2 O Modelo de Regressão Linear Simples

Este capítulo apresenta o modelo de regressão linear simples, que consiste basicamente da equação da reta à qual é adicionada uma variável aleatória normal com média zero e variância σ^2. A partir deste modelo, toda uma teoria é apresentada e discutida. A análise estatística do modelo de regressão linear simples pode ser sintetizada em uma simples pergunta: há evidência estatística de uma relação linear entre a variável preditora x_i e a variável resposta Y_i?

Seja o seguinte modelo de regressão linear simples, ou a seguinte equação de uma reta de regressão:

$$Y = \beta_0 + \beta_1 x + \varepsilon \tag{2.1}$$

em que β_0 e β_1 são parâmetros do modelo (constantes desconhecidas) e ε é uma variável aleatória, $E(\varepsilon) = 0$, $Var(\varepsilon) = \sigma^2$. Como consequência, Y também é uma variável aleatória. Neste caso, a forma *politicamente correta*, considerando x um valor conhecido (ou pré-conhecido) é:

$$Y|x = \beta_0 + \beta_1 x + \varepsilon$$

$$\begin{aligned} E(Y|x) &= E(\beta_0 + \beta_1 x + \varepsilon|x) \\ &= E(\beta_0 + \beta_1 x|x) + E(\varepsilon|x) \\ &= \beta_0 + \beta_1 x \end{aligned} \tag{2.2}$$

$$\begin{aligned} Var(Y|x) &= Var(\beta_0 + \beta_1 x + \varepsilon|x) \\ &= Var(\varepsilon|x) \\ &= \sigma^2 \end{aligned} \tag{2.3}$$

As equações acima podem ser interpretadas da seguinte forma: uma vez conhecido o valor de x o comportamento da variável aleatória Y apresenta uma média definida pela equação de regressão e uma variância (ou dispersão) constante.

2.1 Estimação dos parâmetros β_0 e β_1 a partir de uma amostra de tamanho n

Considere agora, uma amostra de tamanho n de pares de observações independentes, (x_i, y_i). Neste caso a equação de regressão 2.1 deve ser re-definida como:

$$Y_i|x_i = \beta_0 + \beta_1 x_i + \varepsilon_i,\ i = 1,2,\ldots,n \tag{2.4}$$

onde ε_i são variáveis aleatórias independentes e normais com média zero e variância σ^2.

Semelhante à teoria apresentada no capítulo 1, define-se a Soma dos Quadrados dos Erros:

$$\begin{aligned} SQE(\beta_0,\beta_1) &= \sum_{i=1}^{n}[y_i - (\beta_0 + \beta_1 x_i)]^2 \\ &= \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2 \end{aligned} \tag{2.5}$$

onde os estimadores, $\hat{\beta}_0$ e $\hat{\beta}_1$, são os valores que minimizam a soma dos quadrados dos erros. Os estimadores são calculados a partir das derivadas parciais:

$$\begin{aligned} \frac{\partial SQE(\beta_0,\beta_1)}{\partial \beta_0} &= \sum_{i=1}^{n} 2[y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)]^{2-1}(-1) \\ &= -2\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \end{aligned} \tag{2.6}$$

$$\begin{aligned} \frac{\partial SQE(\beta_0,\beta_1)}{\partial \beta_1} &= \sum_{i=1}^{n} 2[y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)]^{2-1}(-x_i) \\ &= -2\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)x_i = 0 \end{aligned} \tag{2.7}$$

Simplificando, obtém-se um sistema linear de duas equações:

$$\begin{cases} n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \\ \hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} y_i x_i \end{cases} \tag{2.8}$$

ou na forma:

$$\begin{cases} n\hat{\beta}_0 + \hat{\beta}_1 n\bar{x} = n\bar{y} \\ \hat{\beta}_0 n\bar{x} + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} y_i x_i \end{cases} \tag{2.9}$$

A Equação 2.9 representa um sistema de equações lineares dos estimadores $\hat{\beta}_0$ e $\hat{\beta}_1$, cuja solução é:

$$\begin{aligned} \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \\ \hat{\beta}_1 &= \frac{\sum_{i=1}^{n} y_i x_i - \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)\left(\sum_{i=1}^{n} x_i\right)}{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}} \end{aligned} \tag{2.10}$$

ou na forma:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} y_i (x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}, \qquad \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad e \qquad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \tag{2.11}$$

2.2 Propriedades dos Estimadores

As equações definidas em 2.9 mostram que os estimadores são funções lineares das variáveis aleatórias Y_i e, como consequência, também são variáveis aleatórias com distribuição normal. Então, podemos determinar seus parâmetros de média e variância como:

$$\hat{\beta}_1 = \sum_{i=1}^{n} \left[\frac{(x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right] Y_i = \sum_{i=1}^{n} c_i Y_i \tag{2.12}$$

$$\begin{aligned} E(\hat{\beta}_1) &= E\left(\sum_{i=1}^{n} c_i Y_i\right) = \sum_{i=1}^{n} c_i E(Y_i) \\ &= \sum_{i=1}^{n} c_i(\beta_0 + \beta_1 x_i) \\ &= \beta_0 \sum_{i=1}^{n} c_i + \beta_1 \sum_{i=1}^{n} c_i x_i \end{aligned} \tag{2.13}$$

onde

$$\sum_{i=1}^{n} c_i = \frac{\sum_{i=1}^{n} (x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{1}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \sum_{i=1}^{n} (x_i - \bar{x}) = 0 \tag{2.14}$$

Uma vez que $\sum_{i=1}^{n} (x_i - \bar{x}) = 0$, a análise do segundo termo, $\sum_{i=1}^{n} c_i x_i$, pode ser realizada separadamente para o numerador e o denominador:

$$\sum_{i=1}^{n} c_i x_i = \frac{1}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \sum_{i=1}^{n} x_i (x_i - \bar{x}) \tag{2.15}$$

$$\begin{aligned} \sum_{i=1}^{n} x_i (x_i - \bar{x}) &= \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i \bar{x} \\ &= \sum_{i=1}^{n} x_i^2 - \bar{x} \sum_{i=1}^{n} x_i \\ &= \sum_{i=1}^{n} x_i^2 - n\bar{x}^2 \end{aligned} \tag{2.16}$$

Por outro lado, analisando o denominador:

$$\begin{aligned} \sum_{i=1}^{n} (x_i - \bar{x})^2 &= \sum_{i=1}^{n} \left(x_i^2 - 2x_i\bar{x} + \bar{x}^2\right) \\ &= \sum_{i=1}^{n} x_i^2 - 2\bar{x} \sum_{i=1}^{n} x_i + n\bar{x}^2 \\ &= \sum_{i=1}^{n} x_i^2 - 2n\bar{x}^2 + n\bar{x}^2 \\ &= \sum_{i=1}^{n} x_i^2 - n\bar{x}^2 \end{aligned} \tag{2.17}$$

logo, é possível concluir que $\sum_{i=1}^{n} c_i x_i = 1$ e, portanto, $E(\hat{\beta}_1) = \beta_1$. No caso do intercepto, $\hat{\beta}_0$:

$$\begin{aligned} E(\hat{\beta}_0) &= E(\bar{y} - \hat{\beta}_1 \bar{x}) = E(\bar{y}) - E(\hat{\beta}_1)\bar{x} \\ &= \frac{1}{n} \sum_{i=1}^{n} y_i - \bar{x}\beta_1 = \frac{1}{n} \left(\sum_{i=1}^{n} \beta_0 + \beta_1 x_i\right) - \bar{x}\beta_1 \\ &= \beta_0 + \frac{1}{n}\beta_1 \sum_{i=1}^{n} x_i - \bar{x}\beta_1 = \beta_0 + \beta_1\bar{x} - \beta_1\bar{x} \\ &= \beta_0 \end{aligned} \tag{2.18}$$

Se as observações Y_i são independentes e possuem a mesma variância $Var(Y_i) = \sigma^2$, então:

$$Var(\hat{\beta}_1) = Var\left(\sum_{i=1}^{n} c_i Y_i\right) = \sum_{i=1}^{n} c_i^2 Var(Y_i) \tag{2.19}$$

$$= \sigma^2 \sum_{i=1}^{n} c_i^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{2.20}$$

e

$$Var(\hat{\beta}_0) = Var(\overline{Y} - \hat{\beta}_1 \bar{x}) \tag{2.21}$$

$$= Var(\overline{Y}) + \bar{x}^2 Var(\hat{\beta}_1) - 2\bar{x} Cov(\overline{Y}, \hat{\beta}_1) \tag{2.22}$$

onde $Var(\overline{Y}) = \sigma^2/n$ e $Cov(\overline{Y}, \hat{\beta}_1) = 0$:

$$Cov(\overline{Y}, \hat{\beta}_1) = Cov\left(\frac{1}{n}\sum_{i=1}^{n} Y_i, \sum_{j=1}^{n} c_j Y_j\right) \tag{2.23}$$

$$= \frac{1}{n} Cov\left(\sum_{i=1}^{n} Y_i, \sum_{j=1}^{n} c_j Y_j\right) \tag{2.24}$$

$$= \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} c_j Cov(Y_i, Y_j) \tag{2.25}$$

$$\tag{2.26}$$

mas, como:

$$Cov(Y_i, Y_j) = \begin{cases} \sigma^2, & \text{se } i = j \\ 0, & \text{se } i \neq j \end{cases} \tag{2.27}$$

então $Cov(\overline{Y}, \hat{\beta}_1) = \frac{1}{n}\sigma^2 \sum_{i=1}^{n} c_i = 0$, pois $\sum_{i=1}^{n} c_i = 0$, como mostrado na equação 2.14.
Conclui-se que:

$$Var(\hat{\beta}_0) = Var(\bar{y}) + \bar{x}^2 Var(\hat{\beta}_1) \tag{2.28}$$

$$= \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right) \tag{2.29}$$

Observação: segundo o **Teorema de Gauss-Markov** (SEBER; LEE, 2012; MONTGOMERY; PECK; VINING, 2012), assumindo que $E(\varepsilon_i) = 0$, $Var(\varepsilon_i) = \sigma^2$ e independência, $Cov(\varepsilon_i, \varepsilon_j) = 0 \; \forall \; i \neq j$, os estimadores de mínimos quadrados são não viciados e de mínima variância entre os estimadores lineares.

O estimador de σ^2 é:

$$S^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2} \quad \text{(maiores detalhes no capítulo 3)} \tag{2.30}$$

2.3 Teste de Hipótese para β_1

Neste caso, vamos assumir que $\varepsilon_i \sim NIID(0,\sigma^2)$ (Normais, Independentes e Identicamente Distribuídos). O objetivo é testar a hipótese nula de que não há correlação linear entre x_i e Y_i, ou seja, $H_0: \ \beta_1 = 0$. Então, recapitulando as propriedades dos estimadores:

$$\hat{\beta}_0 \sim N\left[\beta_0, \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)\right] \tag{2.31}$$

$$\hat{\beta}_1 \sim N\left(\beta_1, \sigma^2 \frac{1}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right) \tag{2.32}$$

e definindo as hipóteses nula e alternativa:

$$\begin{cases} H_0 : \beta_1 = 0 \\ H_1 : \beta_1 \neq 0 \end{cases} \tag{2.33}$$

Neste caso, a estatística de teste é: $t_{obs} = \frac{\hat{\beta}_1 - \beta_1^{H_0}}{s(\hat{\beta}_1)} = \frac{\hat{\beta}_1}{s(\hat{\beta}_1)}$. A hipótese nula é rejeitada se $|t_{obs}| \geq t_{\alpha/2,n-2}$. Também é possível calcular o nível descritivo do teste.

2.4 Intervalo de Confiança para a Resposta Média

Da mesma forma que os estimadores $\hat{\beta}_0$ e $\hat{\beta}_1$ podem ser definidos como variáveis aleatórias, a combinação linear dos mesmos, ou seja, a média estimada: $\hat{\mu}_{Y|x_0} = \hat{\beta}_0 + \hat{\beta}_1 x_0$ também é uma variável aleatória. Como consequência, podemos calcular as suas propriedades (média e variância) e estimar um intervalo de confiança a partir dessas propriedades.

$$\begin{aligned} \hat{\mu}_{Y|x_0} &= \hat{E}(Y|x_0) \\ &= \hat{\beta}_0 + \hat{\beta}_1 x_0 \end{aligned} \tag{2.34}$$

$$\begin{aligned} E(\hat{\mu}_{Y|x_0}) &= E(\hat{\beta}_0 + \hat{\beta}_1 x_0) \\ &= E(\hat{\beta}_0) + E(\hat{\beta}_1) x_0 \\ &= \beta_0 + \beta_1 x_0 \end{aligned} \tag{2.35}$$

$$\begin{aligned} Var(\hat{\mu}_{Y|x_0}) &= Var(\hat{\beta}_0 + \hat{\beta}_1 x_0) \\ &= Var[\bar{y} + \hat{\beta}_1(x_0 - \bar{x})] \end{aligned} \tag{2.36}$$

em que $Cov(\bar{y},\hat{\beta}_1) = 0$, então:

$$Var(\hat{\mu}_{Y|x_0}) = \frac{\sigma^2}{n} + \sigma^2 \frac{(x_0 - \bar{x}^2)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \sigma^2\left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right] \tag{2.37}$$

2.5 Predição de Novas Observações

Neste caso, desejamos definir um intervalo de predição considerando as incertezas associadas à estimativa da reta de regressão e da componente do erro, definida no modelo de regressão segundo Equação 2.1. Podemos então definir a variáveis de interesse como sendo:

$$Y_0|x_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0 + \epsilon \tag{2.38}$$
$$= \hat{\mu}_0 + \epsilon \tag{2.39}$$

A incerteza sobre $Y_0|x_0$ pode ser obtida como:

$$\begin{aligned} Var(Y_0|x_0) &= Var(\hat{\mu}_0 + \epsilon) \\ &= Var(\hat{\mu}_0) + Var(\epsilon) \\ &= \sigma^2 \left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right] \end{aligned} \tag{2.40}$$

De forma alternativa, seja a seguinte variável aleatória:

$$\varphi = Y_0 - \hat{Y}_0 \tag{2.41}$$

$$\begin{aligned} Var(\varphi) &= Var(Y_0 - \hat{Y}_0) \\ &= Var(Y_0) + Var(\hat{Y}_0) \\ &= \sigma^2 \left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right] \end{aligned} \tag{2.42}$$

Se utilizarmos $\hat{Y}_0$ para predizer Y_0, então o erro padrão de $\varphi = Y_0 - \hat{Y}_0$ é a estatística apropriada para calcular o intervalo de predição.

2.6 Análise de Variância do Modelo de Regressão Linear Simples

A análise variância está associada à decomposição da soma dos quadrados totais em duas componentes: a soma dos quadrados de regressão e a soma dos quadrados dos resíduos:

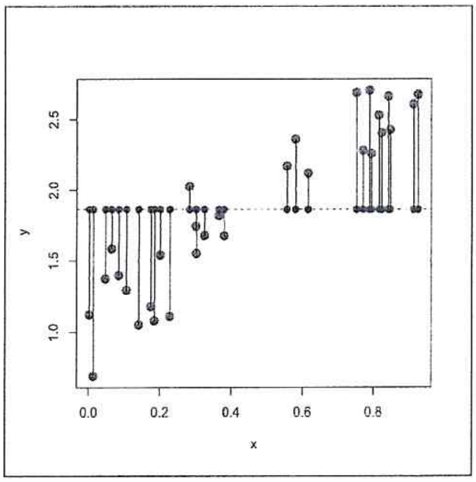
$$\sum_i (y_i - \bar{y})^2 = \sum_i (\hat{y}_i - \bar{y})^2 + \sum_i (y_i - \hat{y}_i)^2 \tag{2.43}$$

ou na forma:

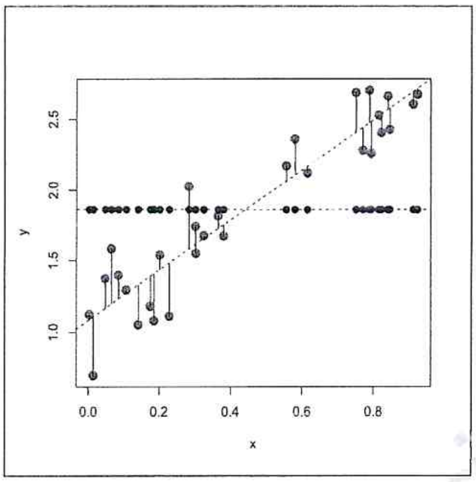
$$SQ_{Total} = SQ_{Regressao} + SQ_{Residuos}$$

O primeiro termo representa a soma dos quadrados totais, $SQ_T = \sum_i (y_i - \bar{y})^2$, o segundo termo é a soma dos quadrados de regressão, $SQ_{Reg} = \sum_i (\hat{y}_i - \bar{y})^2$, e o terceiro termo é a soma dos quadrados dos resíduos, $SQ_{res} = \sum_i (y_i - \hat{y}_i)^2$. **Esta decomposição só é possível se o parâmetro de intercepto (β_0) estiver incluído no modelo, caso contrário a equação 2.43 não se aplica.** A demonstração será apresentada na formulação matricial, no capítulo 3.

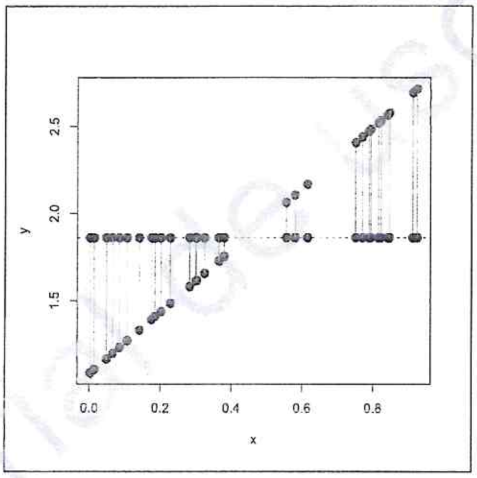
Uma análise visual da decomposição da soma dos quadrados totais é mostrada na figura 2.1.



(a) Soma dos Quadrados Totais é calculada a partir das diferenças entre os valores observados e a média amostral, ou seja, é uma medida da dispersão dos dados com relação à média amostral. Mensura o quanto da dispersão dos dados a média amostral não é capaz de explicar, $SQ_T = \sum_i (y_i - \bar{y})^2$.



(b) Soma dos Quadrados dos Resíduos é calculada a partir das diferenças entre os valores observados e os valores estimados pelo modelo de regressão linear simples. Mensura o quanto da dispersão dos dados o modelo de regressão não é capaz de explicar, $SQ_{res} = \sum_i (y_i - \hat{y}_i)^2$.



(c) Soma dos Quadrados de Regressão é calculada a partir das diferenças entre os valores estimados pelo modelo de regressão linear simples e a média amostral. Mensura o quanto da dispersão dos dados com relação ao modelo da média é explicado pelo modelo de regressão linear simples, $SQ_{Reg} = \sum_i (\hat{y}_i - \bar{y})^2$.

Figura 2.1: Decomposição da soma dos quadrados totais em soma dos quadrados de regressão e soma dos quadrados dos resíduos.

A tabela de variância (ANOVA - *ANalisys Of VAriance*), para o modelo de regressão linear simples, apresenta uma análise estatística da decomposição da soma dos quadrados totais.

A hipótese nula para a tabela de análise de variância (Tabela 2.1) para o modelo de regressão linear simples é $H_0: \beta_1 = 0$ e, portanto, o valor-P obtido pela tabela ANOVA é igual ao resultado

Tabela 2.1: Tabela de análise de variância (ANOVA) para o modelo de regressão linear simples

Fonte	graus de liberdade	Soma dos Quadrados	Quadrados Médios	Estatística F	valor-P
Regressão	1	$SQ_{Reg} = \sum_i (\hat{y}_i - \bar{y})^2$	$QM_{Reg} = SQ_{Reg}$	F=$\frac{QM_{Reg}}{QM_{Res}}$	-
Erro	n-2	$SQ_{Res} = \sum_i (y_i - \hat{y}_i)^2$	$QM_{Res} = \frac{SQ_{Res}}{n-2}$		
Total	n-1	$SQ_T = \sum_i (y_i - \bar{y})^2$	$QM_T = \frac{SQ_T}{n-1}$		

do teste de hipótese utilizando as propriedades estatísticas do estimador $\hat{\beta}_1$. No caso da tabela de variância, a hipótese nula é rejeitada se $F_o > F_{(1,n-1,\alpha)}$

2.7 O Coeficiente de Determinação R^2

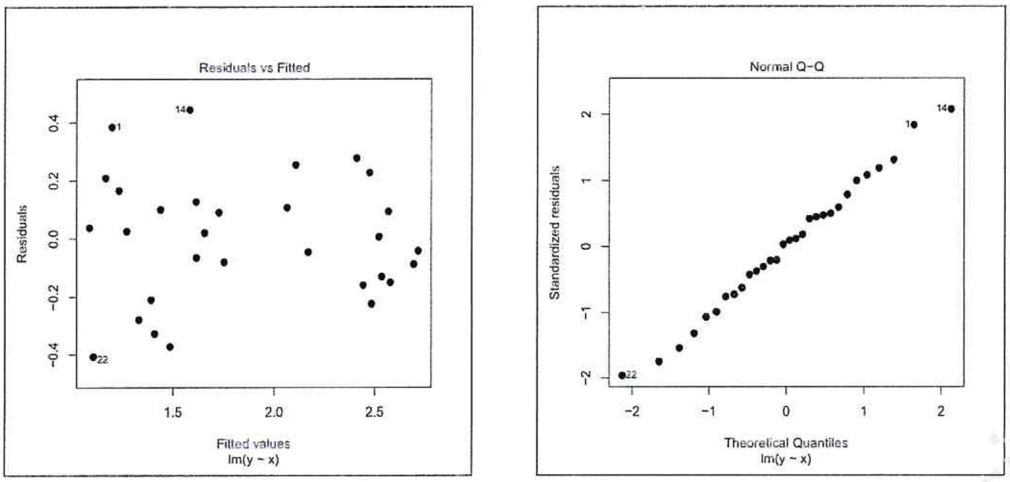
O coeficiente de determinação, conhecido por R^2, é a razão entre a soma dos quadrados de regressão e soma dos quadrados totais. Em termos práticos, o R^2 representa a percentagem da dispersão (ou variabilidade) dos dados com relação ao modelo da média amostral que é explicada (ou *absorvida*) pelo modelo de regressão linear simples.

$$R^2 = \frac{SQ_{Reg}}{SQ_T} = 1 - \frac{SQ_{Res}}{SQ_T} \tag{2.44}$$

onde $0 \leq R^2 \leq 1$. O coeficiente de determinação R^2 é uma medida de ajuste e **não** deve ser utilizado como um critério para validação estatística de um modelo.

2.8 A análise dos resíduos (r_i) no modelo de regressão linear simples

O resíduo é definido como a estimativa do erro, ou seja: $r_i = \hat{e}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$. Vale destacar que a variável aleatória erro (ϵ) define o comportamento aleatória da variável Y_i, e que o estimador da variância ($\hat{\sigma}^2$) é proporcional à soma dos quadrados dos resíduos. Uma vez que a suposição da distribuição do erro é normal, testes de normalidade podem ser aplicados aos resíduos bem como análises empíricas para verificar a independência entre a variância estimada e a média estimada, ou seja, a homocedasticidade dos erros, como ilustra a figura 2.2.



(a) Resíduos versus média estimada.

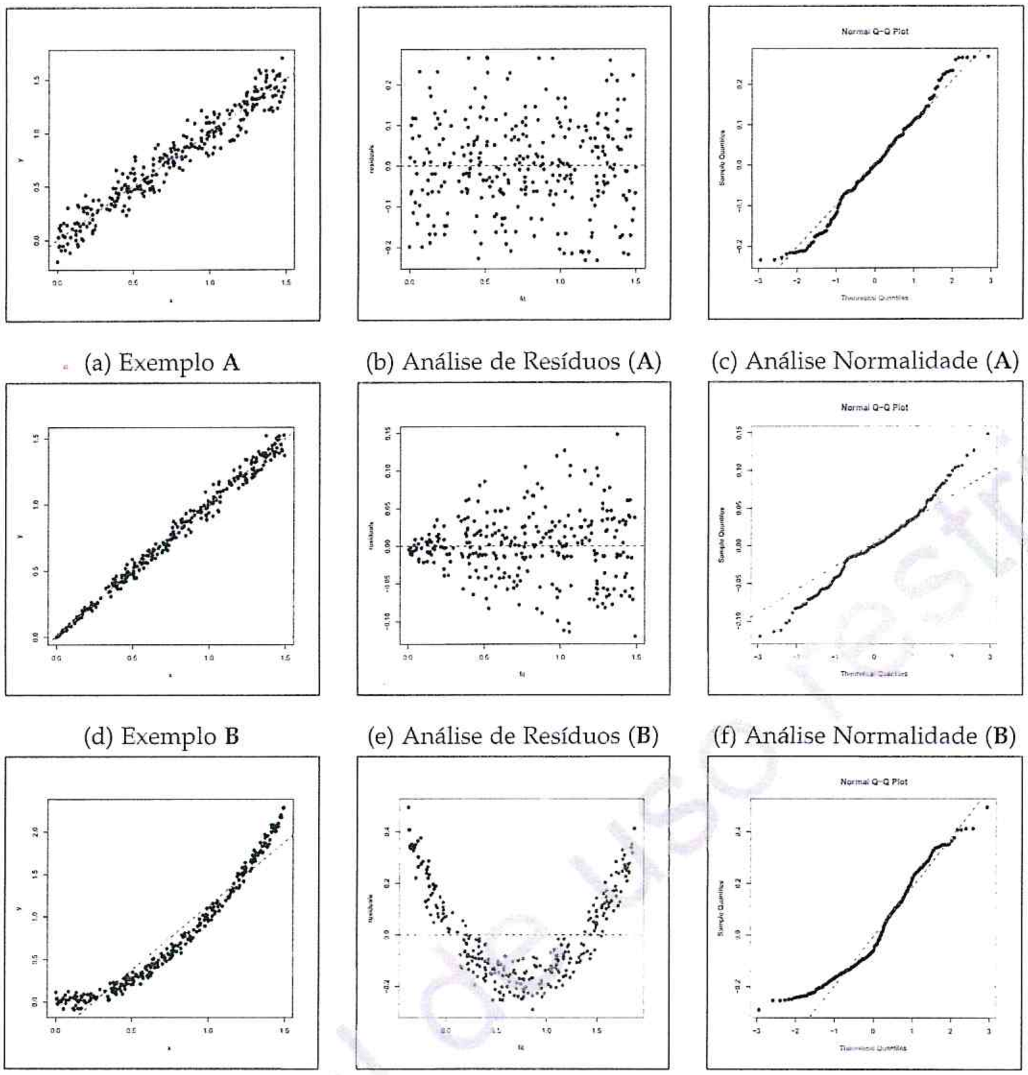
(b) Gráfico de probabilidade normal dos resíduos.

Figura 2.2: Análise de resíduos para verificar a não correlação entre resíduo e média estimada, homocedasticidade, e seu comportamento segundo uma distribuição normal.

A suposição do modelo de regressão linear é que os erros (ϵ_i) são independentes e identicamente distribuídos segundo uma distribuição normal com média zero e variância σ^2. Os resíduos, $r_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$, representam as estimativas dos erros ($r_i = \hat{\epsilon}_i$). Seria intuitivo assumir que os resíduos são variáveis aleatórias com média zero e variância σ^2. Esta afirmação é **falsa** e será posteriormente avaliada. Entretanto, é possível afirmar que os resíduos são variáveis aleatórias com distribuição normal pois representam combinações lineares de variáveis aleatórias normais (Y_i, $\hat{\beta}_0$ e $\hat{\beta}_1$).

Por uma questão de escala, seria desejável analisar os resíduos padronizados: $r_i^* = (r_i - E(r_i))/\sqrt{Var(r_i)}$. Neste caso, os resíduos padronizados seriam comparados com a distribuição normal padronizada ($r_i^* \sim Normal(0,1)$). Por exemplo, considerando um intervalo de confiança de 99,7% os limites superior e inferior para os resíduos padronizados são ± 3.

A figura 2.3 apresenta os resultados da análise visual dos resíduos para diferentes condições dos resíduos. As figuras 2.3 (a),(b) e (c) apresentam o comportamento ideal dos resíduos, caso todas as suposições iniciais do modelo estejam corretas. As figuras 2.3 (d),(e) e (f) apresentam o comportamento dos resíduos em uma condição de heterocedasticidade dos mesmos. Neste caso, a variância dos resíduos é proporcional à resposta média. As figuras 2.3 (g),(h) e (i) apresentam o comportamento dos resíduos em uma condição na qual os erros são homocedásticos mas a resposta média dos dados é não linear. Neste caso, o comportamento anormal dos resíduos reflete a má especificação da equação de regressão, ou seja, da equação do comportamento médio dos dados.



(a) Exemplo **A** (b) Análise de Resíduos (**A**) (c) Análise Normalidade (**A**)

(d) Exemplo **B** (e) Análise de Resíduos (**B**) (f) Análise Normalidade (**B**)

(g) Exemplo **C** (h) Análise de Resíduos (**C**) (i) Análise Normalidade (**C**)

Figura 2.3: Ajuste do modelo de regressão, análise de resíduos e análise de normalidade para três exemplos de regressão linear simples (**A**, **B** e **C**)

2.9 Ajuste do modelo de regressão linear simples utilizando o R

A sequência de comandos a seguir simula um conjunto de dados e realiza o ajuste do modelo de regressão linear simples.

```
n    ← 20
x    ← runif(n,-1,7)
erro ← rnorm(n, mean=0, sd=0.7)
y    ← 3 + 1.5*x + erro
dt   ← data.frame(y, x)
modelo  ← lm(y ~ x, data = dt)
summary(modelo)

plot(x, y, pch=19); grid()
seq.x ← seq(min(x), max(x), length.out=10)

saida ← predict(modelo, newdata=data.frame(x=seq.x),
```

```
        interval="confidence", level=0.95)
lines(seq.x, saida[,"fit"], col="red", lwd=1.5)
lines(seq.x, saida[,"lwr"], col="blue", lty=2)
lines(seq.x, saida[,"upr"], col="blue", lty=2)

saida ← predict(modelo, newdata=data.frame(x=seq.x),
         interval="prediction", level=0.95)
lines(seq.x, saida[,"lwr"], col="dark green")
lines(seq.x, saida[,"upr"], col="dark green")
```

O resultado do ajuste é mostrado na figura 2.4.

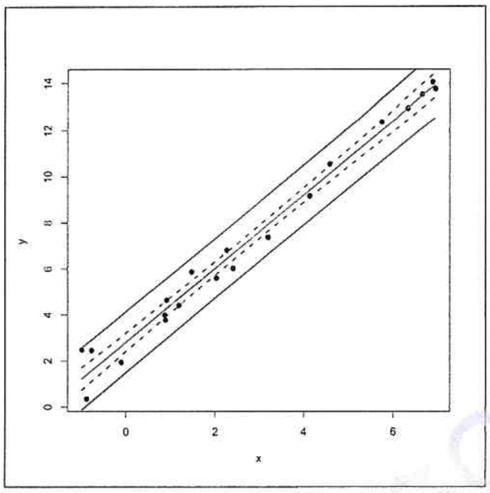


Figura 2.4: Modelo de regressão linear simples ajustado, intervalos de confiança e predição.

o sumário do modelo é apresentado a seguir:

```
> summary(modelo)
  Call:
  lm(formula = y ~ x, data = dt)

  Residuals:
       Min       1Q   Median       3Q      Max
  -1.22325 -0.40824  0.00209  0.36912  1.27780

  Coefficients:
              Estimate Std. Error t value Pr(>|t|)
  (Intercept)  3.07632    0.21001   14.65 1.92e-11 ***
  x            1.45883    0.06608   22.08 1.74e-14 ***
  Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05  `.' 0.1 ` ' 1

  Residual standard error: 0.6397 on 18 degrees of freedom
  Multiple R-squared:  0.9644,     Adjusted R-squared:  0.9624
  F-statistic: 487.4 on 1 and 18 DF,  p-value: 1.739e-14
```

2.10 Estudo de caso

Vamos considerar o exemplo da seção 1.8, procurando estimar o comportamento médio do salário em função da variável preditora: anos de experiência. A figura 2.5 mostra o resultado do ajuste do modelo de regressão linear simples, os intervalos de confiança e de predição.

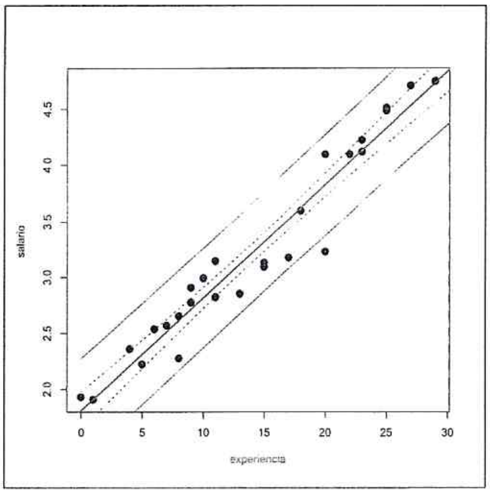


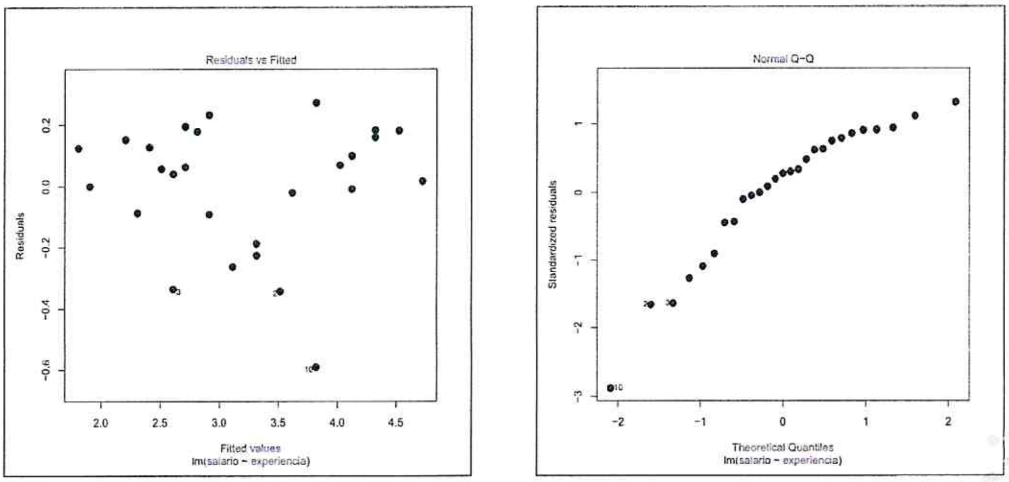
Figura 2.5: Resultado do ajuste do modelo de regressão linear simples, intervalos de confiança e predição.

O sumário do modelo é apresentado a seguir:

```
>  summary(modelo)
   Call:
   lm(formula = salario ~ experiencia, data = dados)

   Residuals:
        Min       1Q   Median       3Q      Max
   -0.59082 -0.08940  0.05755  0.15571  0.27078

   Coefficients:
               Estimate Std. Error t value Pr(>|t|)
   (Intercept) 1.806330   0.081610   22.13   <2e-16 ***
   experiencia 0.100759   0.005014   20.10   <2e-16 ***
   ---
   Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05  `.' 0.1 ` ' 1

   Residual standard error: 0.2113 on 25 degrees of freedom
   Multiple R-squared:  0.9417,    Adjusted R-squared:  0.9394
   F-statistic: 403.8 on 1 and 25 DF,  p-value: < 2.2e-16
```

Os resultados mostram que, segundo os resultados do modelo de regressão, a cada ano de experiência o salário médio é acrescido de 0,100759 mil reais. Este modelo apresenta um coeficiente de determinação (R^2) de 93,94%. Ou seja, o modelo de regressão é capaz de explicar 93,94% da dispersão dos dados em relação ao modelo de média (considerando um modelo contendo somente o intercepto).

Para validar as suposições de normalidade e homocedasticidade (ou variância constante) dos erros, é apresentado a análise dos resíduos na figura 2.6. A figura 2.6 (a) mostra que a observação localizada na linha 10 da base de dados apresenta um valor de resíduo muito baixo, quando comparado aos demais resíduos. Este fato é evidenciado na figura 2.6 (b) que mostra que a observação de número 10 possui um valor de resíduo padronizado próximo de -3.



(a) Resíduos versus valores ajustados ($\hat{y}_i$).

(b) Gráfico de probabilidade normal dos resíduos.

Figura 2.6: Análise dos resíduos do modelo.

Neste caso, é possível realizar um teste de hipótese para a normalidade dos resíduos. O resultado do teste de hipótese é apresentado a seguir:

```
> shapiro.test( residuals(modelo) )

        Shapiro-Wilk normality test

  data:  residuals(modelo)
  W = 0.9012, p-value = 0.01425
```

O resultado do teste de normalidade dos resíduos mostra que os resíduos não apresentam um comportamento de normalidade (valor-P = 0.01425, hipótese nula de normalidade foi rejeitada). Este resultado é potencialmente resultante da observação discrepante identificada na análise anterior.

2.11 O Estimador de Máxima Verossimilhança

Considere a amostra (y_i, x_i), i=1,2,...,n, e o modelo de regressão $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, onde $\varepsilon_i \sim NIID(0,\sigma^2)$ e $Y_i|x_i \sim NIID(\mu = \beta_0 + \beta_1 x_i, \sigma^2)$. A partir da distribuição de probabilidade da variável aleatória Y_i é possível escrever a função de Verossimilhança como:

$$L(\beta_0,\beta_1,\sigma^2) = P(\widetilde{Y}|\widetilde{X}) \tag{2.45}$$

$$= \prod_{i=1}^{n} f_{Y|X=x}(y_i|x_i) \tag{2.46}$$

$$= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[\frac{-1}{2\sigma^2}(y_i - \beta_0 - \beta_1 x_i)^2\right] \tag{2.47}$$

onde $\widetilde{Y}$ e $\widetilde{X}$ representam os vetores das observações, $\widetilde{Y} = \{y_1, \ldots, y_n\}$ e $\widetilde{X} = \{x_1, \ldots, x_n\}$. Pode-se então definir a função log-verossimilhança como:

$$l(\beta_0,\beta_1,\sigma^2) = -\frac{n}{2} ln 2\pi - \frac{n}{2} ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$$

Os estimadores de máxima verossimilhança são as soluções para o seguinte sistema de equações:

$$\frac{\partial l}{\partial \beta_0} = -\frac{1}{2\sigma^2} \cdot 2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)(-1) = 0 \tag{2.48}$$

$$\frac{\partial l}{\partial \beta_1} = -\frac{1}{2\sigma^2} \cdot 2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i) - x_i = 0 \tag{2.49}$$

Resolvendo o sistema de equações, temos os mesmos estimadores do método de Mínimos Quadrados, porém o estimador de máxima verossimilhança para σ^2 é a solução do seguinte sistema:

$$\frac{\partial l}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2 = 0$$

cuja solução é dada por:

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

2.12 Exercícios

Exercício 2.1

Obtenha a expressão para o estimador de mínimos quadrados para o seguinte modelo linear:
$Y_i|x_i = \beta_1 x_i + \epsilon_i$,
onde ϵ_i são variáveis aleatórias independentes e identicamente distribuídas:
$\epsilon_i \sim Normal(0, \sigma^2)$

(a) Para o modelo acima, encontre: $E(\hat{\beta}_1)$ e $Var(\hat{\beta}_1)$

Exercício 2.2

Considere o seguinte modelo de regressão linear simples:
$Y_i|x_i = \beta_0 + \beta_1 \cdot (x_i - \bar{x})$, onde $\bar{x}$ é a média amostral.

(a) Mostre que: $\hat{\beta}_0 = \bar{y}$, onde $\bar{y}$ é a média amostral observada da variável resposta.

(b) Obtenha a expressão do estimador $\hat{\beta}_1$.