dos pacientes apenas utilizando as medidas x_1, x_2 e x_3, que facilmente podem ser encontradas com equipamentos simples dentro dos consultórios.

   (a) Analise e comente sobre a existência ou não de multicolinearidade. Tome as medidas cabíveis para resolver o problema de multicolinearidade.

   (b) Estamos interessados em encontrar o modelo de regressão para explicar y em função de x_2 e x_3. Primeiro faça o gráfico de dispersão das variáveis x_2 e x_3 e tente identificar no desenho os pontos que parecem ter medida h grande.

   (c) Encontre agora a medida h de cada observação e identifique no gráfico os pontos $(x_{i,2}, x_{i,3})$ com medida h maior que $2p/n$.

   (d) Encontre os Resíduos Studentizados e faça o gráfico dos Resíduos Studentizados versus cada variável preditiva x_k e versus o valor ajustado $\hat{y}$ e comente sobre as suposições de linearidade e homocedasticidade.

   (e) Encontre os Resíduos Deletados, faça o gráfico desses valores versus os índices e identifique as observações com maior valor para esse resíduo. A partir dessa resposta junto com a resposta do item 3c quais observações você acredita que terá as maiores Distâncias de Cook?

   (f) Encontre a Distância de Cook, faça o gráfico desses valores versus os índices e identifique as observações com maior Distância de Cook. Comente sobre a existência de pontos influentes.

   (g) A partir do modelo final encontre uma estimativa intervalar para o percentual de gordura médio de pacientes com 50 cm de circunferência da coxa e 28 cm de circunferência do braço.

4. O arquivo `CH03TA07.txt` (`https://netfiles.umn.edu/users/nacht001/www/nachtsheim/5th/KutnerData/Chapter%20%203%20Data%20Sets/CH03TA07.txt`) apresenta os dados de um experimento que busca relacionar o desempenho de vendedores com o número de dias de treino recebidos em situações de venda simuladas. A primeira coluna se refere ao número de dias de treino (x) e a segunda ao desempenho (y) de cada um dos 10 vendedores que passaram pelo experimento.

   (a) Ajuste um modelo de regressão linear para as variáveis x e y e faça o diagnóstico de não-linearidade. Não deixe de analisar os seguintes gráficos: (i) gráfico de dispersão de x e y junto com a reta de regressão estimada; (ii) gráfico dos resíduos studentizados versus x.

(b) Crie uma nova variável $x' = \sqrt{x}$. Agora ajuste um novo modelo de regressão linear para as variáveis x' e y e veja se o problema de não-linearidade foi resolvido. Não deixe de analisar os seguintes gráficos: (i) gráfico de dispersão de x' e y junto com a reta de regressão estimada; (ii) gráfico dos resíduos studentizados versus x'.

(c) Considerando o modelo do item 4b, encontre a função de regressão estimada para a variável original x. Apresente em seguida o gráfico de dispersão de x e y junto com a função de regressão estimada definida pelo modelo do item 4b.

(d) Considerando o modelo do item 4b encontre:

- Uma estimativa pontual para o desempenho de vendedores com 2 dias de treinamento.
- Um intervalo de confiança para a média do desempenho de vendedores com 2 dias de treinamento.
- Um intervalo de predição para o desempenho de um novo vendedor com 2 dias de treinamento.

5. Um grupo de pesquisadores em saúde está interessado em estudar a relação entre a pressão arterial diastólica e a idade de pacientes saudáveis. Participaram desse estudo 54 mulheres saudáveis entre 20 e 60 anos de idade. Os dados recolhidos estão apresentados no arquivo `CH11TA01.txt` (`https://netfiles.umn.edu/users/nacht001/www/nachtsheim/5th/KutnerData/Chapter%2011%20Data%20Sets/CH11TA01.txt`) onde a primeira coluna representa a idade (x) e a segunda coluna a pressão arterial diastólica (y) das pacientes.

(a) Ajuste um modelo de regressão linear entre as variáveis x e y. Encontre as estimativas por mínimos quadrados ordinários para os parâmetros β's e em seguida faça o diagnostico de heterocedasticidade. Não deixe de analisar os seguintes gráficos: (i) gráfico de dispersão de x e y junto com a função de regressão estimada pelo modelo; (ii) gráfico dos resíduos ordinários versus $\hat{y}$; (iii) gráfico do módulo dos resíduos ordinários versus $\hat{y}$.

(b) Ajuste um modelo de regressão linear simples entre as variáveis $|e|$ e $\hat{y}$. Adicione ao gráfico (iii) do item anterior a reta de regressão estimada desse último ajuste. Encontre as estimativas $\hat{s}_i$ e em seguida defina o vetor de pesos w.

(c) Ajuste um novo modelo de regressão linear entre as variáveis x e y, agora considerando os pesos w para encontrar as estimativas por mínimos quadrados ponderados.

(d) Compare as estimativas para $\underline{\beta}$ por mínimos quadrados ponderados, encontradas no item 5c, com as estimativas por mínimos quadrados ordinários, encontradas no item 5a. Compare também os intervalos de confiança para β_1 nos dois modelos.

(e) Considerando o modelo encontrado no item 5c encontre: (i) uma estimativa pontual para a pressão arterial em mulheres com 40 anos; (ii) um intervalo de confiança para a média da pressão arterial em mulheres com 40 anos; (iii) um intervalo de predição para a pressão arterial uma nova paciente com 40 anos.

Lista de Exercícios do Capítulo 3

3.1. O diretor de uma pequena universidade selecionou 120 alunos do primeiro semestre de forma aleatória para um estudo que buscava determinar se o *GPA - grade point average* (equivalente ao CR, mas varia de 0-4) dos alunos ao final do primeiro semestre (y) pode ser previsto pelo *ACT test score - American College Testing* (equivalente à nota do ENEM, mas a nota máxima é 36) (x_1). Além dessas duas notas a amostra também apresenta uma outra variável x_2 que indica se o aluno indicou ou não uma área de concentração no momento da inscrição. Se aluno tiver indicado uma área $x_2 = 1$, caso contrário $x_2 = 0$.

A amostra referente a esse estudo encontra-se nos arquivos `CH01PR19.txt` (`https://netfiles.umn.edu/users/nacht001/www/nachtsheim/5th/KutnerData/Chapter%20%201%20Data%20Sets/CH01PR19.txt`) (variáveis y e x_1) e `CH08PR16.txt` (`https://netfiles.umn.edu/users/nacht001/www/nachtsheim/5th/KutnerData/Chapter%20%208%20Data%20Sets/CH08PR16.txt`). (variável x_2). Para esse problema considere o modelo de regressão

$$y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \varepsilon.$$

(a) Explique como cada coeficiente do modelo de regressão pode ser interpretado nesse problema.

(b) Ajuste os dados para o modelo acima e encontre a função de regressão estimada.

(c) Faça o gráfico de dispersão entre as variáveis y e x_1 de forma que você consiga distinguir os pontos referentes aos alunos que indicaram e que não indicaram uma área de concentração na hora da inscrição. Junto desse gráfico desenhe as retas de regressão estimada para cada caso.

(d) Teste se podemos considerar semelhante a relação entre o GPA e o ACT dos alunos que indicaram e não indicaram uma área de concentração no momento da inscrição.

3.2. Este exercício continua com os dados do Exercício 1 acima. Mas agora considere o modelo

$$y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,1} x_{i,2} + \varepsilon_i.$$

(a) Ajuste os dados para o modelo acima e encontre a função de regressão estimada.

(b) Faça o gráfico de dispersão entre as variáveis y e x_1 de forma que você consiga distinguir os pontos referentes aos alunos que indicaram e que não indicaram uma área de concentração na hora da inscrição. Junto desse gráfico desenhe as retas de regressão estimada para cada caso.

(c) Teste se o termo cruzado deve ou não ser retirado do modelo. Use $\alpha = 0.05$. Se este termo não puder ser retirado do modelo, interprete o seu efeito no problema em questão.

(d) Independente da resposta do item acima continue com o modelo definido no enunciado. Teste se podemos considerar semelhante a relação entre o GPA e o ACT dos alunos que indicaram e não indicaram uma área de concentração no momento da inscrição.

3.3. Seja x_1 uma variável quantitativa e x_2 uma variável indicadora, que representa alguma variável qualitativa de duas classes. Considere os dois modelos de regressão a seguir:

$$y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \varepsilon_i$$
$$y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,1} x_{i,2} + \varepsilon_i$$

A conclusão de que $\beta_2 = 0$ tem a mesma implicação nos dois modelos acima? Explique.

3.4. Em uma análise de regressão para dados de sobre acidentes de trabalho causados por queda de objetos dentro de um armazém seja y a variável que define uma medida sobre a gravidade do acidente, x_1 um índice que indica de forma simultânea o peso do objeto e a distância da qual ele caiu, e x_2 e x_3 variáveis indicadoras para representar se o trabalhador usava alguma proteção na cabeça no instante do acidente.

Tipo de proteção	x_2	x_3
Capacete de segurança	1	0
Protetor anti-impacto	0	1
Nenhuma proteção	0	0

A diferença entre o capacete de proteção e o protetor anti-impacto é que o primeiro segue as normas de segurança do trabalho.

A função de regressão a ser usada nesse estudo é $E[y] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$.

(a) Desenvolva a função de regressão para cada uma das três categorias de proteção.

(b) Para cada uma das perguntas a seguir especifique quais hipóteses H_0 e H_1 devem ser testadas a fim de se chegar a uma conclusão.

i. Considerando x_1 fixo no modelo, usar a proteção anti-impacto reduz a gravidade esperada do acidente quando comparado com a gravidade esperado sem proteção alguma?

ii. Considerando x_1 fixo no modelo, a gravidade esperada do acidente é a mesma quando usado o capacete de segurança ou o protetor anti-impacto?

3.5. Considere um modelo onde y é o desgaste de uma ferramenta e x_1 a velocidade em que esta foi empregada. Além disso cada ferramente pode ser classificada de acordo com seu modelo: M1, M2, M3 e M4. Para incluir essas categorias no modelos de regressão serão criadas três variáveis indicadoras:

$$X_2 = \begin{cases} 1, & \text{se a ferramente é do modelo M2} \\ 0, & \text{caso contrário} \end{cases}$$

$$X_3 = \begin{cases} 1, & \text{se a ferramente é do modelo M3} \\ 0, & \text{caso contrário} \end{cases}$$

$$X_4 = \begin{cases} 1, & \text{se a ferramente é do modelo M4} \\ 0, & \text{caso contrário} \end{cases}$$

Dessa forma o modelo de regressão de primeira ordem é definido por:

$$y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,3} + \beta_4 x_{i,4} + \varepsilon_i.$$

Baseado no modelo acima, indique o significado de cada parâmetro a seguir no problema em questão.

(a) β_3

(b) $\beta_4 - \beta_3$

(c) β_1

3.6. Como processo de admissão em uma agência governamental 25 candidatos foram submetidos à quatro diferentes testes de aptidão e suas notas registradas. Propositalmente os 25 candidatos foram aceitos, independente das notas obtidas nos quatro testes. Depois de um período probatório cada candidato foi avaliado de acordo com a sua proficiência no trabalho. As quatro notas dos testes (x_1, x_2, x_3 e x_4) assim como a avaliação com relação à proficiência no trabalho (y) de cada candidato se encontram no arquivo `CH09PR10.txt` (`https://netfiles.umn.edu/users/nacht001/www/nachtsheim/5th/KutnerData/Chapter%20%209%20Data%20Sets/CH09PR10.txt`), onde a primeira coluna indica o valor de y e as outras quatro colunas os valores das variáveis preditivas x_1 x_2, x_3 e x_4.

Analise se existe ou não multicolinearidade quando consideradas as quatro variáveis preditivas. Caso haja, tome as medidas cabíveis para corrigir esse problema antes de realizar a seleção do modelo.

3.7. Continue trabalhando com os dados do exercício 3.6. Utilize a comparação entre todos os possíveis modelos e decida quais variáveis selecionar para modelo de regressão linear. Quantos modelos você vai ter que ajustar?

(a) Faça a seleção do modelo utilizando como medida de comparação o valor de R^2_a. Como há um pequena diferença entre o R^2_a dos melhores modelos, quais outros critérios você pode usar para comparação?

(b) Faça a seleção do modelo utilizando como medida de comparação a informação de Akaike (AIC).

(c) Faça a seleção do modelo utilizando como medida de comparação a informação Bayesiana (BIC)

OBS: (i) No R já existe as funções `AIC` e `BIC`, elas devem ser aplicadas nos objetos da classe `lm`; (ii) Use o critério gráfico para ajudá-lo nas comparações.

3.8. Continue trabalhando com os dados do exercício 3.6.

(a) Faça a seleção do modelo utilizando o Método da Inclusão Progressiva. Use $\alpha = 0.05$ para incluir variáveis e $\alpha = 0.10$ para excluir (considerando que após a inclusão de uma nova variável você vai verificar se alguma das variáveis já incluídas deve ser eliminada).

(b) Faça a seleção do modelo utilizando o Método da Eliminação Progressiva. Use $\alpha = 0.05$.

3.9. Continue trabalhando com os dados do exercício 3.6.

Nesse exercício vamos estudar com detalhes o modelo formado pelas variáveis preditivas X_1 e X_3. Então antes de começar os itens a seguir ajuste os dados ao modelo

$$y = \beta_0 + \beta_1 x_1 + \beta_3 x_3 + \varepsilon.$$

(a) Obtenha os Resíduos Studentizados e faça o gráfico deles contra: $\hat{y}_i$ e cada uma das quatro variáveis preditivas. Analise os gráficos e comente sobre as suposições de linearidade e homocedasticidade.

(b) Faça o gráfico `qqnorm` para os Resíduos Studentizados e comente sobre a suposição de normalidade dos erros.

(c) Encontre a medida h para cada observação. Faça o gráfico de dispersão das variáveis preditivas x_1 e x_3 e identifique os pontos com medida h maior que $2p/n$.

(d) Encontre a Distância de Cook para cada observação. Em seguida faça o gráfico da Distância de Cook contra o índice da observação. O que você pode dizer sobre a existência de pontos influentes?

3.10. Continue trabalhando com os dados do exercício 3.1 e considere o modelo apresentado no exercício 3.2 com o termo cruzado:

$$y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,1} x_{i,2} + \varepsilon_i.$$

(a) Faça o gráfico de dispersão entre as variáveis y e x_1 de forma que seja possível identificar os pontos com $x_2 = 1$ e os com $x_2 = 0$. Junto com o gráfico desenhe as retas de regressão estimada pelo modelo. Obs: este é o mesmo gráfico do exercício 3.2(a).

(b) Obtenha os Resíduos Studentizados. Identifique no gráfico do item (a) os pontos com Resíduos Studentizados maiores que 2 e menores que -2.

(c) Faça os gráficos adequados para analisar as suposições de linearidade e homocedasticidade.

(d) Faça o gráfico `qqnorm` para os Resíduos Studentizados e comente sobre a suposição de normalidade dos erros.

(e) Encontre a medida h para cada observação. Identifique no gráfico do item (a) os pontos com medida h maior que $2p/n$.

(f) Encontre a Distância de Cook para cada observação. Faça o gráfico da Distância de Cook contra o índice da observação. O que você pode dizer sobre a existência de pontos influentes? Identifique no gráfico do item (a) os pontos com maior Distância de Cook, mesmo que não sejam pontos influentes.

3.11. Este é uma continuação do Exercício 7 das aulas práticas do Capítulo 1. Em uma manufatura deseja-se estudar a relação entre o tamanho dos lotes produzidos e o tempo gasto em sua produção. Para isso 111 lotes de tamanhos diferentes forma produzidos e tempo gasto em sua produção anotado, esses valores podem ser encontrados em `CH03PR18.txt` (`https://netfiles.umn.edu/users/nacht001/www/nachtsheim/5th/KutnerData/Chapter%20%203%20Data%20Sets/CH03PR18.txt`), onde a primeira coluna é o tempo de produção de cada lote (y) e na segunda o tamanho do lote (x).

a) Ajuste um modelo de regressão linear para as variáveis x e y e faça o diagnóstico de não-linearidade. Não deixe de analisar os seguintes gráficos: (i) gráfico de dispersão de x e y junto com a reta de regressão estimada; (ii) gráfico dos resíduos studentizados versus x.

b) Escolha uma transformação x' de x e ajuste agora um novo modelo de regressão linear para as variáveis x' e y de forma a minimizar (ou até resolver) o problema de não-linearidade. Não deixe de analisar os seguintes gráficos: (i) gráfico de dispersão de x' e y junto com a reta de regressão estimada; (ii) gráfico dos resíduos studentizados versus x'.

c) Expresse a função de regressão estimada pelo modelo do item 3.11b na variável original do problema (x). Faça também o gráfico de dispersão de x e y e junto com ele a curva desta função.

d) Considerando o modelo do item 3.11b encontre:

- Uma estimativa pontual para o tempo de produção de lotes de tamanho 10.
- Um intervalo de confiança para a média do tempo de produção de lotes de tamanho 10.
- Um intervalo de predição para o tempo de produção de um novo lote de tamanho 10.

3.12. Um endocrinologista está interessado em explorar a relação entre o nível de esteroide (y) e a idade dos pacientes (x). Para esse isso 27 pacientes mulheres saudáveis entre 8 e 25 anos tiveram seu nível de esteroide medidos. Os dados desse estudo podem ser encontrados no arquivo `CH08PR06.txt` (`https://netfiles.umn.edu/users/nacht001/www/nachtsheim/5th/KutnerData/Chapter%20%208%20Data%20Sets/CH08PR06.txt`), onde a primeira coluna guarda os valores de y e a segunda os valores de x.

a) Faça o gráfico de dispersão das variáveis x e y e comente sobre a relação entre essas variáveis.

b) Ajuste os dados no modelo quadrático $y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i$ e encontre as estimativas para $\underline{\beta}$.

c) Expresse a função de regressão estimada pelo modelo do item 3.12b na variável x. Adicione ao gráfico de dispersão de x e y a curva desta função.

d) Teste se existe ou não alguma relação entre x e y de acordo com o modelo proposto. Enuncie as hipóteses, a regra de decisão, o p-valor e a conclusão para cada teste. Interprete o resultado.

e) Teste se o termo quadrático deve ou não ser retirado do modelo proposto. Enuncie as hipóteses, a regra de decisão, o p-valor e a conclusão para cada teste. Interprete o resultado.

f) A parir do modelo ajustado encontre os intervalos de confiança para os níveis médio de esteroide em mulheres com 10, 15 e 20.

g) Encontre os resíduos ordinários e resíduos studentizados. Faça o gráfico de x versus os resíduos (ordinários ou studentizados) e em seguida o gráfico de $\hat{y}$ versus os resíduos (ordinários ou studentizados). Faça também o gráfico `qqnorm` para os resíduos studentizados. Comente o gráfico.

3.13. O número de peças defeituosas produzidas em uma máquina (y) se relaciona de forma linear com a velocidade dessa máquina (x). Os dados no arquivo `CH11PR07.txt` (`https://netfiles.umn.edu/users/nacht001/www/nachtsheim/5th/KutnerData/Chapter%2011%20Data%20Sets/CH11PR07.txt`) se referem aos registros obtidos pelo setor de qualidade. A primeira coluna se refere a variável y e a segunda a x.

a) Ajuste um modelo de regressão linear entre as variáveis x e y. Encontre as estimativas por mínimos quadrados ordinários para os parâmetros β's e em seguida faça o diagnostico de heterocedasticidade.

b) Seja e os resíduos ordinários e $\hat{y}$ os valores ajustados obtidos pelo último modelo ajustado. Ajuste um modelo de regressão linear simples entre as variáveis $|e|$ e $\hat{y}$. Encontre as estimativas $\hat{s}_i$ e em seguida defina o vetor de pesos w.

c) Ajuste um novo modelo de regressão linear entre as variáveis x e y, agora considerando os pesos w e as estimativas por mínimos quadrados ponderados.

d) Compare as estimativas para $\underline{\beta}$ por mínimos quadrados ponderados (item 3.13c) com as estimativas por mínimos quadrados ordinários (item 3.13a). Compare também os intervalos de confiança para β_1 nos dois modelos.

e) Repita novamente os itens 3.13b, 3.13c e 3.13d. Houve uma mudança significativa nas estimativas dos parâmetros β's? Se sim, o que você deve fazer?

3.14. Uma empresa de RH está interessada em estudar a relação entre o salário de um funcionário (y - coluna 1) e as variáveis: título, classificado por: 1=bacharel, 2=mestre e 3=doutor (coluna 2); anos de experiência desde o último título (coluna 3); número de funcionários atualmente sob a sua supervisão (coluna 4). Tais informações referentes a 65 funcionários de uma mesma empresa estão disponíveis no arquivo `CH11PR08.txt` (`https://netfiles.umn.edu/users/nacht001/www/nachtsheim/5th/KutnerData/Chapter%2011%20Data%20Sets/CH11PR08.txt`).

a) Ajuste um modelo de regressão linear considerando todas as variáveis preditivas do problema. Não deixe de criar as variáveis indicadoras necessárias. Encontre as estimativas por mínimos quadrados ordinários e em seguida verifique a suposição de homocedasticidade. Você deve concluir que há heterocedasticidade.

b) Seja e os resíduos ordinários e $\hat{y}$ os valores ajustados obtidos pelo último modelo ajustado. Ajuste um modelo de regressão linear simples entre as variáveis $|e|$ e $\hat{y}$. Encontre as estimativas $\hat{s}_i$ e em seguida defina o vetor de pesos w_i.

c) Ajuste novamente um modelo de regressão linear considerando todas as variáveis preditivas do problema, agora considerando os pesos w_i para encontrar as estimativas por mínimos quadrados ponderados.

d) Compare as estimativas para $\underline{\beta}$ por mínimos quadrados ponderados (item 3.14c) com as estimativas por mínimos quadrados ordinários (item 3.14a).

e) Repita novamente os itens 3.14b, 3.14c e 3.14d. Houve uma mudança significativa nas estimativas dos parâmetros β's? Se sim, o que você deve fazer?

f) Considerando o último modelo ajustado por mínimos quadrados ponderados encontre uma estimativa pontual para o salário de funcionários com metrado, 5 anos de experiência desde o último título e sem funcionários sob sua supervisão. Em seguida encontre um intervalo de confiança para a média do salário de funcionários com esse mesmo perfil.

Capítulo 4

Alguns Modelos Lineares Generalizados

4.1 Regressão Logística

Nessa seção veremos um modelo adequado para o caso em que a variável resposta é binária, isto é, quando y só pode assumir os valores 0 ou 1. Veja que nesse caso o Modelo de Regressão Linear Normal, estudado nos capítulos anteriores, não é mais apropriado, pois não podemos considerar que a distribuição de probabilidade de y seja a distribuição Normal. A solução será criar uma nova relação entre a variável resposta e as variáveis preditivas.

4.1.1 Modelo de regressão para variável resposta binária

É muito comum surgir o interesse na relação entre uma variável binária (variável resposta) e outras variáveis (preditivas). Veja alguns exemplos.

Exemplo 4.1.1 *Suponha que estejamos interessados em analisar a participação de mulheres casadas no mercado de trabalho, ou seja, queremos entender quais fatores (características) aumentam as chances das mulheres casadas entrarem no mercado de trabalho. Para isso foi recolhido uma amostra com as seguintes informações de algumas mulheres casadas: idade (x_1), número de filhos (x_2), renda do marido (x_3) e se ela está ou não no mercado de trabalho (y). Veja que y é uma variável binária. A questão nesse estudo é entender como as variáveis preditivas influenciam na probabilidade de uma mulher casada estar no mercado de trabalho, ou seja, entender como podemos encontrar $E[y_i] = P(y_i = 1) = \pi_i$, probabilidade da $i-$ésima mulher casada estar no mercado de trabalho, em função das variáveis preditivas $x_{i,1}$, $x_{i,2}$ e $x_{i,3}$.*

Exemplo 4.1.2 *Suponha que estejamos interessados em analisar/encontrar alguns fatores de risco para uma determinada doença, ou seja, queremos encontrar fatores que aumentam ou diminuam a chance de um individuo ter a doença. Para realizar esse estudo foram entrevistadas pessoas saudáveis ($y = 0$) e com a doença ($y = 1$) de forma a se obter as seguintes informações de cada indivíduo: idade (x_1), sexo (x_2), nível de colesterol (x_3), hábito de fumar (x_4), entre outros fatores. A questão nesse estudo é entender quais desses fatores de risco (x_1, x_2, x_3, x_4, ...) são importantes para aumentar ou diminuir a chance de um individuo ter a doença, ou seja, como encontrar $E[y_i] = P(y_i = 1) = \pi_i$, probabilidade do $i-$ésimo indivíduo ter a doença, em função de $x_{i,1}$, $x_{i,2}$, $x_{i,3}$, $x_{i,4}$...*

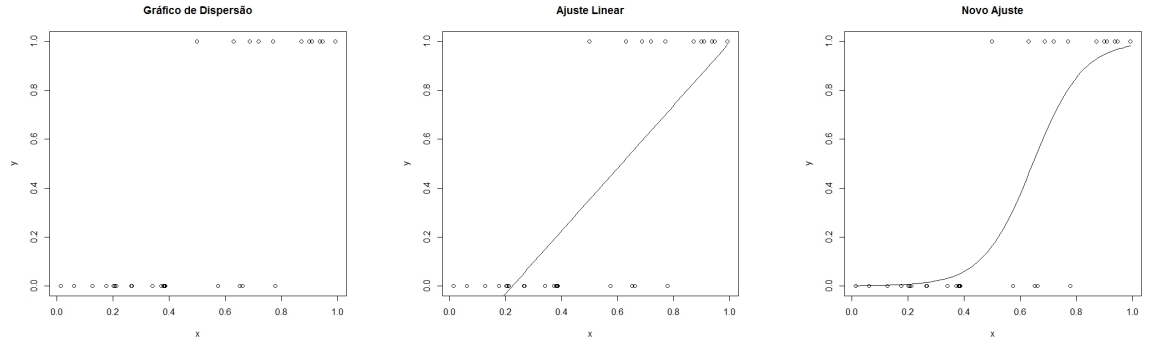
Veja que nesses dois exemplos a variável resposta y é binária, isto é, $y = 0$ ou $y = 1$, e por isso não podemos supor que a sua distribuição seja Normal. Então o modelo Normal não é mais adequado e vamos supor

$$y_i \sim Bernoulli(\pi_i).$$

Veja que o nosso interesse continua sendo analisar a relação entre a média da variável respostas y_i e as variáveis preditivas $\underline{x}_i$. Mas agora $y_i \sim Bernoulli(\pi_i)$, logo $E[y_i] = \pi_i$, ou seja, queremos estudar a relação entre π_i e as variáveis preditivas $\underline{x}_i$.

Além de y não ser mais uma variável aleatória Normal a relação entre $E[y_i] = \pi_i$ e as variáveis preditivas não pode mais ser considerada linear. Veja o Exemplo 4.1.3 a seguir, que ilustra essa afirmação.

Exemplo 4.1.3 *Suponha que queremos estudar como a média de uma variável resposta binária y_i varia com a variação de uma variável preditiva x_i (quantitativa neste exemplo). Suponha que os valores de y_i e x_i sejam aquelas apresentados no gráfico de dispersão da Figura 4.1(a). O gráfico da Figura 4.1(b) apresenta a reta de regressão estimada considerando o Modelo de Regressão Linear para os pontos (x_i, y_i). O gráfico da Figura 4.1(c) apresenta a curva de regressão estimada para o novo modelo que vamos aprender nesse capítulo.*



(a) Gráfico de Dispersão (b) Ajuste Linear (c) Novo Ajuste

Figura 4.1: Gráfico de Dispersão para Variável Resposta Binária

Alguns comentários relevantes sobre o gráfico. Primeiro veja na Figura 4.1(a) que $P(y = 1) = E[y_i] = \pi_i$ aumenta conforme x cresce. Isso nos faz acreditar que a variável preditiva x exerce influencia na variável resposta y.

Veja na Figura 4.1(b) que a reta de regressão estimada não está bem ajustada aos pontos. Além disso, se usássemos essa reta para realizar previsões a média de y seria negativa quando $x_h \approx 0$, isto é, se $x_h \approx 0$ então $\hat{\pi}_h < 0$, o que não faz sentido uma vez que $0 \leq \pi_h \leq 1$ qualquer que seja o x_h.

Já o ajuste (não linear) apresentado na Figura 4.1(c) parece bem mais adequado do que o ajuste linear. Além de nesse ajuste o valor de $\hat{\pi}_i$ aumentar conforme x_i cresce sempre teremos $0 \leq \hat{\pi}_i \leq 1$, o que torna este novo modelo bastante apropriado para resolver o problema em questão.

Então, em vez de supor uma relação linear entre a média de y e x vamos escolher outra relação, que tenha um gráfico parecido com o apresentado na Figura 4.1(c). Existem

algumas alternativas para definir f com esse padrão de gráfico, veremos aqui uma delas, a função logística.

4.1.2 A Função Logística

A Equação 4.1 apresenta a definição da Função Logística e a Figura 4.2 o seu gráfico.

$$\begin{array}{rccl} f: & \mathbb{R} & \to & [0,1] \\ & x & \mapsto & f(x) = \frac{1}{1+e^{-x}} \end{array} \tag{4.1}$$

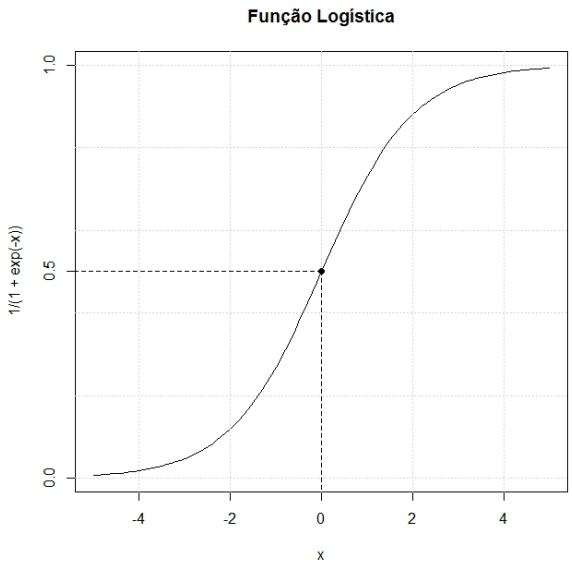


Figura 4.2: Gráfico da Função Logística

Veja que se a variável resposta for binária essa é uma boa alternativa para modelar a relação entre $E[y] = P(y = 1) = \pi$ e x, pois a imagem de f é o intervalo $[0, 1]$. Além disso a Função Logística trás algumas outras vantagens, como por exemplo, poder trabalhar com a razão de chance, como veremos em breve. Antes de apresentar o Modelo Logístico vejamos algumas propriedades da Função Logística que serão úteis mais a frente.

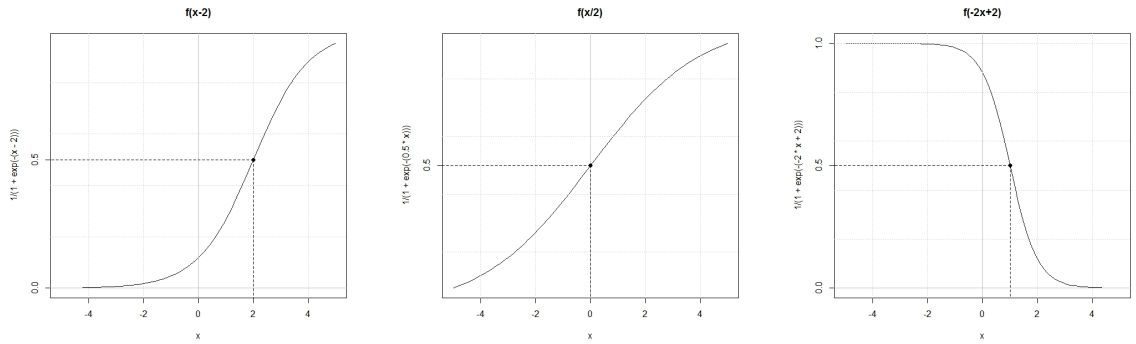
$$f(x) = \frac{1}{1+e^{-x}} = \frac{e^x}{e^x+1} \tag{4.2}$$

$$f'(x) = \frac{e^{-x}}{(1+e^{-x})^2} = \frac{1}{1+e^{-x}}\frac{e^{-x}}{1+e^{-x}} = f(x)(1-f(x)) \tag{4.3}$$

$$f^{-1}(x) = \ln\left(\frac{x}{1-x}\right), \quad 0 < x < 1 \tag{4.4}$$

Combinação da função logística com uma transformação linear em x pode mudar o padrão do gráfico, mas continua valendo $\lim_{x\to\pm\infty} f(x) = 0$ ou 1. Veja na Figura 4.3 alguns exemplos dos gráficos de $f(ax + b)$.

A constante b gera um deslocamento horizontal no gráfico da função logística. Já a constante a modifica a amplitude da curva. Se $a < 0$ a função muda de crescente para decrescente.



(a) $f(x-2)$ (b) $f(x/2)$ (c) $f(-2x+2)$

Figura 4.3: Variações nos gráficos de $f(ax+b)$.

4.1.3 O Modelo Logístico

O modelo de regressão que vamos trabalhar quando a variável y for binária será o Modelo Logístico. Nele supomos que a relação entre $E[y_i] = \pi_i$ e x_i é definida pela Função Logística. Vamos primeiro definir o modelo simples, para uma variável preditiva, e depois generalizar para o modelo múltiplo, com $p-1$ variáveis preditivas.

Definição 4.1.4 *Seja $y_i \sim Bernoulli(\pi_i)$, o Modelo Logístico Simples define*

$$y_i = E[y_i] + \varepsilon_i$$

e supõe a seguinte relação entre $E[y_i] = \pi_i$ e a variável preditiva x_i:

$$\pi_i = \frac{1}{1+e^{-(\beta_0+\beta_1 x_i)}} \qquad ou \qquad \ln\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_i.$$

Nosso objetivo é estimar β_0 e β_1 de forma a melhor ajustar o modelo aos dados. Veja que definimos $\pi_i = f(\beta_0 + \beta_1 x)$, então os valores de β_0 e β_1 são responsáveis por ajustar a curva melhor aos pontos (x_i, y_i).

Definição 4.1.5 *Seja $y_i \sim Bernoulli(\pi_i)$, o Modelo Logístico Múltiplo define*

$$y_i = E[y_i] + \varepsilon_i$$

e supõe a seguinte relação entre $E[y_i] = \pi_i$ e as $p-1$ variáveis preditivas $\underline{x}_i = (x_{i,1}, x_{i,2}, \ldots, x_{i,p-1})$:

$$\pi_i = \frac{1}{1+e^{-\underline{x}_i^T \underline{\beta}}} = \frac{1}{1+e^{-(\beta_0+\beta_1 x_{i,1}+\beta_2 x_{2,i}+\ldots+\beta_{p-1}x_{i,p-1})}}$$

$$ou \quad \ln\left(\frac{\pi_i}{1-\pi_i}\right) = \underline{x}_i^T \underline{\beta} = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{2,i} + \ldots + \beta_{p-1} x_{i,p-1}$$

O estimador adotado para $\underline{\beta}$ será o de máxima verossimilhança. Vamos encontrá-lo.

$$\begin{aligned} L(\underline{\beta}|\underline{y},\underline{x}) &= \prod_{i=1}^{n} \pi_i^{y_i}(1-\pi_i)^{1-y_i} \\ l(\underline{\beta}|\underline{y},\underline{x}) &= \ln\left(\prod_{i=1}^{n} \pi_i^{y_i}(1-\pi_i)^{1-y_i}\right) = \sum_{i=1}^{n} \ln(\pi_i^{y_i}(1-\pi_i)^{1-y_i}) \end{aligned}$$

$$
\begin{aligned}
&= \sum_{i=1}^{n} \left(\ln(\pi_i^{y_i}) + \ln((1-\pi_i)^{1-y_i})\right) = \sum_{i=1}^{n} \left(y_i \ln(\pi_i) + (1-y_i)\ln(1-\pi_i)\right) \\
&= \sum_{i=1}^{n} \left(y_i \left(\ln(\pi_i) - \ln(1-\pi_i)\right) + \ln(1-\pi_i)\right) \\
&= \sum_{i=1}^{n} y_i \left(\ln(\pi_i) - \ln(1-\pi_i)\right) + \sum_{i=1}^{n} \ln(1-\pi_i) \\
&= \sum_{i=1}^{n} y_i \ln\left(\frac{\pi_i}{1-\pi_i}\right) + \sum_{i=1}^{n} \ln(1-\pi_i)
\end{aligned}
$$

Assim chegamos na seguinte equação para a função de log-verossimilhança.

$$
l(\underline{\pi}|\underline{y},\underline{x}) = \sum_{i=1}^{n} y_i \ln\left(\frac{\pi_i}{1-\pi_i}\right) + \sum_{i=1}^{n} \ln(1-\pi_i) \tag{4.5}
$$

onde π_i é definido na Definição 4.1.5. Se quisermos escrever a função de log-verossimilhança em função de β devemos apenas substituir π_i e seguir com as contas.

$$
\begin{aligned}
l(\underline{\beta}|\underline{y},\underline{x}) &= \sum_{i=1}^{n} y_i \left(\underline{x}_i^T \underline{\beta}\right) + \sum_{i=1}^{n} \ln\left(1 - \frac{1}{1+e^{-\underline{x}_i^T \underline{\beta}}}\right) = \sum_{i=1}^{n} y_i \left(\underline{x}_i^T \underline{\beta}\right) + \sum_{i=1}^{n} \ln\left(\frac{e^{-\underline{x}_i^T \underline{\beta}}}{1+e^{-\underline{x}_i^T \underline{\beta}}}\right) \\
&= \sum_{i=1}^{n} y_i \left(\underline{x}_i^T \underline{\beta}\right) + \sum_{i=1}^{n} \ln\left(\frac{1}{1+e^{\underline{x}_i^T \underline{\beta}}}\right) = \sum_{i=1}^{n} y_i \left(\underline{x}_i^T \underline{\beta}\right) - \sum_{i=1}^{n} \ln\left(1+e^{\underline{x}_i^T \underline{\beta}}\right)
\end{aligned}
$$

Como alternativa, temos a expressão de l em função de $\underline{\beta}$

$$
l(\underline{\beta}|\underline{y},\underline{x}) = \sum_{i=1}^{n} y_i \left(\underline{x}_i^T \underline{\beta}\right) - \sum_{i=1}^{n} \ln\left(1+e^{\underline{x}_i^T \underline{\beta}}\right) \tag{4.6}
$$

O que faríamos agora é derivar em relação a cada β_k e igualar a zero para encontrar o ponto de máximo de l, que será o estimador de máxima verosimilhança. Mas isso não é possível, não existe solução analítica para o seu ponto de máximo, apesar de existir um ponto de máximo. Ou seja, não temos uma expressão para $\hat{\underline{\beta}}$ em função da amostra. Nesse caso, para cada amostra será usado um método iterativo para encontrar uma aproximação para a estimativa de máxima verossimilhança de $\underline{\beta}$. Programas como o R já tem esses métodos iterativos implementados e por isso não teremos problema para encontrar as estimativas $\hat{\underline{\beta}}$.

Dada a estimativa $\hat{\underline{\beta}}$, a estimativa pontual para a média da variável resposta, nesse caso para π_i, é dada por:

$$
\hat{\pi}_i = \frac{1}{1+e^{-\underline{x}_i^T \hat{\underline{\beta}}}} = \frac{1}{1+e^{-(\hat{\beta}_0 + \hat{\beta}_1 x_{i,1} + \hat{\beta}_2 x_{2,i} + \ldots + \hat{\beta}_{p-1} x_{i,p-1})}}
$$

4.1.4 Interpretação para β_k (razão de chance)

Vejamos agora como é a interpretação dos parâmetros β_k na Regressão Logística. Antes disso vamos precisar de algumas definições.

Definição 4.1.6 *A chance (odds) de ocorrência de um evento é definida pela razão entre a probabilidade de ocorrência e de não ocorrência deste evento. Em outras palavras, se p é a probabilidade de um evento ocorrer a chance de ocorrência desse evento é*

$$odds = \frac{p}{1-p}.$$

Por exemplo, se dizemos que a chance de ocorrência de uma determinada doença é 1 significa que o evento "ter a doença" tem mesma probabilidade de ocorrer e de não ocorrer, isto é este evento tem probabilidade 1/2 de ocorrer. Por outro lado, se a chance de ocorrência dessa determinada doença for 1/2 significa que a probabilidade do evento "ter a doença" não ocorrer é 2 vezes a probabilidade dele ocorrer, ou seja, a probabilidade de um indivíduo ter a doença é 1/3. Para terminar, se a chance de ocorrência dessa determinada doença for 2 significa que a probabilidade do evento "ter a doença" ocorrer é 2 vezes a probabilidade dele não ocorrer, ou seja, a probabilidade de um indivíduo ter a doença é 2/3.

Veja que quanto maior a probabilidade do evento ocorrer maior será a chance de ocorrência desse evento. Veja também que $0 \leq p \leq 1$ e $0 \leq odds \leq \infty$.

Definição 4.1.7 *A razão de chance (odds ration) entre dois grupos para um determinado evento é definida como a razão entre a chance de ocorrência desse evento em um grupo e a chance de ocorrência do mesmo evento no outro grupo.*

$$OR = \frac{odds_1}{odds_2}$$

Por exemplo, se a razão de chance entre as mulheres (grupo 1) e os homens (grupo 2) para a ocorrência de uma determinada doença (evento) for 1 ($OR = 1$) significa que a chance de ocorrência dessa doença nas mulheres e nos homens é a mesma, ou seja, nos dois grupos a probabilidade de ocorrência é a mesma. Por outro lado, se a razão de chance entre as mulheres (grupo 1) e os homens (grupo 2) para a ocorrência dessa determinada doença for 1/2 ($OR = 1/2$) significa que a chance de ocorrência dessa doença entre as mulheres é metade da chance de ocorrência entre os homens, ou seja, a probabilidade de ocorrência entre os homens é maior que a probabilidade da doença ocorrer entre as mulheres. Para terminar, se a razão de chance entre as mulheres (grupo 1) e os homens (grupo 2) para a ocorrência dessa doença (evento) for 2 significa que a chance de ocorrência dessa doença entre as mulheres é o dobro da chance de ocorrência entre os homens, ou seja, a probabilidade de ocorrência entre as mulheres é maior que a probabilidade da doença ocorrer entre os homens.

Modelo Simples

Já vimos na Definição 4.1.4 que

$$E[y_i] = \pi_i = \frac{1}{1+e^{-(\beta_0+\beta_1 x_i)}} \qquad \text{ou} \qquad \ln\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_i$$

onde $\pi_i = P(y_i = 1)$, de acordo com o modelo. Então a chance $\{y_i = 1\}$ supondo o nível x_i é $\pi_i/(1-\pi_i)$. Podemos também definir a chance de ocorrência do evento $\{y = 1\}$ para um nível x qualquer como

$$odds(x) = \frac{\pi(x)}{1-\pi(x)}.$$

onde $\pi(x) = 1/\left(1 + e^{-(\beta_0+\beta_1 x)}\right)$. Assim podemos expressar a razão de chance entre os níveis $x+1$ e x para a ocorrência do evento $\{y = 1\}$ por:

$$OR = \frac{odds(x+1)}{odds(x)} = \frac{\pi(x+1)/(1-\pi(x+1))}{\pi(x)/(1-\pi(x))}.$$

Veja como fica a expressão $\ln(OR)$:

$$\begin{aligned}
\ln(OR) &= \ln\left(\frac{odds(x+1)}{odds(x)}\right) = \ln\left(odds(x+1)\right) - \ln\left(odds(x)\right) \\
&= \ln\left(\frac{\pi(x+1)}{1-\pi(x+1)}\right) - \ln\left(\frac{\pi(x)}{1-\pi(x)}\right) \\
&= \beta_0 + \beta_1(x+1) - (\beta_0 + \beta_1 x) = \beta_1.
\end{aligned}$$

Então,

$$\beta_1 = \ln\left(\frac{odds(x+1)}{odds(x)}\right) = OR \Rightarrow e_1^{\beta} = OR$$

onde OR é a razão de chance entre os níveis $x+1$ e x (qualquer que seja o x) para o evento $\{y = 1\}$. Logo, $e^{\hat{\beta}_1}$ será uma estimativa para essa razão de chance e será essa a interpretação que vamos usar para este parâmetro.

Exemplo 4.1.8 *Suponha que ajustamos o modelo de Regressão Logística Simples para os dados y = ter ou não ter diabetes tipo 2 e x = idade do paciente. Para esse modelo encontramos a seguinte função de regressão estimada*

$$\hat{\pi} = \frac{1}{1 + e^{-(-4.5+0.1x)}}$$

ou seja, as estimativas de máxima verossimilhança para os parâmetros foram $\hat{\beta}_0 = -4.5$ *e* $\hat{\beta}_1 = 0.1$*. Nesse caso a estimativa para a razão de chace é* $e^{\hat{\beta}_1} = e^{0.1} = 1.105171$*, ou seja, a razão de chance entre os níveis* $x+1$ *e* x *para o evento* $\{y = 1\}$ *é* 1.105171*. Isso significa que a cada ano de idade a chance do paciente ter diabetes tipo 2 aumenta em torno de 10%.*

Exemplo 4.1.9 *Suponha agora que ajustamos o modelo de Regressão Logística Simples novamente para y = ter ou não ter diabetes tipo 2, mas agora x = ter ou não ter história familiar de diabetes tipo 2. Vamos codificar* $x = 1$ *quando há histórico familiar e* $x = 0$ *caso contrário. Para esse modelo encontramos a seguinte função de regressão estimada*

$$\hat{\pi} = \frac{1}{1 + e^{-(-3.4+2.2x)}}$$

ou seja, as estimativas de máxima verossimilhança para os parâmetros foram $\hat{\beta}_0 = -3.4$ *e* $\hat{\beta}_1 = 2.2$*. Nesse caso a estimativa para a razão de chace é* $e^{\hat{\beta}_1} = e^{2.2} = 9.025013$*, ou seja, a razão de chance entre os níveis* $x+1$ *e* x *para o evento* $\{y = 1\}$ *é* 9.025013*. Isso significa que quando* $x = 1$ *a chance de ter diabetes é 9 vezes maior do que quando* $x = 0$*, ou seja, a chance de pacientes com histórico familiar ter diabetes é 9 vezes maior que a chance de pacientes sem histórico familiar.*

Exemplo 4.1.10 *Para terminar essa sequencia de exemplo suponha que ajustamos o modelo de Regressão Logística Simples mais uma vez para y = ter ou não ter diabetes tipo 2, mas agora x = praticar ou não atividades física regularmente. Vamos codificar* $x = 1$ *quando o paciente pratica atividades físicas regularmente e* $x = 0$ *caso contrário. Para esse modelo encontramos a seguinte função de regressão estimada*

$$\hat{\pi} = \frac{1}{1 + e^{-(-0.51-1.5x)}}$$

ou seja, as estimativas de máxima verossimilhança para os parâmetros foram $\hat{\beta}_0 = -0.51$ *e* $\hat{\beta}_1 = -1.5$. *Nesse caso a estimativa para a razão de chace é* $e^{\hat{\beta}_1} = e^{-1.5} = 0.2231302$, *ou seja, a razão de chance entre os níveis* $x + 1$ *e* x *para o evento* $\{y = 1\}$ *é* 0.2231302. *Aquando a razão de chance for menor que 1 fica mais fácil fazer a sua interpretação pelo inverso. Se a OR entre os níveis* $x + 1$ *e* x *é* 0.2231302 *então a razão de chance entre os níveis* x *e* $x + 1$ *é* $1/0.2231302 = 4.481688$. *Isso significa que quando* $x = 0$ *a chance de ter diabetes é 4 vezes maior do que quando* $x = 1$, *ou seja, a chance de pacientes que não praticam atividade física ter diabetes é 4 vezes maior do que a chance de pacientes que praticam atividade física.*

Para terminar a interpretação de β_1 no Modelo Logístico Simples, a razão de chance entre os níveis $x + k$ e x é definida como $e^{k\beta_1}$. Ou seja, para o Exemplo 4.1.8 uma estimativa para a razão de chance entre os níveis $x + 10$ e x é $e^{10\hat{\beta}_1} = e^1 = 2.718282$, o que significa que a cada 10 anos de idade que o paciente ganha a chance de ter diabetes tipo 2 aumenta mais quase 3 vezes.

Modelo Múltiplo

No modelo múltiplo a interpretação de β_k é equivalente a do modelo simples. A razão de chance entre os níveis $x_k + 1$ e x_k é e^{β_k}. Mas aqui temos que tomar mais um cuidado, pois essa razão de chance vale considerando que as demais variáveis preditivas não sejam alteradas.

Exemplo 4.1.11 *Suponha que ajustamos o modelo de Regressão Logística Simples para os dados y = ter ou não ter diabetes tipo 2,* x_1 *= idade do paciente,* x_2 *= ter ou não ter história familiar de diabetes tipo 2 e* x_3 *= praticar ou não atividades física regularmente. Para esse modelo encontramos a seguinte função de regressão estimada*

$$\hat{\pi} = \frac{1}{1 + e^{-(-5.6+0.12x_1+1.99x_2-1.3x_3)}}.$$

Nesse caso faremos uma interpretação para cada variável preditiva. Considerando x_1, *idade, temos* $OR = e^{0.12} = 1.127497$, *ou seja, a cada ano de vida a chance de um paciente ter diabetes tipo 2 aumenta em 12%, considerando pacientes com mesmo histórico familiar e mesma prática de exercícios físicos. Se quisermos ver o quanto aumenta a chance de ter diabetes entre pacientes com diferença de 10 anos temos que calcular* $e^{10\times0.12} = 3.320117$, *isso significa que a chance de ter diabetes tipo 2 triplica em 10 anos, considerando pacientes com mesmo histórico familiar e mesma prática de exercícios físicos.*

Agora vejamos a interpretação para x_2. *Como* $OR = e^{1.99} = 7.315534$ *podemos concluir que a chance de ter diabetes do tipo 2 entre os pacientes com histórico familiar é 7*

vezes maior que para aqueles que não tem histórico familiar, considerando pacientes da mesma idade e com a mesma prática de exercícios físicos.

Para terminar, considere x_3. Como $OR = e^{-1.3} = 0.2725318$ podemos considerar a razão de chance entre x e $x+1$, que será $1/0.2725318 = 3.669297$. Então a chance de ter diabetes tipo 2 entre os pacientes sem prática de exercícios físicos é 3 vezes maior do que a chance em pacientes com prática de exercícios físicos, considerando pacientes com mesma idade e mesmo histórico familiar de diabetes tipo 2.

4.1.5 Teste de Hipótese e IC para cada β_k: Teste de Wald

Resultados teóricos garantem que os estimadores de máxima verossimilhança são assintoticamente não tendenciosos e assintoticamente Normais, ou seja, para grandes amostras (n grande) podemos considerar $E[\hat{\beta}_k] = \beta_k$ e $\hat{\beta}_k \sim Normal$. Baseado nesse resultado vamos assumir

$$\frac{\hat{\beta}_k - \beta_k}{Var(\hat{\beta}_k)} \sim N(0,1).$$

Assim podemos construir o seguinte intervalo de confiança para cada β_k:

$$\hat{\beta}_k \pm z_{1-\alpha/2}\sqrt{Var(\hat{\beta}_k)} \tag{4.7}$$

E também as seguintes regras para decidir entre $H_0 : \beta_k = 0$ contra $H_1 : \beta_k \neq 0$,

$$\begin{array}{c} \text{Se } |z^*| \leq z_{1-\alpha/2} \text{ conclui } H_0; \\ \text{Se } |z^*| > z_{1-\alpha/2} \text{ conclui } H_1. \end{array}$$

onde a estatística de teste z^* é definida por

$$z^* = \frac{\hat{\beta}_k}{Var(\hat{\beta}_k)}$$

e $Var(\hat{\beta}_k)$ será fornecido pelo R.

4.1.6 Intervalo de confiança para OR

Na Seção 4.1.4 definimos $OR_k = e^{\beta_k}$, ou seja, OR_k é função estritamente crescente de β_k. Então podemos encontrar um intervalo de confiança para OR_k a partir do intervalo de confiança para β_k.

Se $[L, U]$ é um intervalo de confiança para β_k com confiabilidade $(1-\alpha)$, então $[e^L, e^U]$ é um intervalo de confiança para OR_k, também com confiabilidade $(1-\alpha)$.

4.1.7 Teste da Razão de Verossimilhança

Suponha que queremos testar:

$$\begin{array}{ll} H_0: & \beta_q = \beta_{q+1} = \ldots = \beta_{p-1} = 0 \\ H_1: & \text{algum } \beta_k \neq 0,\ k = q, q+1, \ldots, p-1 \end{array}$$

Nesse caso o modelo completo (F) é aquele com todas as variáveis preditivas e o modelo reduzido (R) é aquele somente com as variáveis $x_1, \ldots, x_{q-1}$. Para escolher entre os dois modelos vamos usar a estatística de teste G^2 definida por:

$$G^2 = -2 \ln \left(\frac{L(R)}{L(F)} \right) = -2 \left(l(R) - l(F) \right) \tag{4.8}$$

onde a função l é aquela definida nas Equações 4.5 e 4.6.

Veja que se $L(R)/L(F)$ for pequeno, significa que $L(R) << L(F)$. Logo vamos preferir o modelo completo. Veja também que quando $L(R)/L(F)$ é pequeno G^2 é grande. Já se $L(R)/L(F)$ for grande, G^2 será pequeno e vamos preferir o modelo reduzido.

Quando temos uma amostra grande podemos considerar, sob H_0, G^2 aproximadamente uma variável aleatória χ^2_{p-q}. Então a regra de decisão deste teste será:

$$\begin{array}{c} \text{Se } G^2 \leq q_{1-\alpha} \text{ conclui } H_0; \\ \text{Se } G^2 > q_{1-\alpha} \text{ conclui } H_1. \end{array}$$

onde $q_{1-\alpha}$ é o quantil da distribuição χ^2_{p-q}.

Para encontrar $l(R)$ primeiro ajuste o modelo reduzido, depois encontre os valores ajustados $\hat{\pi}_i$ e substitua π_i por $\hat{\pi}_i$ na Equação 4.5. Outra alternativa é ajustar o modelo, encontrar as estimativas $\underline{\hat{\beta}}$ e então substituir cada β_k por $\hat{\beta}_k$ na Equação 4.6. De forma semelhante podemos encontrar o valor de $l(F)$, a única diferença é que para isso o modelo ajustado será o completo.

4.1.8 Seleção do modelo

Assim como no Modelo Linear Normal, a seleção do modelo para Regressão Logística pode ser feita comparando todos os possíveis modelos ou utilizando os métodos da Inclusão ou Eliminação de variáveis.

Comparação entre todos os modelos possíveis

A comparação entre todos os possíveis modelos para a Regressão Logística pode ser feita a partir do Critério da Informação de Akaike (AIC) ou do Critério da Informação Baysiana (BIC), que passam a depender da função de verossimilhança, como mostram as Equações 4.9 e 4.10

$$AIC = -2 \ln \left(L \left(\underline{\hat{\beta}} \right) \right) + 2p \tag{4.9}$$

$$BIC = -2 \ln \left(L \left(\underline{\hat{\beta}} \right) \right) + p \ln(n) \tag{4.10}$$

Métodos de seleção passo-a-passo

Os métodos da inclusão e eliminação de variáveis seguem os mesmos algoritmos apresentados na Seção 3.4.2. A única diferença é que no caso da Regressão Logística o p-valor usado será o do Teste de Wald.

4.1.9 Intervalo de confiança para a média da variável resposta

Queremos agora encontrar um intervalo de confiança para a media da variável resposta, isto é, para π_h dado um nível $\underline{x_h}$ qualquer. Já vimos que a Regressão Logística supõe

$$\pi_h = \frac{1}{1 + e^{-\underline{x_h}^T \underline{\beta}}}. \tag{4.11}$$

Então, para encontrar um intervalo de confiança para π_h vamos primeiro encontrar um intervalo de confiança para $\underline{x_h}^T \underline{\beta}$ e depois encontramos um para π_h.

Já vimos, pelas propriedades assintóticas dos estimadores de máxima verossimilhança, que para n grande podemos considerar $\underline{\hat{\beta}} \sim N_p$ com $E[\underline{\hat{\beta}}] = \underline{\beta}$. Vamos chamar de $Var(\underline{\hat{\beta}})$ a matriz de variância-covariância desse estimador e de $\hat{Var}(\underline{\hat{\beta}})$ uma estimativa para essa matriz. Então,

$$\underline{x_h}^T \underline{\hat{\beta}} \sim N\left(\underline{x_h}^T \underline{\beta} \ , \ \underline{x_h}^T Var(\underline{\hat{\beta}}) \underline{x_h}\right)$$

Padronizando,

$$\frac{\underline{x_h}^T \underline{\hat{\beta}} - \underline{x_h}^T \underline{\beta}}{\sqrt{\underline{x_h}^T Var(\underline{\hat{\beta}}) \underline{x_h}}} \sim N(0,1).$$

Como n é supostamente grande podemos considerar mais uma aproximação, $Var(\underline{\hat{\beta}}) \approx \hat{Var}(\underline{\hat{\beta}})$. Assim chegamos na seguinte quantidade pivotal para $\underline{x_h}^T \underline{\beta}$:

$$\frac{\underline{x_h}^T \underline{\hat{\beta}} - \underline{x_h}^T \underline{\beta}}{\sqrt{\underline{x_h}^T \hat{Var}(\underline{\hat{\beta}}) \underline{x_h}}} \sim N(0,1).$$

A partir desta quantidade pivotal podemos construir o seguinte intervalo de confiança para $\underline{x_h}^T \underline{\beta}$:

$$\underline{x_h}^T \underline{\hat{\beta}} \pm z_{1-\alpha/2} \sqrt{\underline{x_h}^T \hat{Var}(\underline{\hat{\beta}}) \underline{x_h}} \tag{4.12}$$

A matriz $\hat{Var}(\underline{\hat{\beta}})$, aproximação da matriz de variância e covariância de $\underline{\hat{\beta}}$ pode ser encontrada no R a partir do comando `summary(mlogit)$cov.unscaled`.

Como π_h é função estritamente crescente de $\underline{x_h}^T \underline{\beta}$, veja Equação 4.11, podemos afirmar que se $[L, U]$ é um intervalo de confiança para $\underline{x_h}^T \underline{\beta}$ com confiabilidade de $1 - \alpha$ então o intervalo $[\frac{1}{1+e^{-L}}, \frac{1}{1+e^{-U}}]$ é um intervalo de confiança para π_h com a mesma confiabilidade de $1 - \alpha$.

4.1.10 Previsão para uma nova observação

No caso da Regressão Logística temos $y_i \sim Bernoulli(\pi_i)$ e a partir do modelo estimado somos capazes de encontrar $\hat{\pi}_i$, uma estimativa para $\pi_i = P(y_i = 1)$. Então, dado um nível $\underline{x_h}$ qualquer sabemos encontrar $\hat{\pi}_h = P(y_h = 1 | \underline{x_h})$. Mas como fazer previsões a partir do modelo estimado? Ou seja, como encontrar $\hat{y}_h$ para um nível $\underline{x_h}$ qualquer?

Talvez a alternativa mais intuitiva para realizar uma previsão seja:

$$\hat{y}_h = \begin{cases} 1 & \text{, se } \hat{\pi}_h > \frac{1}{2} \\ 0 & \text{, se } \hat{\pi}_h \leq \frac{1}{2}. \end{cases}$$

Mas essa não é a única alternativa. Poderíamos escolher outro ponto de corte qualquer, não necessariamente o 1/2. Ou seja, de forma geral podemos definir a seguinte regra para realizar as previsões:

$$\hat{y}_h = \begin{cases} 1 & \text{, se } \hat{\pi}_h > \pi^* \\ 0 & \text{, se } \hat{\pi}_h \leq \pi^*. \end{cases} \tag{4.13}$$

Dado um ponto de corte π^* é possível criar a Tabela de Classificação 4.1 a seguir, que vai ajudar a medir o quanto o modelo está bem ajustado aos dados da amostra.

	$\hat{y}_i = 0$	$\hat{y}_i = 1$
$y_i = 0$	t_{00}	t_{01}
$y_i = 1$	t_{10}	t_{11}

Tabela 4.1: Tabela de Classificação

t_{00} é número de observações dentro da amostra tais que $y_i = 0$ e $\hat{y}_i = 0$
t_{01} é número de observações dentro da amostra tais que $y_i = 0$ e $\hat{y}_i = 1$
t_{10} é número de observações dentro da amostra tais que $y_i = 1$ e $\hat{y}_i = 0$
t_{11} é número de observações dentro da amostra tais que $y_i = 1$ e $\hat{y}_i = 1$

Veja que o número de previsões corretas dentro da amostra é $t_{00} + t_{11}$ e o número de previsões erradas é $t_{01} + t_{10}$. Na área médica costuma-se usar os termos verdadeiro positivo (t_{11}), verdadeiro negativo (t_{00}), falso positivo (t_{01}) e falso negativo (t_{10}). Outros dois termos muito usados são: a sensibilidade e a especificidade, que ajudam a medir o quão preciso é o teste.

A sensibilidade mede a capacidade do teste em identificar corretamente $\hat{y}_i = 1$ entre as observações com $y_i = 1$, ou seja, o quão sensível é o teste. Podemos então definir:

$$\text{sensibilidade} = P(\hat{y}_i = 1 | y_i = 1) = \frac{P(\hat{y}_i = 1 \cap y_i = 1)}{P(y_i = 1)} = \frac{t_{11}}{t_{10} + t_{11}}.$$

A especificidade mede a capacidade do teste em identificar corretamente $\hat{y}_i = 0$ entre as observações com $y_i = 0$, ou seja, o quão específico o teste é. Podemos então definir:

$$\text{especificidade} = P(\hat{y}_i = 0 | y_i = 0) = \frac{P(\hat{y}_i = 0 \cap y_i = 0)}{P(y_i = 0)} = \frac{t_{00}}{t_{00} + t_{01}}.$$

Veja que quanto maior a sensibilidade ou quanto maior a especificidade melhor o modelo está ajustado. Mas o aumento de uma medida gera a diminuição da outra. Se escolhemos $\pi^* = 1$ temos sensibilidade = 0, mas especificidade = 1. Já se escolhemos $\pi^* = 0$ temos sensibilidade = 1, mas especificidade=0. A escolha do valor de π^* pode ser feita de forma a maximizar a soma sensibilidade + especificidade.

Para encontrar esse valor de π^* é muito comum usar a Curva ROC, apresentada na Figura 4.4. Essa curva é construída a partir dos valores da sensibilidade e de 1-especificidade, para diferentes valores de π^*. Veja que

$$1 - \text{especificidade} = 1 - P(\hat{y}_i = 0 | y_i = 0) = P(\hat{y}_i = 1 | y_i = 0)$$

é a probabilidade do modelo fornecer um falso positivo. Então a curva usa os valores de sensibilidade (probabilidade de um verdadeiro positivo) e os valores de 1-especificidade (probabilidade de um falso positivo). O ponto desejado é aquele que maximiza sensibilidade e minimiza 1-especificidade.

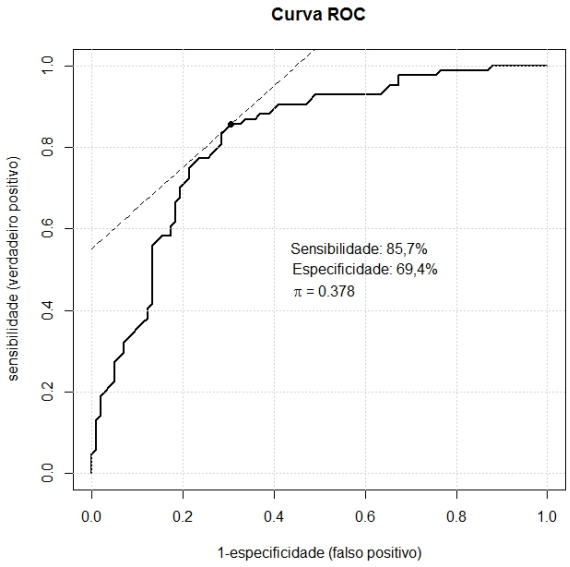


Figura 4.4: Comparação entre sensibilidade e especificidade para a escolha de π^*

O valor escolhido para π^* será aquele que gerou o ponto destacado no gráfico da Figura 4.4, ou seja, aquele que maximizou a sensibilidade e minimizou 1-especificidade. Para esse exemplo o valor escolhido foi $\pi^* = 0.378$ e, para esse valor de π^*, a sensibilidade do ajuste foi 85,7% e a especificidade foi de 69,4%. Ou seja, a probabilidade do modelo ajustado fornecer um falso positivo é de 30,6%.

Podemos gerar a Curva ROC no R a partir dos comandos gráficos ou a partir do comando `ROC` do pacote `Epi`. A sintaxe desse comando é: `ROC(form = y ~ x, plot = "ROC")`, onde y guarda os valores observados da variável resposta e x a variável preditiva.

4.1.11 Várias observações para cada nível - Modelo Binomial

Em alguns experimentos é possível observar muitas vezes a variável resposta y para um mesmo nível x. Por exemplo, considere o exercício da aula prática em que y indicava se um programador conseguiu ou não executar uma dada tarefa e x o número de meses de experiência desse programador. Para esse experimento poderíamos buscar vários programadores para cada possível valor de x, ou seja, vários programadores com 5 meses de experiência, vário com 6 meses, vários com 7, .. todos eles tentariam executar a tarefa e assim teríamos várias observações de y para cada nível x. Mas atenção, não poderíamos colocar um mesmo programador para tentar executar várias vezes a tarefa, pois nesse caso as observações não seriam independentes.

Para esse tipo de problema vamos definir a seguinte notação. Seja c o número de diferentes níveis da variável preditiva $\underline{x}$. No caso do exemplo do programador teríamos $x \in \{5, 6, \ldots, 30\}$, supondo que os programadores da amostra tinham de 5 a 30 meses de experiência, e nesse caso $c = 26$. Para cada $j = 1, 2, \ldots, c$ seja n_j o número de observações da variável resposta no nível j. Voltando ao exemplo, se $j = 1$ representa os programadores com 5 meses de experiência n_1 seria o número de programadores com 5 meses de experiência na amostra. Vamos usar a notação $y_{i,j}$ para representar a i-ésima observação do nível j e p_j para representar a proporção de sucessos na amostra para o

nível j. Assim temos:

$$p_j = \frac{\sum_{i=1}^{n_j} y_{i,j}}{n_j} \tag{4.14}$$

Veja que $\sum_{i=1}^{n_j} y_{i,j} \sim Binomial(n_j, \pi_j)$ com n_j conhecido. O problema continua o mesmo, buscamos a partir das observações de y a curva logística para estimar π. Mas agora, com várias observações em cada nível, a curva é aquela que melhor se ajusta aos pontos (x_j, p_j).

É claro que nem sempre é possível criar uma amostra com medidas repetidas, mas sempre que isso for possível esta é uma medida recomendada. Quando temos várias observações para cada nível possível temos mais estabilidade nas estimativas, isto é, menor variância, e ainda podemos realizar o teste da qualidade do ajuste descrito a seguir.

Função de log-verossimilhança

No caso que termos várias observações independentes para cada nível podemos re-escrever a função de log-verossimilhança l definida na Equação 4.5 em função dos parâmetros n_j, p_j e c, como mostrado a Equação 4.15 a seguir.

$$\begin{aligned} l(\underline{\pi}|\underline{y}, \underline{x}) &= \sum_{i=1}^{n} \left(y_i \ln\left(\frac{\pi_i}{1-\pi_i}\right) + \ln(1-\pi_i) \right) \\ &= \sum_{j=1}^{c} \left(\sum_{i=1}^{n_j} \left(y_{i,j} \ln\left(\frac{\pi_j}{1-\pi_j}\right) + \ln(1-\pi_j) \right) \right) \\ &= \sum_{j=1}^{c} \left(\ln\left(\frac{\pi_j}{1-\pi_j}\right) \sum_{i=1}^{n_j} y_{i,j} + n_j \ln(1-\pi_j) \right) \\ &= \sum_{j=1}^{c} \left(n_j p_j \ln\left(\frac{\pi_j}{1-\pi_j}\right) + n_j \ln(1-\pi_j) \right) \end{aligned} \tag{4.15}$$

Teste da qualidade do ajuste

O teste da qualidade do ajuste verifica se o modelo está bem ajustado. Neste teste as as hipóteses são:

$$\begin{aligned} H_0&: \quad E[y_{i,j}] = 1/(1 + e^{-\underline{x_j}^T \underline{\beta}}) \\ H_1&: \quad E[y_{i,j}] = \pi_j, \ j = 1, 2, \ldots, c \end{aligned}$$

A ideia é testar H_0, hipótese do modelo bem ajustado, contra H_1, hipótese de que para cada valor do nível x existe uma probabilidade π_j que não pode ser descrita pela função logística.

Considerando o modelo reduzido aquele definido na hipótese H_0 e o modelo completo aquele definido na hipótese H_1 o teste segue com o teste da razão de máxima verosimilhança, definido na Equação 4.8. Vamos construir então a estatística de teste G^2, para isso precisamos de $l(R)$ e $l(F)$. Veja que $l(R)$ pode ser encontrada substituindo π_j da Equação 4.15 por $\hat{\pi}_j$, valor ajustado pelo modelo reduzido definido pela regressão logística. Já $l(F)$ pode ser encontrada substituindo π_j da Equação 4.15 por p_j. Assim chegamos nas seguintes expressões.

$$
\begin{aligned}
l(R) &= \sum_{j=1}^{c}\left(n_j p_j \ln\left(\frac{\hat{\pi}_j}{1-\hat{\pi}_j}\right) + n_j \ln(1-\hat{\pi}_j)\right) \\
l(F) &= \sum_{j=1}^{c}\left(n_j p_j \ln\left(\frac{p_j}{1-p_j}\right) + n_j \ln(1-p_j)\right)
\end{aligned}
$$

onde $\hat{\pi_j}$ é o valor ajustado considerando o modelo reduzido. Logo, $G^2 = -2\left(l(R) - l(F)\right)$ pode ser expressa como na Equação 4.16:

$$
G^2 = -2\sum_{j=1}^{c}\left(n_j p_j \ln\left(\frac{\hat{\pi}_j}{p_j}\right) + n_j(1-p_j)\ln\left(\frac{1-\hat{\pi}_j}{1-p_j}\right)\right) \tag{4.16}
$$

Nesse caso a estatística de teste G^2 é chamada de Função Desvio, Desviância ou *Deviance*. Sob H_0 podemos considerar $G^2 \sim \chi^2_{c-p}$ somente se cada n_j for suficientemente grande. Se isso acontecer podemos definir a seguinte regra de decisão:

$$
\begin{array}{c}
\text{Se } G^2 \leq q_{1-\alpha} \text{ conclui } H_0; \\
\text{Se } G^2 > q_{1-\alpha} \text{ conclui } H_1.
\end{array}
$$

onde $q_{1-\alpha}$ é o quantil da distribuição χ^2_{c-p}.

4.2 Regressão de Poisson

A Regressão de Poisson é mais um modelo linear generalizado que veremos nesse curso. Esta é uma opção de regressão quando a variável resposta y se refere a dados de contagem, ou seja, $y = 0, 1, 2, \ldots$ A suposição do modelo é que $y_i \sim Poisson(\lambda_i)$ e o nosso interesse é estudar como varia λ_i em função das variáveis preditivas. Antes de formalizar o modelo vejamos um exemplo.

Exemplo 4.2.1 *Suponha que estamos interessado em estudar o número de vezes que uma família vai ao mercado por mês. Queremos saber se, em média, esse número varia em função de algumas variáveis preditivas, como por exemplo, o tamanho da família, se a família mora no centro ou não, a quantidade de crianças, renda, etc. Nesse caso devemos definir y_i = número de idas ao mercado da família i no mês de observação e considerar $y_i \sim Poisson(\lambda_i)$. O que estamos interessados em estudar é a influência das variáveis preditivas em $E[y_i] = \lambda_i$.*

4.2.1 O Modelo da Regressão de Poisson

No caso da Seção 4.1, quando y era uma variável binária, estabelecemos que a relação entre $E[y_i] = \pi_i$ e as variáveis preditivas era dada pela Função Logística. Aqui também precisamos definir uma relação entre $E[y_i] = \lambda_i$ e as variáveis preditivas e, assim como antes, não temos uma única opção para isso. A opção mais comum é considerar a relação exponencial, uma vez que dessa forma garantimos $\hat{\lambda}_i > 0$.

Definição 4.2.2 *Seja $y_i \sim Poisson(\lambda_i)$, o Modelo da Regressão de Poisson define*

$$y_i = E[y_i] + \varepsilon_i$$

e supõe a seguinte relação entre $E[y_i] = \lambda_i$ e as $p-1$ variáveis preditivas $\underline{x}_i = (x_{i,1}, x_{i,2}, \ldots, x_{i,p-1})$:

$$E[y_i] = \lambda_i = e^{\underline{x}_i^T \underline{\beta}} = e^{\beta_0 + \beta_1 x_{i,1} + \beta_2 x_{2,i} + \ldots + \beta_{p-1} x_{i,p-1}}$$

$$ou \qquad \ln(\lambda_i) = \underline{x}_i^T \underline{\beta} = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{2,i} + \ldots + \beta_{p-1} x_{i,p-1}$$

O estimador adotado para $\underline{\beta}$ será o de máxima verossimilhança. Vamos encontrá-lo.

$$\begin{aligned} L(\underline{\beta}|\underline{y}, \underline{x}) &= \prod_{i=1}^{n} e^{-\lambda_i} \frac{\lambda_i^{y_i}}{y_i!} \\ l(\underline{\beta}|\underline{y}, \underline{x}) &= \ln\left(\prod_{i=1}^{n} e^{-\lambda_i} \frac{\lambda_i^{y_i}}{y_i!}\right) = \sum_{i=1}^{n} \ln\left(e^{-\lambda_i} \frac{\lambda_i^{y_i}}{y_i!}\right) = \sum_{i=1}^{n} \ln\left(e^{-\lambda_i}\right) + \ln(\lambda_i^{y_i}) - \ln(y_i!) \\ &= -\sum_{i=1}^{n} \lambda_i + \sum_{i=1}^{n} y_i \ln(\lambda_i) - \sum_{i=1}^{n} \ln(y_i!) \end{aligned}$$

Chegamos assim na seguinte função de log-verossimilhança para a Regressão de Poisson.

$$l(\underline{\beta}|\underline{y_i}, \underline{x_i}) = \sum_{i=1}^{n} y_i \ln(\lambda_i) - \sum_{i=1}^{n} \lambda_i - \sum_{i=1}^{n} \ln(y_i!) \tag{4.17}$$

Para encontrar os estimadores de máxima verossimilhança é preciso encontrar o ponto de máximo da função l definida na Equação 4.17. Mas também não existe solução analítica para esse problema e por isso será necessário métodos iterativos para encontrar as estimativas dada uma certa amostra.

Dada a estimativa $\hat{\underline{\beta}}$, a estimativa pontual para a média da variável resposta, nesse caso para λ_i, é dada por:

$$\hat{\lambda}_i = e^{\underline{x_i}^T \hat{\underline{\beta}}} = e^{\hat{\beta}_0 + \hat{\beta}_1 x_{i,1} + \hat{\beta}_2 x_{2,i} + \ldots + \hat{\beta}_{p-1} x_{i,p-1}}$$

As inferências feitas na Regressão de Poisson são bem semelhante aquelas feitas na Regressão Logística, por isso as Seções 4.1.5 (teste de hipótese e intervalo de confiança para cada β_k) e 4.1.7 (teste da razão de verossimilhança) também valem para a Regressão de Poisson. A seleção do modelo também não muda, por isso a Seção 4.1.8 também serve para a Regressão de Poisson. A única adaptação que deve ser feita nessas seções é considerar a função l como a função de log-verossimilhança do modelo de Poisson, definida pela Equação 4.17.

4.2.2 Interpretação para β_k

Já vimos que a Regressão de Poisson supõe que a média de y_i para um certo nível $\underline{x_i}$ seja modelada por:

$$\lambda_i = e^{\underline{x}_i^T \underline{\beta}} = e^{\beta_0 + \beta_1 x_{i,1} + \beta_2 x_{2,i} + \ldots + \beta_{p-1} x_{i,p-1}}.$$

Modelo Simples

Definimos o risco relativo (RR) como a razão entre a média de y_i para $x_i = x+1$ e $x_i = x$. Se estamos no modelo simples temos $E[y_i|x_i] = \lambda_i = e^{\beta_0+\beta_1 x_i}$ e

$$RR = \frac{E[y_i|x_i = x+1]}{E[y_i|x_i = x]} = \frac{e^{\beta_0+\beta_1(x+1)}}{e^{\beta_0+\beta_1 x}} = \frac{e^{\beta_0}e^{\beta_1 x}e^{\beta_1}}{e^{\beta_0}e^{\beta_1 x}} = e^{\beta_1}.$$

Veja que se $RR = e^{\beta_1} > 1$ significa que a média de y_i cresce quando x cresce e se $RR = e^{\beta_1} < 1$ significa que a média de y_i diminui quando x cresce. Vejamos dois exemplos.

Exemplo 4.2.3 *Suponha que ajustamos o modelo de Regressão de Poisson para os dados y = o número de vezes que uma família vai ao mercado por mês e x = tamanho da família. Para esse modelo encontramos a seguinte função de regressão estimada*

$$\hat{\lambda}_i = e^{-0.61+0.58x_i}$$

ou seja, as estimativas de máxima verossimilhança para os parâmetros foram $\hat{\beta}_0 = -0.61$ *e* $\hat{\beta}_1 = 0.58$*. Nesse caso a estimativa para o risco relativo é* $RR = e^{\hat{\beta}_1} = e^{0.58} = 1.786038$*, ou seja, a cada acréscimo de um indivíduo na família o número médio de idas ao mercado por mês aumenta em 78%.*

Exemplo 4.2.4 *Suponha que ajustamos o modelo de Regressão de Poisson novamente para y = o número de vezes que uma família vai ao mercado por mês, mas agora considere e x = 1 se a família mora no centro e 0 caso contrário. Para esse modelo encontramos a seguinte função de regressão estimada*

$$\hat{\lambda}_i = e^{0.09+1.01x_i}$$

ou seja, as estimativas de máxima verossimilhança para os parâmetros foram $\hat{\beta}_0 = 0.09$ *e* $\hat{\beta}_1 = 1.01$*. Nesse caso a estimativa para o risco relativo é* $RR = e^{\hat{\beta}_1} = e^{1.01} = 2.745601$*, ou seja, as famílias que moram no centro vão 2.74 vezes mais ao mercado por mês quando comparadas com as famílias que não moram no centro.*

Modelo Múltiplo

No modelo múltiplo a interpretação de β_k é equivalente a do modelo simples. O risco relativo entre as médias para $x_k + 1$ e x_k é e^{β_k}. Mas aqui temos que tomar mais um cuidado, pois temos que considerar as demais variáveis preditivas constantes.

Exemplo 4.2.5 *Suponha que ajustamos o modelo de Regressão de Poisson para os dados y = o número de vezes que uma família vai ao mercado por mês,* x_1 *= tamanho da família e* $x_2 = 1$ *se a família mora no centro e 0 caso contrário. Para esse modelo encontramos a seguinte função de regressão estimada*

$$\hat{\lambda}_i = e^{-0.16+0.61x_{i,1}+0.92x_{i,2}}$$

ou seja, as estimativas de máxima verossimilhança para os parâmetros foram $\hat{\beta}_0 = -0.16$*,* $\hat{\beta}_1 = 0.61$ *e* $\hat{\beta}_2 = 0.92$*. A interpretação deve ser feita para uma variável de cada vez.*

A estimativa para o risco relativo referente a variável x_1 *é* $RR_1 = e^{\hat{\beta}_1} = e^{0.61} = 1.840431$*, ou seja, a cada acréscimo de um indivíduo na família o número médio de idas ao mercado por mês aumenta em 84%, considerando a mesma região de moradia.*

A estimativa para o risco relativo referente a variável x_2 *é* $RR_2 = e^{\hat{\beta}_2} = e^{0.92} = 2.50929$*, ou seja, uma família que mora no centro vai cerca de 2.5 vezes mais ao mercado por mês do que uma família que não mora no centro, considerando famílias com mesma quantidade de indivíduos.*

4.2.3 Teste da qualidade do ajuste

Para a Regressão de Poisson podemos realizar o teste da qualidade do ajuste e verificar as seguintes hipóteses:

$$\begin{array}{ll} H_0: & E[y_i] = e^{\underline{x_i}^T \underline{\beta}} \\ H_1: & E[y_i] = \lambda_i,\ i = 1, 2, \ldots, n \end{array}$$

ou seja, a hipótese H_0 considera o ajuste bom e a hipótese H_1 considera que o ajuste não é adequado. O teste a ser realizado é o teste da razão de máxima verossimilhança supondo que o modelo reduzido é aquele definido pela hipótese H_0 e o modelo completo aquele definido por H_1. Veja cmo fica a função de verossimilhança l, definida na Equação 4.17, para cada um desses modelos.

$$\begin{aligned} l(R) &= \sum_{i=1}^{n} y_i \ln\left(\hat{\lambda}_i\right) - \sum_{i=1}^{n} \hat{\lambda}_i - \sum_{i=1}^{n} \ln\left(y_i!\right) \\ l(F) &= \sum_{i=1}^{n} y_i \ln\left(y_i\right) - \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \ln\left(y_i!\right) \end{aligned}$$

onde $\hat{\lambda}_i$ é o valor ajustado considerando o modelo reduzido. A estatística de teste $G^2 = -2\left(l(R) - l(F)\right)$ é chamada de Função Desvio (Deviance ou *Deviance*) para a Regressão de Poisson e pode ser expressa como na Equação 4.18:

$$G^2 = -2\left[\sum_{i=1}^{n} y_i \ln\left(\frac{\hat{\lambda}_i}{y_i}\right) + \sum_{i=1}^{n}\left(y_i - \hat{\lambda}_i\right)\right]. \tag{4.18}$$

Sob H_0 G^2 tem distribuição χ_{n-p} graus de liberdade. O termo $y_i \ln\left(\frac{\hat{\lambda}_i}{y_i}\right)$ será considerado 0 sempre que $y_i = 0$.

4.2.4 Componentes da Função Desvio

Para a Regressão de Poisson ainda podemos usar a componente do Desvio, definida por dev_i na Equação 4.19, para verificar se o modelo está bem ajustado. Essa verificação pode ser feita a partir do gráfico de dev_i versus o índice. Se o modelo está bem ajustado as componentes do desvio devem estar aleatórias em torno de zero.

$$dev_i = \pm\left[-2y_i \ln\left(\frac{\hat{\lambda}_i}{y_i}\right) - 2(y_i - \hat{\lambda}_i)\right]^{1/2} \tag{4.19}$$

onde o sinal é definido pelo sinal de $y_i - \hat{\lambda}_i$ o termo $y_i \ln\left(\frac{\hat{\lambda}_i}{y_i}\right)$ será considerado 0 sempre que $y_i = 0$. Veja que,

$$G^2 = \sum_{i=1}^{n} dev_i^2.$$

4.2.5 Intervalo de confiança para a média da variável resposta

Já vimos que na Regressão de Poisson tem-se:

$$\lambda_i = e^{\underline{x_i}^T \underline{\beta}} = e^{\beta_0 + \beta_1 x_{i,1} + \beta_2 x_{2,i} + \ldots + \beta_{p-1} x_{i,p-1}}.$$

Para encontrar um intervalo de confiança para λ_i vamos primeiro encontrar o intervalo de confiança para $\underline{x_i}^T \underline{\beta}$ definido na Equação 4.12, denominado por $[L, U]$. Então $[e^L, e^U]$ é um intervalo de confiança para λ_i, uma vez que e^x é uma função estritamente crescente.

4.3 Modelos Lineares Generalizados

Sejam $y_1, y_2, \ldots, y_n$ variáveis aleatórias independentes com média $E[y_i] = \mu_i$. Sejam $\underline{x_1}, \underline{x_2}, \ldots, \underline{x_n}$ observações das variáveis preditivas, onde $\underline{x_i} = \{x_{i,1}, x_{i,2}, \ldots, x_{i,p-1}\}$.

Um modelo que relaciona y_i com $\underline{x_i}$ pertence a família dos Modelos Lineares Generalizados se ele supõe que:

$$g(\mu_i) = \underline{x_i}^T \underline{\beta} = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \ldots + \beta_{p-1} x_{i,p-1} \tag{4.20}$$

Para alguma função g, chamada de função de ligação.

Veja que os modelos vistos nesse curso são modelos lineares generalizados.

- Modelo Linear Normal:

  $$E[y_i] = \mu_i \text{ e } \mu_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \ldots + \beta_{p-1} x_{i,p-1}.$$

  Função de ligação identidade: $g(\mu_i) = \mu_i$.

- Modelo Logístico:

  $$E[y_i] = \pi_i \text{ e } \frac{\pi_i}{1-\pi_i} = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \ldots + \beta_{p-1} x_{i,p-1}.$$

  Função de ligação: $g(\pi_i) = \frac{\pi_i}{1-\pi_i}$.

- Modelo Poisson:

  $$E[y_i] = \lambda_i \text{ e } \ln(\lambda_i) = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \ldots + \beta_{p-1} x_{i,p-1}.$$

  Função de ligação: $g(\lambda_i) = \ln(\lambda_i)$.

Exercícios para Aulas Práticas do Capítulo 4

1. Uma analista de sistemas está interessado em estudar o efeito da experiência de um programador na capacidade de finalizar uma tarefa complexa em um tempo específico. Para isso foram selecionadas 25 pessoas com diferentes meses de experiência em programação e para cada uma delas foi solicitado que uma mesma tarefa fosse executada. Na tabela apresentada no arquivo `CH14TA01.txt` (`https://netfiles.umn.edu/users/nacht001/www/nachtsheim/5th/KutnerData/Chapter%2014%20Data%20Sets/CH14TA01.txt`) a primeira coluna indica os meses de experiência de cada indivíduo (x) e a segunda coluna se a tarefa foi finalizada no tempo ou não (y), caso a tarefa tenha sido executado no tempo tem-se $y = 1$. Esqueça a terceira coluna.

   (a) Faça o gráfico de dispersão das variáveis x e y. Discuta o gráfico. Você acha que o modelo logístico é apropriado?

   (b) Usando o comando `glm` encontre estimativas para os parâmetros β_0 e β_1 do modelo logístico.

   (c) Adicione ao gráfico da questão (a) a curva de regressão estimada a partir do modelo logístico.

   (d) Encontre uma estimativa pontual para a probabilidade de um programador com 14 meses de experiência finalizar a tarefe no tempo determinado.

   (e) Encontre a razão de chances quando o tempo de experiência muda em 1 mês. Interprete o resultado.

   (f) Encontre a razão de chances quando o tempo de experiência muda em 15 meses. Interprete o resultado.

   (g) Espera-se que o parâmetro β_1 seja positivo, uma vez que mais tempo de experiência deve resultar em maior chance de terminar a tarefa. Formule um teste estatístico a fim de confirmar essa hipótese.

   (h) Encontre um Intervalo de Confiança para o parâmetro β_1. Use $\alpha = 0.05$.

   (i) Encontre um Intervalo de Confiança para a razão de chance e interprete o valor encontrado. Use $\alpha = 0.05$.

2. Em um estudo de saúde para investigar um surto epidêmico de uma doença que é transmitida por mosquitos indivíduos foram amostrados aleatoriamente em dois setores da cidade. Cada paciente foi entrevistado e os entrevistadores fizeram perguntas pertinentes para avaliar se alguns sintomas específicos associados com as doenças estavam presentes durante o período de teste. A variável resposta y foi codificado em 1 se a doença foi determinada e 0 caso contrário.

   Três variáveis preditivas foram incluídas no modelo: idade (x_1), status socioeconômico da família e setor das cidade. O status socioeconômico da família foi categorizado usando duas variáveis indicadoras, como mostra a tabela abaixo.

Classe	x_2	x_3
Alta	0	0
Média	1	0
Baixa	0	1

O setor da cidade também foi categorizado, mas como o estudo foi realizado em apenas dois setores uma única variável indicadora x_4 foi usada: $x_4 = 0$ se o paciente é do setor 1 e $x_4 = 1$ se ele é do setor 2.

Os dados referentes aos 98 pacientes entrevistados encontram-se no arquivo `CH14TA03.txt` (`https://netfiles.umn.edu/users/nacht001/www/nachtsheim/5th/KutnerData/Chapter%2014%20Data%20Sets/CH14TA03.txt`). A primeira coluna guarda um índice de identificação do paciente e as demais colunas apresentam as variáveis do problema na seguinte ordem: x_1, x_2, x_3, x_4 e y.

(a) Usando o comando `glm` encontre estimativas para os parâmetros do modelo logístico.

(b) Verifique se a variável idade deve ser retirada do modelo.

(c) Verifique se a variável status socioeconômico deve ser retirada do modelo.

(d) Faça a seleção do modelo a partir do método da eliminação de variáveis.

(e) Encontre uma estimativa para a razão de chance de cada variável do modelo final. Interprete os resultados.

(f) Encontre uma estimativa pontual para a probabilidade de uma pessoa de 33 anos, de classe alta e do setor 1 contrair essa doença.

(g) Encontre uma estimativa intervalar para a probabilidade de uma pessoa de 33 anos, de classe alta e do setor 1 contrair essa doença. Use 95% de confiabilidade. Observe que o intervalo encontrado não é simétrico em relação à estimação pontual.

(h) Para cada regra de previsão a seguir encontre a sensibilidade e especificidade do modelo. Qual das duas regras você prefere?

i. "Preveja 1 se $\hat{\pi}_h > 0.5$ e 0 caso contrário"

ii. "Preveja 1 se $\hat{\pi}_h > 0.35$ e 0 caso contrário"

(i) Usando a Curva ROC decida qual o valor de π^* que deve ser usado na regra de previsão. Para esse valor determine a sensibilidade, especificidade e probabilidade de falso positivo do ajuste.

3. Um supermercado realizou um estudo para avaliar a eficácia de seus cupons de descontos. Para isso 1.000 famílias foram selecionas e cada uma delas recebeu pelos correios 1 cupom de desconto. Os cupons tinham diferentes valores de desconto: 5, 10, 15, 20 e 30 reais. Cada tipo de cupom foi enviado para 200 famílias. Ao final da validade dos cupons o supermercado contabilizou quantos cupons de cada tipo foram utilizados, como mostra a tabela abaixo.

valor do cupom	nº de famílias com esse cupom	nº de cupons utilizados
5	200	30
10	200	55
15	200	70
20	200	100
30	200	137

(a) Ajuste os dados a um modelo de Regressão Logística e encontre as estimativas para os parâmetros β_0 e β_1.

(b) Faça o gráfico de p_j versus x_j e junto com ele coloque a curva ajustada pelo modelo logístico.

(c) Determine a expressão para a função de regressão estimada pelo modelo logístico e em seguida encontre estimativas pontuais para $\pi_h = P(Y_h = 1)$ considerando os cinco diferentes níveis encontrados na amostra: $x_j \in \{5, 10, 15, 20, 30\}$.

(d) Encontre uma estimativa para a razão de chance da variável preditiva do modelo. Interprete o resultado.

(e) Encontre uma estimativa para a razão de chance entre os cupons com R$20 de diferença, por exemplo de R$10 e de R$30. Interprete o resultado.

(f) Faça o teste de Wald e verifique se a utilização ou não do cupom está relacionada com o valor do desconto. Enuncie as hipóteses a serem testadas, encontre o valor da estatística de teste, indique a sua distribuição sob H_0 e tome a decisão final baseada no p-valor do teste.

(g) Ache o valor da *Deviance* do modelo e faça o teste da qualidade do ajuste. Quais as hipóteses que estão sendo testadas e qual o p-valor encontrado?

(h) Construa a tabela de classificação do modelo considerando como critério de corte $\pi^* = 0,5$.

(i) A partir da Curva ROC encontre o melhor valor de π^* para ser usado como critério de corte.

4. Uma grande rede de lojas de material de construção decidiu fazer uma pesquisa para verificar o movimento dos clientes em suas lojas. Para isso a região de atuação dessa rede foi dividida em 110 setores, cada um com raio de 10 milhas. Durante duas semanas o movimentos das lojas foram analisados e, para cada setor, as seguintes informações recolhidas:

   x_1: Número de casas
   x_2: Renda média, em dólares
   x_3: Idade média das casas
   x_4: Distância para o concorrente mais próximo, em milhas
   x_5: Distância para a loja mais próxima, em milhas
   y: Número de clientes do setor que visitaram uma das lojas da rede.

   Os dados referentes aos 110 setores encontram-se no arquivo `CH14TA14.txt` (`https://netfiles.umn.edu/users/nacht001/www/nachtsheim/5th/KutnerData/Chapter%2014%20Data%20Sets/CH14TA14.txt`) e as colunas apresentam as variáveis do problema na seguinte ordem: y, x_1, x_2, x_3, x_4 e x_5.

   (a) Usando o comando `glm` ajuste os dados ao Modelo de Regressão de Poisson considerando a função de ligação logarítmica.

   (b) Encontre as estimativas para cada parâmetro β e em seguida o intervalo de confiança para esses parâmetros.

   (c) Encontre a estimativas pontuais e intervalares para cada risco relativo e faça as devidas interpretações.

   (d) Faça o teste da qualidade do ajuste.

   (e) Encontre as componentes do Desvio, dev_i, e faça o gráfico desses valores versus os índices.

(f) Para os itens a seguir considere um setor com as seguintes características:500 casas, renda média de 30.000 dólares, idade média das casas de 15 anos, distância para o concorrente de 2 milhas e distância para a loja de 3 milhas.

- Encontre uma estimativa pontual para o número médio de visitas em uma das lojas de moradores deste setor em um período de 2 semanas.
- Encontre uma estimativa intervalar para o número médio de visitas em uma das lojas de moradores deste setor em um período de 2 semanas.
- Encontre uma estimativa pontual para a probabilidade de nenhum morador deste setor visitar uma das lojas em um período de 2 semanas.
- Encontre uma estimativa intervalar para a probabilidade de nenhum morador deste setor visitar uma das lojas em um período de 2 semanas.

Lista de Exercícios do Capítulo 4

4.1. Um psicólogo conduziu um estudo para investigar a relação, caso exista, entre a estabilidade emocional de um funcionário (x) e sua capacidade em realizar tarefas de grupo (y). A estabilidade emocional dos funcionários foi medida através de um teste escrito de forma que quanto maior a pontuação no teste maior a estabilidade emocional do funcionário. A capacidade de realizar tarefas em grupo ($y = 1$ se capaz e $y = 0$ se incapaz) foi avaliada a partir do relato de seu supervisor. Foram analisados 27 empregados e as amostras das variáveis x e y encontram-se no arquivo `CH14PR09.txt` (`https://netfiles.umn.edu/users/nacht001/www/nachtsheim/5th/KutnerData/Chapter%2014%20Data%20Sets/CH14PR09.txt`). Assuma que o modelo logístico simples é adequado para esses dados e faça o que se pede.

(a) Encontre os estimadores de máxima verissimilhança para β_0 e β_1. Em seguida determine a expressão para a função de regressão estimada.

(b) Faça o gráfico de dispersão das variáveis x e y e, junto com ele, trace a função de regressão estimada.

(c) Faça a interpretação adequada a partir do valor de $\hat{\beta}_1$.

(d) Qual a probabilidade estimada de que funcionários com nota 550 no teste de estabilidade emocional sejam capaz de realizar tarefas em grupo?

(e) Encontre um intervalo de confiança para $OR_1 = e^{\beta_1}$ com 90% de confiabilidade. Interprete o resultado.

(f) Faça o teste de Wald e verifique se a estabilidade emocional dos funcionário está relacionada com a capacidade do funcionário realizar tarefas em grupo. Enuncie as hipóteses a serem testadas, encontre o valor da estatística de teste, indique a sua distribuição sob H_0 e tome a decisão final baseada no p-valor do teste.

4.2. Uma pesquisa de marketing foi realizada a fim de investigar a possibilidade de uma família comprar um carro novo no próximo ano a partir do modelo logístico. Para isso 33 famílias foram selecionadas e as informações referente sua renda anual (x_1, em milhares de dólares) e a idade do seu veículo mais velho (x_2, em anos) foram recolhidas. Tais famílias foram acompanhadas durante os 12 meses seguintes de forma a constatar se ela comprou um carro no ($y = 1$) ou não ($y = 0$). Os dados referentes a essa pesquisa encontram-se no arquivo `CH14PR13.txt` (`https://netfiles.umn.edu/users/nacht001/www/nachtsheim/5th/KutnerData/Chapter%2014%20Data%20Sets/CH14PR13.txt`), onde a primeira coluna indica os valores de y, a segunda os valores de x_1 e a terceira os valores de x_2 para cada família entrevistada.

Assuma que o modelo logístico múltiplo é adequado para esses dados e faça o que se pede.

(a) Encontre os estimadores de máxima verissimilhança para β_0, β_1 e β_2. Em seguida determine a expressão para a função de regressão estimada.

(b) Faça as interpretações adequadas a partir dos valores de $\hat{\beta}_1$ e $\hat{\beta}_2$.

(c) Qual a probabilidade estimada de uma família comprar um carro novo no próximo ano se a sua renda anual é de $ 50 mil dólares e o seu carro mais velho está com 3 anos?

(d) Encontre um intervalo de confiança para a razão de chance referente a variável renda familiar considerando uma variação de 20 mil dólares com 90% de confiabilidade. Interprete o resultado.

(e) Encontre um intervalo de confiança para a razão de chance referente a variável idade do carro mais velho considerando variação de 2 anos com 90% de confiabilidade. Interprete o resultado.

(f) Faça o teste de Wald e verifique se x_2, a idade do carro mais velho, pode ser retirado do modelo. Enuncie as hipóteses a serem testadas, encontre o valor da estatística de teste, indique a sua distribuição sob H_0 e tome a decisão final baseada no p-valor do teste.

(g) Faça o teste da razão de verossimilhança e verifique se x_2, a idade do carro mais velho, pode ser retirado do modelo. Apresente o modelo completo e o modelo reduzido. Enuncie as hipóteses a serem testadas, encontre o valor da estatística de teste, indique a sua distribuição sob H_0 e tome a decisão final. Qual o p-valor desse teste?

(h) Encontre um intervalo de confiança para a probabilidade de famílias com renda anual 50 mil dólares e carro mais velho com 3 anos comprar um carro novo no próximo ano. Interprete o resultado.

4.3. Uma clinica de saúde local enviou informativos para seus clientes encorajando todos a tomares a vacina contra gripe a tempo de se prevenir contra possível epidemia. O encorajamento foi destinado principalmente aos mais velhos, que têm mais chances de complicação devido a uma gripe. Depois de alguns meses um estudo piloto selecionou 159 clientes dessa clinica, que foram perguntados se eles se vacinaram ou não. O clientes que tomaram a vacina foram codificados por $y = 1$ enquanto os que não tomaram por $Yy0$. Outros dados referentes aos clientes forma coletados: idade (x_1), consciência de saúde (x_2) e sexo (x_3), $x_3 = 1$ para os homens e $x_3 = 0$ para as mulheres. A variável x_2 está sendo representada por um índice tal que quanto maior o seu valor mais consciente em relação a sua saúde o paciente é. Os dados referentes a essa pesquisa pode ser encontrado em `CH14PR14.txt` (`https://netfiles.umn.edu/users/nacht001/www/nachtsheim/5th/KutnerData/Chapter%2014%20Data%20Sets/CH14PR14.txt`), onde a primeira coluna apresenta os valores de y, a segunda de x_1, a terceira de x_2 e a última coluna os valores de x_3.

Assuma que o modelo logístico múltiplo é adequado para esses dados e faça o que se pede.

(a) Encontre os estimadores de máxima verissimilhança para β_0, β_1, β_2 e β_3. Em seguida determine a expressão para a função de regressão estimada.

(b) Faça as interpretações adequadas a partir dos valores de $\hat{\beta}_1$, $\hat{\beta}_2$ e $\hat{\beta}_3$.

(c) Qual a probabilidade estimada de um paciente homem de 55 anos de idade e índice de consciência de saúde 60 ter tomado a vacina contra a gripe?

(d) Encontre um intervalo de confiança para a razão de chance referente a variável idade considerando uma variação de 30 anos com 90% de confiabilidade. Interprete o resultado.

(e) Faça o teste de Wald e verifique se x_3, o sexo do paciente, pode ser retirado do modelo. Enuncie as hipóteses a serem testadas, encontre o valor da estatística de teste, indique a sua distribuição sob H_0 e tome a decisão final baseada no p-valor do teste.

(f) Faça o teste da razão de verossimilhança e verifique se x_3, o sexo do paciente, pode ser retirado do modelo. Apresente o modelo completo e o modelo reduzido. Enuncie as hipóteses a serem testadas, encontre o valor da estatística de teste, indique a sua distribuição sob H_0 e tome a decisão final. Qual o p-valor desse teste?

(g) Encontre um intervalo de confiança para a probabilidade de mulheres de 65 anos e índice de consciência de saúde igual à 50 ter tomado a vacina contra a gripe. Interprete o resultado.

4.4. Para os bancos de dados dos exercícios 4.2 e 4.3 faça o que se pede.

(a) Escolha um método de seleção de variáveis e determine quais variáveis preditivas devem fazer parte do modelo de Regressão Logística.

(b) Apresente a função de regressão estimada do modelo final e faça as interpretações adequadas a partir dos valores estimados para cada β_k.

(c) Encontre o valor de corte π^* de forma a maximizar a soma da sensibilidade e especificidade do modelo ajustado.

4.5. Um experimento busca avaliar o efeito de uma certa substância tóxica. Para isso foram usados 1.500 insetos, divididos em 6 grupos de 250. Cada grupo recebeu uma dosagem diferente da substancia tóxica e depois de 24 horas foram contabilizados o número de insetos mortos em cada grupo. A tabela a seguir apresenta o resultado desse experimento.

Dosagem :	1	2	3	4	5	6
Total de insetos:	250	250	250	250	250	250
Nº de insetos mortos:	28	53	93	126	172	197

(a) Faça o gráfico de p_j versus x_j para cada um dos 6 níveis, $x_j \in \{1, 2, 3, 4, 5, 6\}$.

(b) Ajuste um modelo logístico aos dados e em seguida faça a função de regressão estimada junto com o gráfico do item acima.

(c) Encontre um intervalo de confiança para a razão de chance referente a variável dosagem. Interprete o resultado.

(d) De acordo com o modelo estimado, o quando aumenta as chances de um inseto morrer se a dosagem do substância tóxica aumentar em 3 unidade?

(e) Qual a probabilidade de um inseto morrer se ele for exposto a uma dosagem de 3.5 dessa substância, de acordo com o modelo ajustado?

(f) Encontre o valor da sensibilidade e da especificidade do modelo se o critério de corte adotado for $\pi^* = 0,5$.

(g) A partir da Curva ROC encontre o valor para π^* de forma a maximizar a soma da sensibilidade com a especificidade. Para esse valor de π^* determine a sensibilidade, a especificidade e a probabilidade do modelo retornar um falso positivo.

4.6. Este exercício se refere aos dados do exercício 1.3 do Capítulo 1. Em vez de usar o Modelo Normal vamos adotar o Modelo de Poisson, uma vez que y é o número de ampolas quebradas e por isso é uma variável de contagem.

(a) Faça o gráfico de dispersão de y versus x.

(b) Ajuste uma Regressão de Poisson aos dados e junto com o gráfico do item acima desenhe a curva da função de regressão estimada.

(c) Obtenha as componentes do desvio dev_i e faça seu gráfico versus o índice. O modelo parece bem ajustado?

(d) Faça o teste da qualidade do ajuste e verifique se o modelo está bem ajustado.

(e) Encontre a estimativas pontuais e intervalares para cada risco relativo e faça as devidas interpretações.

(f) A partir do modelo ajustado encontre uma estimativa pontual para o número médio de ampolas quebras quando $x = 0, 1, 2, 3$.

(g) Encontre uma estimativa pontual para a probabilidade de haver 10 ou menos ampolas quebradas quando $x = 0$.

4.7. Um grupo de geriatras está interessado em estudar o efeito de algumas intervenções no número de quedas em idosos. Para isso 100 indivíduos, com pelos menos 65 anos e boa saúde, foram divididos em 2 grupo que receberam diferentes intervenções: um grupo recebeu somente instruções ($x_1 = 0$) e o outro grupo recebeu instruções e treinos aeróbicos ($x_1 = 1$). Todos os indivíduos foram acompanhados durante 6 meses e o número de quedas neste período contabilizado (y). Outras informações referentes a cada indivíduo também foram coletadas: o sexo ($x_2 = 0$ para as mulheres e $x_2 = 1$ para os homens), um índice de equilíbrio (x_3, quanto maior mais equilibrado é o indivíduo) e um índice de força (x_4, quanto maior maios forte é o indivíduo). Os dados referentes a essa pesquisa pode ser encontrado em `CH14PR39.txt` (`https://netfiles.umn.edu/users/nacht001/www/nachtsheim/5th/KutnerData/Chapter%2014%20Data%20Sets/CH14PR39.txt`), onde as variáveis são apresentadas na seguinte ordem: y, x_1, x_2, x_3 e x_4.

(a) Ajuste uma Regressão de Poisson aos dados.

(b) Faça a seleção do modelo e, para o modelo final, apresente as estimativas para os coeficientes e a função de regressão estimada.

(c) Obtenha as componentes do desvio dev_i e faça seu gráfico versus o índice. O modelo parece bem ajustado?

(d) Faça o teste da qualidade do ajuste e verifique se o modelo está bem ajustado.

(e) Encontre a estimativas pontuais e intervalares para cada risco relativo e faça as devidas interpretações.

(f) Encontre uma estimativa pontual e uma intervalar para o número médio de quedas em mulheres que receberam apenas instruções (e não receberam treino aeróbico) com índice de equilíbrio igual a 50 e índice de força igual a 60.

(g) Encontre uma estimativa pontual para a probabilidade de mulheres, com as mesmas características do item acima, não sofrerem queda alguma no período de 6 meses.

Referências Bibliográficas

[Kutner et al., 2005] Kutner, M. H., Nachtsheim, C. J., Neter, J., and Li, W. (2005). *Applied Linear Statistical Models*. McGraw-Hill/Irwin, 5th edition.

[Montgomery et al., 2012] Montgomery, D. C., Peck, E. A., and Vining, G. G. (2012). *Introduction to Linear Regression Analysis*. A John Wiley & Sons, 5th edition.

Apêndice A

Tabelas

Tabela A.1: Percentis da Distribuição Normal Padrão

$$P(Z \leq z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx$$

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998

Tabela A.2: Percentis da Distribuição t

$$\mathrm{P}(T \leq t) = \int_{-\infty}^{t} \frac{\Gamma((r+1)/2)}{\sqrt{\pi r}\,\Gamma(r/2)} \frac{1}{(1+x^2/r)^{(r+1)/2}}\, dx$$

	P($T \leq t$)						P($T \leq t$)				
df	**0.900**	**0.950**	**0.975**	**0.990**	**0.995**	**df**	**0.900**	**0.950**	**0.975**	**0.990**	**0.995**
1	3.078	6.314	12.706	31.821	63.657	**35**	1.306	1.690	2.030	2.438	2.724
2	1.886	2.920	4.303	6.965	9.925	**40**	1.303	1.684	2.021	2.423	2.704
3	1.638	2.353	3.182	4.541	5.841	**45**	1.301	1.679	2.014	2.412	2.690
4	1.533	2.132	2.776	3.747	4.604	**50**	1.299	1.676	2.009	2.403	2.678
5	1.476	2.015	2.571	3.365	4.032	**55**	1.297	1.673	2.004	2.396	2.668
6	1.440	1.943	2.447	3.143	3.707	**60**	1.296	1.671	2.000	2.390	2.660
7	1.415	1.895	2.365	2.998	3.499	**65**	1.295	1.669	1.997	2.385	2.654
8	1.397	1.860	2.306	2.896	3.355	**70**	1.294	1.667	1.994	2.381	2.648
9	1.383	1.833	2.262	2.821	3.250	**75**	1.293	1.665	1.992	2.377	2.643
10	1.372	1.812	2.228	2.764	3.169	**80**	1.292	1.664	1.990	2.374	2.639
11	1.363	1.796	2.201	2.718	3.106	**85**	1.292	1.663	1.988	2.371	2.635
12	1.356	1.782	2.179	2.681	3.055	**90**	1.291	1.662	1.987	2.368	2.632
13	1.350	1.771	2.160	2.650	3.012	**95**	1.291	1.661	1.985	2.366	2.629
14	1.345	1.761	2.145	2.624	2.977	**100**	1.290	1.660	1.984	2.364	2.626
15	1.341	1.753	2.131	2.602	2.947	**105**	1.290	1.659	1.983	2.362	2.623
16	1.337	1.746	2.120	2.583	2.921	**110**	1.289	1.659	1.982	2.361	2.621
17	1.333	1.740	2.110	2.567	2.898	**115**	1.289	1.658	1.981	2.359	2.619
18	1.330	1.734	2.101	2.552	2.878	**120**	1.289	1.658	1.980	2.358	2.617
19	1.328	1.729	2.093	2.539	2.861	**130**	1.288	1.657	1.978	2.355	2.614
20	1.325	1.725	2.086	2.528	2.845	**140**	1.288	1.656	1.977	2.353	2.611
21	1.323	1.721	2.080	2.518	2.831	**150**	1.287	1.655	1.976	2.351	2.609
22	1.321	1.717	2.074	2.508	2.819	**160**	1.287	1.654	1.975	2.350	2.607
23	1.319	1.714	2.069	2.500	2.807	**170**	1.287	1.654	1.974	2.348	2.605
24	1.318	1.711	2.064	2.492	2.797	**180**	1.286	1.653	1.973	2.347	2.603
25	1.316	1.708	2.060	2.485	2.787	**190**	1.286	1.653	1.973	2.346	2.602
26	1.315	1.706	2.056	2.479	2.779	**200**	1.286	1.653	1.972	2.345	2.601
27	1.314	1.703	2.052	2.473	2.771	**210**	1.286	1.652	1.971	2.344	2.599
28	1.313	1.701	2.048	2.467	2.763	**220**	1.285	1.652	1.971	2.343	2.598
29	1.311	1.699	2.045	2.462	2.756	**230**	1.285	1.652	1.970	2.343	2.597
30	1.310	1.697	2.042	2.457	2.750	∞	1.282	1.645	1.960	2.326	2.576

Tabela A.3: Percentis da Distribuição Qui-quadrado

$$\mathrm{P}(\chi_r \leq q) = \int_0^q \frac{1}{2^{r/2}\Gamma(r/2)}\, x^{\frac{r}{2}-1} e^{-\frac{x}{2}}\, dx$$

	P($T \leq t$)									
df	**0.005**	**0.010**	**0.025**	**0.050**	**0.100**	**0.900**	**0.950**	**0.975**	**0.990**	**0.995**
1	0.000	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.60
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.34	12.84
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.14	13.28	14.86
5	0.412	0.554	0.831	1.145	1.610	9.236	11.07	12.83	15.09	16.75
6	0.676	0.872	1.237	1.635	2.204	10.64	12.59	14.45	16.81	18.55
7	0.989	1.239	1.690	2.167	2.833	12.02	14.07	16.01	18.48	20.28
8	1.344	1.646	2.180	2.733	3.490	13.36	15.51	17.53	20.09	21.95
9	1.735	2.088	2.700	3.325	4.168	14.68	16.92	19.02	21.67	23.59
10	2.156	2.558	3.247	3.940	4.865	15.99	18.31	20.48	23.21	25.19
11	2.603	3.053	3.816	4.575	5.578	17.28	19.68	21.92	24.72	26.76
12	3.074	3.571	4.404	5.226	6.304	18.55	21.03	23.34	26.22	28.30
13	3.565	4.107	5.009	5.892	7.042	19.81	22.36	24.74	27.69	29.82
14	4.075	4.660	5.629	6.571	7.790	21.06	23.68	26.12	29.14	31.32
15	4.601	5.229	6.262	7.261	8.547	22.31	25.00	27.49	30.58	32.80
16	5.142	5.812	6.908	7.962	9.312	23.54	26.30	28.85	32.00	34.27
17	5.697	6.408	7.564	8.672	10.09	24.77	27.59	30.19	33.41	35.72
18	6.265	7.015	8.231	9.390	10.86	25.99	28.87	31.53	34.81	37.16
19	6.844	7.633	8.907	10.12	11.65	27.20	30.14	32.85	36.19	38.58
20	7.434	8.260	9.591	10.85	12.44	28.41	31.41	34.17	37.57	40.00
21	8.034	8.897	10.28	11.59	13.24	29.62	32.67	35.48	38.93	41.40
22	8.643	9.542	10.98	12.34	14.04	30.81	33.92	36.78	40.29	42.80
23	9.260	10.20	11.69	13.09	14.85	32.01	35.17	38.08	41.64	44.18
24	9.886	10.86	12.40	13.85	15.66	33.20	36.42	39.36	42.98	45.56
25	10.52	11.52	13.12	14.61	16.47	34.38	37.65	40.65	44.31	46.93
26	11.16	12.20	13.84	15.38	17.29	35.56	38.89	41.92	45.64	48.29
27	11.81	12.88	14.57	16.15	18.11	36.74	40.11	43.19	46.96	49.64
28	12.46	13.56	15.31	16.93	18.94	37.92	41.34	44.46	48.28	50.99
29	13.12	14.26	16.05	17.71	19.77	39.09	42.56	45.72	49.59	52.34
30	13.79	14.95	16.79	18.49	20.60	40.26	43.77	46.98	50.89	53.67
40	20.71	22.16	24.43	26.51	29.05	51.81	55.76	59.34	63.69	66.77
50	27.99	29.71	32.36	34.76	37.69	63.17	67.50	71.42	76.15	79.49
60	35.53	37.48	40.48	43.19	46.46	74.40	79.08	83.30	88.38	91.95
70	43.28	45.44	48.76	51.74	55.33	85.53	90.53	95.02	100.4	104.2
80	51.17	53.54	57.15	60.39	64.28	96.58	101.9	106.6	112.3	116.3
90	59.20	61.75	65.65	69.13	73.29	107.6	113.1	118.1	124.1	128.3
100	67.33	70.06	74.22	77.93	82.36	118.5	124.3	129.6	135.8	140.2
110	75.55	78.46	82.87	86.79	91.47	129.4	135.5	140.9	147.4	151.9
120	83.85	86.92	91.57	95.70	100.6	140.2	146.6	152.2	159.0	163.6
130	92.22	95.45	100.3	104.7	109.8	151.0	157.6	163.5	170.4	175.3
140	100.7	104.0	109.1	113.7	119.0	161.8	168.6	174.6	181.8	186.8
150	109.1	112.7	118.0	122.7	128.3	172.6	179.6	185.8	193.2	198.4
160	117.7	121.3	126.9	131.8	137.5	183.3	190.5	196.9	204.5	209.8
170	126.3	130.1	135.8	140.8	146.8	194.0	201.4	208.0	215.8	221.2
180	134.9	138.8	144.7	150.0	156.2	204.7	212.3	219.0	227.1	232.6
190	143.5	147.6	153.7	159.1	165.5	215.4	223.2	230.1	238.3	244.0
200	152.2	156.4	162.7	168.3	174.8	226.0	234.0	241.1	249.4	255.3
210	161.0	165.3	171.8	177.5	184.2	236.7	244.8	252.0	260.6	266.5
220	169.7	174.2	180.8	186.7	193.6	247.3	255.6	263.0	271.7	277.8
230	178.5	183.1	189.9	195.9	203.0	257.9	266.4	273.9	282.8	289.0