1 Princípios elementares para a análise de dados utilizando os modelos de regressão lineares

Antes de adentrar o universo dos modelos lineares, este capítulo tem como objetivo revisar importantes tópicos estatísticos: propriedades do estimador média amostral, o estimador de mínimos quadrados, o estimador de máxima verossimilhança, dentre outros. O objetivo é apresentar conceitos e propriedades em um contexto mais simples para, nos capitulos seguintes, estender esses conceitos aos modelos lineares e aos modelos lineares generalizados. Neste capítulo, assume-se que existe apenas uma variável aleatória de interesse Y para a qual uma amostra de tamanho n está disponível. Distingue-se uma variável aleatória de uma instância, ou um valor observado, pela nomenclatura: valores maiúsculos são utilizados para variáveis aleatória de valores minúsculos para suas instâncias.

Seja Y_i variáveis aleatória independentes com distribuição normal, com média μ e variância σ^2, $Y_i \sim Normal\left(\mu, \sigma^2\right)$. Vamos definir y_i como uma instância de Y_i. Seja também uma amostra de tamanho n: $\{y_1, y_2, \ldots, y_n\}$.

1.1 O estimador média amostral

Define-se a média amostral como: $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$. Pode-se então caracterizar a variável aleatória $\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$. A distribuição de $\overline{Y}$ é normal com média μ e variância $\frac{\sigma^2}{n}$, $\overline{Y} \sim Normal\left(\mu, \frac{\sigma^2}{n}\right)$. Este fato pode ser verificado utilizando princípios básicos de probabilidade. Utilizando o operador esperança matemática, $E(\cdot)$, têm-se que:

$$
\begin{aligned}
E(\overline{Y}) &= E\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) = \frac{1}{n}E\left(\sum_{i=1}^{n} Y_i\right) \\
&= \frac{1}{n}\sum_{i=1}^{n} E(Y_i) = \frac{1}{n}\sum_{i=1}^{n} \mu \\
&= \frac{1}{n} n\mu \\
&= \mu
\end{aligned}
\tag{1.1}
$$

Aplicando o operador variância, $Var(\cdot)$, novamente assumindo que as variáveis aleatórias Y_i são independentes e identicamente distribuídas, então:

$$
\begin{aligned}
Var(\overline{Y}) =& Var\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) = \frac{1}{n^2} Var\left(\sum_{i=1}^{n} Y_i\right) \\
=& \frac{1}{n^2}\sum_{i=1}^{n} Var(Y_i) = \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 \\
=& \frac{1}{n^2} n\sigma^2 \\
=& \frac{\sigma^2}{n}
\end{aligned} \tag{1.2}
$$

Ambas as equações 1.1 e 1.2 podem ser interpretadas da seguinte forma: caso seja coletada uma amostra de tamanho n de uma variável aleatória normal com média μ e variância σ^2, mesmo desconhecendo os parâmetros μ e σ^2, é possível afirmar que a média amostral oscila em torno do parâmetro da média. Além disso, à medida que o tamanho da amostra aumenta, menor seja a dispersão da média amostral com relação ao parâmetro μ.

1.2 O ESTIMADOR DE MÍNIMOS QUADRADOS

Vamos considerar a estrutura do modelo de regressão mais simples possível: $Y_i = \beta_0 + \epsilon$, onde ϵ é uma variável aleatória normal com média 0 (zero) e variância σ^2, $\epsilon \sim Normal\,(0, \sigma^2)$, e β_0 é uma constante. Este modelo é uma representação alternativa para o caso onde $Y_i \sim Normal(\mu, \sigma^2)$. É possível mostrar que a variável aleatória Y_i pode ser escrita a partir de um modelo aditivo: $Y_i = \mu + \epsilon$, onde μ é uma constante que define a média de Y_i e ϵ é uma variável aleatória com média zero e variância σ^2 que define a variância de Y_i. No caso inicialmente apresentado, o parâmetro de média μ e o parâmetro β_0 são equivalentes.

Neste caso, queremos encontrar a solução para o seguinte problema de otimização:

$$
\hat{\beta}_0 = arg\,\min_{\beta_0}\sum_{i=1}^{n}(y_i - \beta_0)^2 \tag{1.3}
$$

A solução para a equação (1.3) é:

$$
\frac{\partial \sum_{i=1}^{n}\left(y_i - \hat{\beta}_0\right)^2}{\partial \beta_0} = 0
$$

$$
\begin{aligned}
\sum_{i=1}^{n} 2\left(y_i - \hat{\beta}_0\right)\cdot(-1) =& 0 \\
-2\sum_{i=1}^{n} y_i + 2\sum_{i=1}^{n}\hat{\beta}_0 =& 0 \\
2n\hat{\beta}_0 =& 2\sum_{i=1}^{n} y_i \\
\hat{\beta}_0 =& \frac{1}{n}\sum_{i=1}^{n} y_i
\end{aligned} \tag{1.4}
$$

Note que a solução de mínimos quadrados para o parâmetro β_0 também é a média amostral, $\hat{\beta}_0 = \overline{y}$. Neste caso, uma vez que ϵ é uma variável aleatória, então Y_i também é uma variável aleatória normal com média e variância definidas por:

$$
\begin{aligned}
E(Y_i) &= E(\beta_0 + \epsilon) \\
&= \beta_0 + E(\epsilon) \\
&= \beta_0
\end{aligned}
\tag{1.5}
$$

$$
\begin{aligned}
Var(Y_i) &= Var(\beta_0 + \epsilon) \\
&= Var(\beta_0) + Var(\epsilon) \\
&= 0 + Var(\epsilon) \\
&= \sigma^2
\end{aligned}
\tag{1.6}
$$

então: $Y_i \sim Normal(\mu = \beta_0, \sigma^2)$. Neste caso, como o parâmetro de média é desconhecido, utiliza-se o estimador de mínimos quadrados ou a média amostral.

1.3 O estimador de máxima verossimilhança

O estimador de máxima verossimilhança de um parâmetro populacional requer a especificação da distribuição de probabilidade da variável de interesse. Seja então uma amostra de tamanho n de variáveis aleatórias independentes e identicamente distribuídas, $\{Y_1, Y_2, \ldots, Y_n\}$, $Y_i \sim Normal(\beta_0, \sigma^2)$. A função de verossimilhança é definida pela probabilidade de observar a amostra $\{Y_1 = y_1, Y_2 = y_2, \ldots, Y_n = y_n\}$ dado o modelo de distribuição de probabilidade especificado:

$$
\begin{aligned}
F(\beta_0, \sigma) &= P(Y_1 = y_1, Y_2 = y_2, \ldots, Y_n = y_n), \text{ assumindo independência:} \\
&= P(Y_1 = y_1) \times P(Y_2 = y_2) \times \ldots \times P(Y_n = y_n) \\
&= \prod_{i=1}^{n} P(Y_i = y_i), \text{ sendo identicamente distribuídos:} \\
&= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{y_i - \beta_0}{\sigma}\right)^2} \\
&= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n e^{-\frac{1}{2}\frac{\sum_{i=1}^{n}(y_i - \beta_0)^2}{\sigma^2}}
\end{aligned}
\tag{1.7}
$$

O estimador de máxima verossimilhança é a solução que maximiza a função de verossimilhança, ou seja, os valores de β_0 e σ^2 que maximizam a probabilidade de ser observada a amostra $\{Y_1 = y_1, Y_2 = y_2, \ldots, Y_n = y_n\}$. A maximização pode ser realizada, sem alteração da solução, do logarítmo da função de verossimilhança:

$$
log\, F(\beta_0, \sigma) = -n \log\left(\sqrt{2\pi\sigma^2}\right) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \beta_0)^2
$$

A solução de máxima verossimilhança para o parâmetro β_0 é:

$$
\begin{aligned}
\frac{\partial log\, F(\beta_0,\sigma)}{\partial \beta_0} &= 0 \\
\frac{\partial}{\partial \beta_0}\left[-n\log\left(\sqrt{2\pi\sigma^2}\right)-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i-\hat{\beta}_0\right)^2\right] &= 0 \\
\frac{\partial}{\partial \beta_0}\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i-\hat{\beta}_0\right)^2\right] &= 0 \\
\frac{\partial}{\partial \beta_0}\left[-\sum_{i=1}^{n}\left(y_i-\hat{\beta}_0\right)^2\right] &= 0 \\
-2\sum_{i=1}^{n}\left(y_i-\hat{\beta}_0\right)(-1) &= 0 \\
+2\sum_{i=1}^{n}y_i-2\sum_{i=1}^{n}\hat{\beta}_0 &= 0 \\
-2n\hat{\beta}_0 &= -2\sum_{i=1}^{n}y_i \\
\hat{\beta}_0 &= \frac{1}{n}\sum_{i=1}^{n}y_i
\end{aligned}
\tag{1.8}
$$

Ou seja, a solução de mínimos quadrados e o estimador de máxima verossimilhança são equivalentes e iguais à média amostral, supondo o modelo de distribuição normal para a variável aleatória de interesse. Modelos mais complexos serão objetos futuros de averiguação e exercícios.

1.4 Inferência estatística com relação ao parâmetro μ

Inicialmente, vamos considerar a distribuição da média amostral para o caso de variáveis aleatórias com distribuição normal, $Y_i \sim Normal(\mu, \sigma^2)$ e $\overline{Y} \sim Normal\left(\mu, \frac{\sigma^2}{n}\right)$. Se os parâmetros μ e σ^2 fossem conhecidos, então poderíamos inferir sobre possíveis valores da média amostral a partir da construção de faixas de referência:

$$FR(\overline{Y}, \alpha) : \mu \pm Z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}} \tag{1.9}$$

Por exemplo, assumindo um nível de confiança: $1-\alpha = 0.997$ (99.7%) então $Z_{\alpha/2} = 3$. Uma vez que o parâmetro populacional μ é desconhecido, podemos construir um intervalo centrado no valor observado da média amostral, $\overline{Y}$, e então inferir sobre os valores mais prováveis para o parâmetro populacional μ. Neste caso, o intervalo é conhecido como intervalo de confiança (IC) para a média populacional:

$$IC(\mu, \alpha) : \overline{Y} \pm Z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}} \tag{1.10}$$

É importante destacar que, em ambos os casos mostrados nas equações 1.9 e 1.10, assume-se que o parâmetro de variância σ^2 é conhecido. Tal suposição é irreal na grande maioria dos casos. A figura 1.1 apresenta uma ilustração das diferenças entre faixa de referência e intervalo de confiança.

Nos casos onde a variância populacional é desconhecida, faz-se a substituição pela estimativa do desvio padrão amostral, $S = \sqrt{\frac{1}{n-1}\sum_i (x_i - \bar{x})^2}$. Neste caso, a incerteza associada ao estimador do desvio padrão é considerada nos intervalos da confiança e na faixa de referência, a partir da subsituição do escore $Z_{\alpha/2}$ pelo escore $t_{\left(\frac{\alpha}{2}, n-1\right)}$:

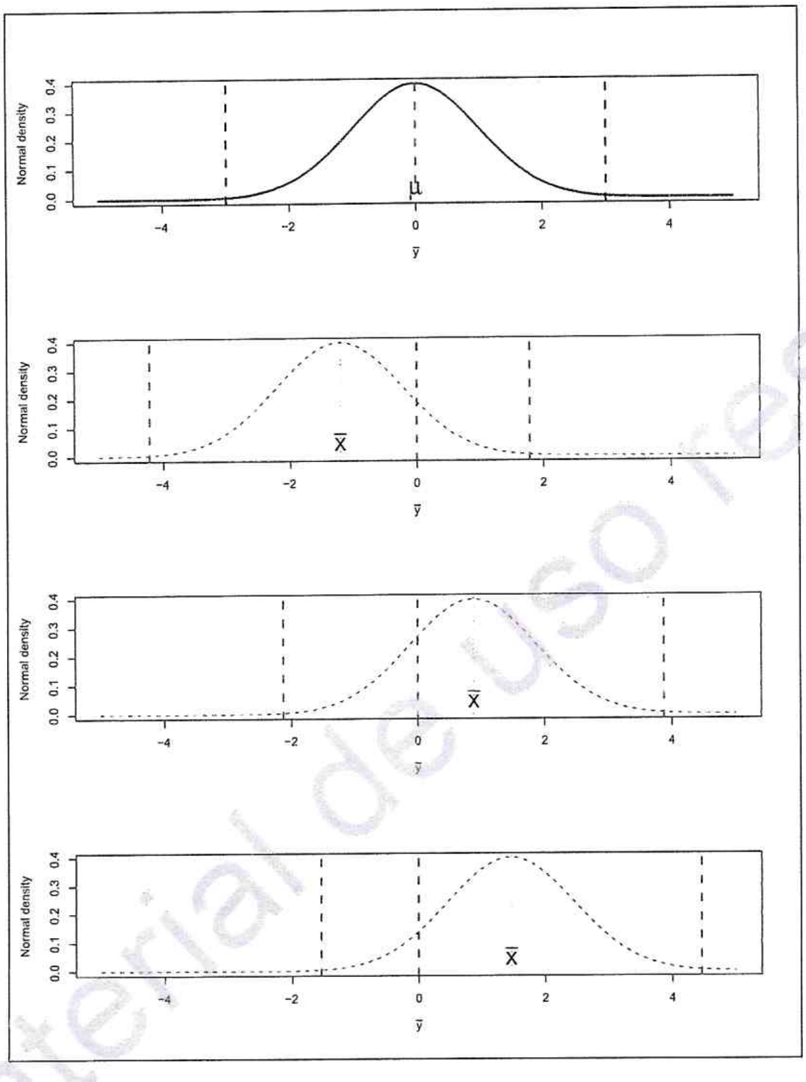


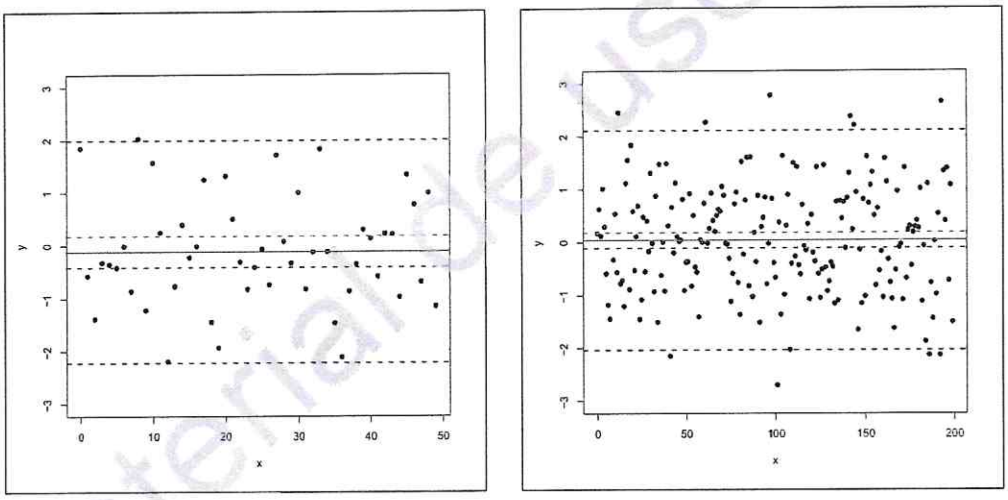
Figura 1.1: Diferenças entre faixa de referência para a média amostral e intervalo de confiança para o parâmetro média populacional. A faixa de referência está centrada no valor de μ e indica a faixa de valores mais prováveis para a média amostral. O intervalo de confiança está centrado na média amostral e indica o intervalo mais provável de conter o verdadeiro valor da média populacional, μ.

$$FR(\overline{Y},\alpha) : \mu \pm t_{\left(\frac{\alpha}{2},n-1\right)} \times \frac{S}{\sqrt{n}} \tag{1.11}$$

$$IC(\mu,\alpha) : \overline{Y} \pm t_{\left(\frac{\alpha}{2},n-1\right)} \times \frac{S}{\sqrt{n}} \tag{1.12}$$

1.5 Intervalo de confiança versus intervalo de predição

Supondo que $Y \sim N\left(\mu,\sigma^2\right)$ e que $\overline{Y} \sim N\left(\mu,\frac{\sigma^2}{n}\right)$, o intervalo de confiança para o parâmetro μ procura caracterizar o intervalo mais provável para a ocorrência do parâmetro da média, μ : $\overline{Y} \pm t_{\left(\frac{\alpha}{2},n-1\right)} \times \frac{S}{\sqrt{n}}$. Ou seja, a largura do intervalo de confiança é afetado pelo tamanho da amostra n. Já o intervalo de predição procura caracterizar o intervalo mais provável para a ocorrência de valores futuros, ou seja, novos valores para Y_i, $i > n$. Uma vez definidos os estimadores da média, $\hat{\mu}$, e desvio padrão amostral, S, o intervalo pode ser construído assumindo $Y_i \sim N\left(\hat{\mu},S^2\right)$. Em geral, o intervalo de predição requer ajustes adicionais em função das incertezas associadas aos estimadores da média e desvio padrão. Mas, o intervalo de predição não é tão afetado pelo tamanho da amostra n do que o intervalo de confiança. A figura 1.2 ilustra as diferenças entre o intervalo de confiança e o intervalo de predição. Sucintamente, o intervalo de confiança diminiu com o aumento do tamanho da amostra. O intervalo de predição se torna estável com o aumento do tamanho da amostra, mas não diminui.



(a) Intervalo de predição e intervalo de confiança para amostra de tamanho $n = 50$.

(b) Intervalo de predição e intervalo de confiança para amostra de tamanho $n = 200$.

Figura 1.2: Estimativas de intervalo de confiança para a média populacional (linhas horizontais em vermelho) e intervalo de predição (linhas horizontais em preto) para novas observações considerando tamanhos de amostra $n = 50$ e $n = 100$.

1.6 Teste de hipóteses para a média populacional

Uma vez que seja especificada uma condição de igualdade para o parâmetro de média, $\mu = \mu_0$ é possível adaptar a faixa de referência para avaliar a hipótese nula $H_0 : \ \mu = \mu_0$. Ou seja, neste contexto, uma hipótese nula consiste em definir uma condição de igualdade para um parâmetro de interesse. Caso a hipótese nula seja verdadeira a média amostral estará contida no intervalo $\mu_0 \pm t_{(\frac{\alpha}{2},n-1)} \times \frac{S}{\sqrt{n}}$ com probabilidade $1-\alpha$. Caso a média amostral não esteja no intervalo, têm-se evidência a favor da hipótese alternativa, que é o complemento da hipótese nula, $H_1 : \mu \neq \mu_0$

1.7 Vício e variância do estimador desvio padrão amostral

Até o momento, foram avaliadas as propriedades do estimados média amostral $(\overline{Y})$. Neste ponto, queremos avaliar propriedades do estimador variância amostral. Considere Y uma variável aleatória com distribuição normal, $Y \sim Normal(\mu, \sigma^2)$, queremos avaliar a esperança do seguinte estimador de σ^2:

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

$$\begin{aligned}
E\left[\frac{1}{n}\sum_{i=1}^{n}(Y_i-\overline{Y})^2\right] &= \frac{1}{n}E\left[\sum_{i=1}^{n}(Y_i-\overline{Y})^2\right] \\
&= \frac{1}{n}E\left[\sum_{i=1}^{n}(Y_i-\mu+\mu-\overline{Y})^2\right] \\
&= \frac{1}{n}E\left\{\sum_{i=1}^{n}\left[(Y_i-\mu)-(\overline{Y}-\mu)\right]^2\right\} \\
&= \frac{1}{n}E\left[\sum_{i=1}^{n}(Y_i-\mu)^2 - 2\sum_{i=1}^{n}(Y_i-\mu)(\overline{Y}-\mu) + \sum_{i=1}^{n}(\overline{Y}-\mu)^2\right] \\
&= \frac{1}{n}E\left[\sum_{i=1}^{n}(Y_i-\mu)^2\right] - 2\frac{1}{n}E\left[\underbrace{\sum_{i=1}^{n}(Y_i-\mu)(\overline{Y}-\mu)}_{a}\right] + \frac{1}{n}E\left[n(\overline{Y}-\mu)^2\right]
\end{aligned}$$

$$\begin{aligned}
a: \sum_{i=1}^{n}(Y_i-\mu)(\overline{Y}-\mu) &= \sum_{i=1}^{n}\left(Y_i\overline{Y} - \mu Y_i - \mu\overline{Y} + \mu^2\right) \\
&= \overline{Y}\sum_{i=1}^{n}(Y_i) - \mu\sum_{i=1}^{n}(Y_i) - n\mu\overline{Y} + n\mu^2 \\
&= n\overline{Y}^2 - n\mu\overline{Y} - n\mu\overline{Y} + n\mu^2 \\
&= n(\overline{Y}-\mu)^2
\end{aligned}$$

logo, considerando a seguinte propriedade: $Var(Y) = E\left[(Y - E[Y])^2\right]$, podemos concluir que:

$$\begin{aligned}
E\left[\frac{1}{n}\sum_{i=1}^{n}(Y_i-\overline{Y})^2\right] &= \frac{1}{n}E\left[\sum_{i=1}^{n}(Y_i-\mu)^2\right] - 2\frac{1}{n}E\left[n(\overline{Y}-\mu)^2\right] + \frac{1}{n}E\left[n(\overline{Y}-\mu)^2\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}E(Y_i-\mu)^2 - \frac{1}{n}E\left[n(\overline{Y}-\mu)^2\right] \\
&= \frac{1}{n}n\sigma^2 - \underbrace{E\left[(\overline{Y}-\mu)^2\right]}_{Var(\overline{Y})} \\
&= \frac{1}{n}n\sigma^2 - \frac{1}{n}\sigma^2 \\
&= \frac{n-1}{n}\sigma^2
\end{aligned}$$

Ou seja, $\frac{1}{n}\sum_{i=1}^{n}\left(Y_i-\overline{Y}\right)^2$ é um estimador viciado para σ^2. Se $E\left[\frac{1}{n}\sum_{i=1}^{n}\left(Y_i-\overline{Y}\right)^2\right]=\frac{n-1}{n}\sigma^2$, podemos encontrar um estimador não viesado multiplicando o estimador viciado por $\frac{n}{n-1}$:

$$\begin{aligned}\frac{n}{n-1}E\left[\frac{1}{n}\sum_{i=1}^{n}\left(Y_i-\overline{Y}\right)^2\right]&=\frac{n}{n-1}\times\frac{n-1}{n}\sigma^2\\&=\sigma^2\end{aligned}$$

ou seja:

$$E\left[\frac{1}{n-1}\sum_{i=1}^{n}\left(Y_i-\overline{Y}\right)^2\right]=\sigma^2 \tag{1.13}$$

Diferente do estimador média amostral, o estimador não viciado para a variância utiliza $\frac{1}{n-1}$ no denominador ao contrário do tradicional $\frac{1}{n}$. A justificativa para isso, como demonstrado, é o fato de que $\frac{1}{n}$ gera um estimador viciado, ao fato que $\frac{1}{n-1}$ gera um estimador não viciado.

1.8 O Teorema Central do Limite / Teorema do Limite Central

Seja $Y_1, Y_2, \ldots, Y_n$ variáveis independentes com média μ_i e variância σ_i^2. Se $T = Y_1 + Y_2 + \ldots + Y_n$, então a distribuição de

$$\frac{T-\sum_{i=1}^{n}\mu_i}{\sqrt{\sum_{i=1}^{n}\sigma_i^2}}$$

se aproxima da distribuição normal padrão $N(\mu=0,\sigma^2=1)$ à medida que n tende a infinito.

Uma descrição alternativa para o Teorema Central do Limite dada uma amostra independente e identicamente distribuída $Y_1, Y_2, \ldots, Y_n$, onde $E(Y_i)=\mu$ e $Var(Y_I)=\sigma^2$ é:

$$\overline{Y} \underset{n\to\infty}{\sim} N\left(\mu, \frac{\sigma^2}{n}\right) \tag{1.14}$$

Embora existam vários elementos técnicos com relação à equação 1.8, é importante ressaltar a importância de sua interpretação. A equação não requer a definição de uma distribuição de probabilidade para Y_i, mas apenas as propriedades de independência e *identicamente distribuídas*. Ou seja, para uma variável aleatória com distribuição desconhecida, a média amostral para uma amostra independente e identicamente distribuída, suficientemente grande, converge para uma distribuição normal com média $E(Y_i)$ e variância $\frac{Var(Y_i)}{n}$. Esta é uma importante propriedade aplicada aos modelos de regressão quando a variável resposta não possui uma distribuição normal.

Exemplo 1-1

Um investigador deseja estudar a possível relação entre o salário (em mil reais) e o tempo de experiência (em anos) no cargo de gerente de agências bancárias de uma grande empresa. Os dados são mostrados na figura 1.3.

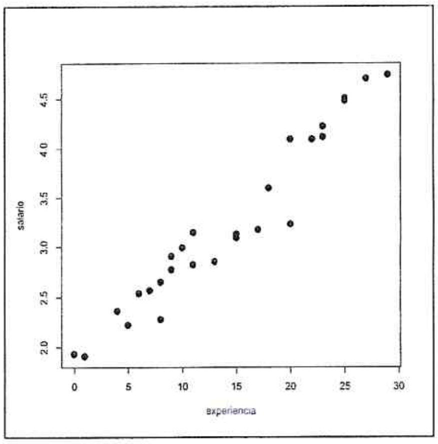
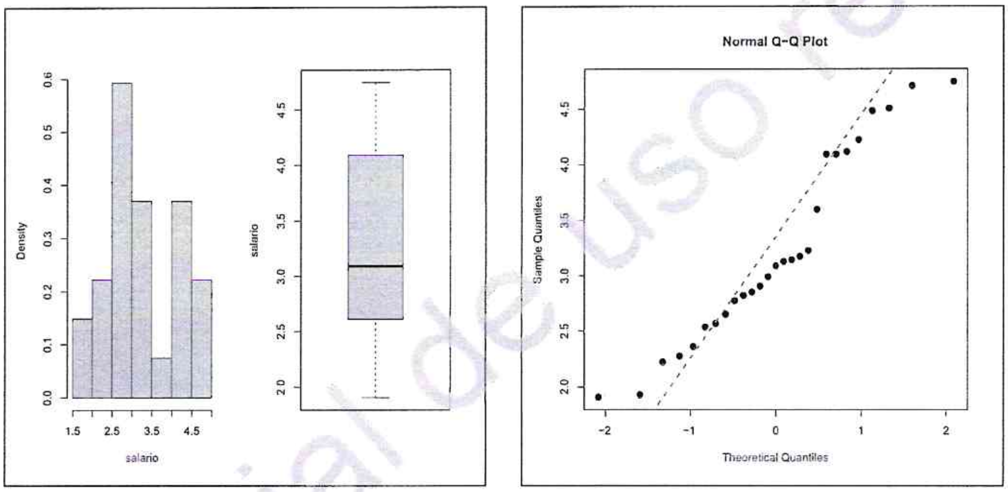


Figura 1.3: Exemplo: salário em função do tempo de experiência

Vamos considerar, inicialmente, que a informação sobre o tempo de experiência é inexistente. A variável aleatória de interesse é o salário (Y). A figura 1.4 mostra o histograma e o boxplot (fig. 1.4 (a)) e o gráfico de normalidade (qqplot) (fig. 1.4 (b)) de y.



(a) Histograma e Boxplot de Y

(b) Gráfico de probabilidade normal de Y

Figura 1.4: Análise de resíduos para verificar a não correlação entre resíduo e média estimada, homocedasticidade, e seu comportamento segundo uma distribuição normal.

Neste caso, é possível assumir que a variável aleatória Y segue uma distribuição de probabilidade com média μ e variância σ^2. A suposição de normalidade da variável aleatória Y pode ser investigada a partir de um gráfico de probabilidade normal, mostrado na figura 1.4 (b). Além da análise visual da característica de normalidade dos valores, é possível aplicar um teste de normalidade, mostrado a seguir.

```
> shapiro.test(salario)
    Shapiro-Wilk normality test
    data:  salario
    W = 0.9344, p-value = 0.0884
```

O resultado do teste de normalidade indica a não-rejeição da hipótese nula (valor-p = 0.0884) de que os dados apresentam uma distribuição normal. Portanto, há evidência de que o comportamento da variável aleatória Y pode ser definido a partir de uma distribuição normal, $Y \sim Normal\left(\mu, \sigma^2\right)$. Os estimadores não viciados são: $\hat{\mu} = 3.228$ e $S^2 = 0.7367$.

O modelo equivalente de regressão para este caso é definido como $Y = \beta_0 + \epsilon$, onde β_0 é um parâmetro (determinístico) desconhecido e $\epsilon \sim Normal(0, \sigma^2)$. Este modelo apresenta as seguintes propriedades: $E(Y) = \beta_0$ e $Var(Y) = \sigma^2$. Ou seja, neste modelo o parâmetro β_0 representa a média da variável aleatória Y e a variável aleatória ϵ caracteriza somente a componente de dispersão (σ^2). Este é o modelo básico de uma análise de regressão. *Na ausência de potenciais variáveis preditoras, o modelo de regressão para o comportamento de uma variável aleatória Y com distribuição normal pode ser definido pela soma de uma constante e uma variável aleatória normal de média zero e variância σ^2.*

■

1.9 Exercícios

Exercício 1.1

Considere uma amostra independente e identicamente distribuída de n variáveis normais com média μ e variância σ^2. Obtenha a expressão do estimador de máxima verossimilhança para o parâmetro de variância, $\hat{\sigma}^2$.

Exercício 1.2

Considere uma amostra independente e identicamente distribuída de n variáveis aleatórias de bernoulli com parâmetro p, $Y_i \sim Bernoulli(p)$. Obtenha a expressão do estimador de máxima verossimilhança para o parâmetro p.

Exercício 1.3

Considere uma amostra independente e identicamente distribuída de n variáveis aleatórias de Poisson com parâmetro μ, $Y_i \sim Poisson(\mu)$. Obtenha a expressão do estimador de máxima verossimilhança para o parâmetro μ.