Universidade Federal de Minas Gerais
Instituto de Ciências Exatas
Departamento de Estatística

Análise de Regressão no *software* R: propriedades dos estimadores via método de Monte Carlo, aplicações e exercícios

Guilherme Lopes de Oliveira
Rosangela Helena Loschi
Magda Carvalho Pires

RELATÓRIO TÉCNICO
SÉRIE ENSINO
RTE 01/2018

Análise de Regressão no *software* R

Universidade Federal de Minas Gerais
Instituto de Ciências Exatas
Departamento de Estatística

Este material é resultado do projeto PIFD2017-64 "Análise de Regressão: Aplicações utilizando o *software* R"do Programa de Incentivo à Formação Docente (PIFD) da Pró-reitoria de Graduação da UFMG (Edital 01/2017).

Guilherme Lopes de Oliveira é atualmente aluno de doutorado em Estatística junto ao DEST-UFMG e bolsista do projeto.

Rosangela Helena Loschi é professora do DEST-UFMG e responsável pela disciplina *EST035 - Análise de Regressão* durante a vigência do projeto PIFD2017-64.

Magda Carvalho Pires é professora do DEST-UFMG, coordenadora do curso de Graduação em Estatística da UFMG e coordenadora do projeto PIFD2017-64.

Belo Horizonte, MG - Brasil
Agosto 2018

Sumário

1 Introdução

Análise de regressão linear é uma técnica estatística utilizada para investigar a relação existente entre variáveis e está entre as ferramentas estatísticas mais utilizadas na prática. De forma sucinta, a meta é modelar a relação entre uma variável dependente contínua (variável resposta) e uma ou mais variáveis explicativas (covariáveis). A especificação de uma distribuição de probabilidade para a variável resposta (por meio do termo de erro) permite que se faça inferência para os parâmetros do modelo além de predições. Portanto, essa técnica pode ser utilizada com vários objetivos, dentre os quais se pode destacar: descrever a relação entre variáveis para entender um processo ou fenômeno; prever o valor de uma variável a partir do conhecimento dos valores das outras variáveis; substituir a medição de uma variável pela observação dos valores de outras variáveis; controlar os valores de uma variável em uma faixa de interesse.

Pela sua importância e grande utilização em diversas áreas, as técnicas de análise de regressão linear são usualmente abordadas em disciplinas de graduação e pós graduação. Na UFMG, por exemplo, os Cursos de Graduação em Estatística e em Ciências Atuariais oferece em sua grade curricular a disciplina obrigatória "EST035 - Análise de Regressão", que é optativa para o Curso de Ciências Econômicas e alguns cursos de Engenharia. Com caráter teórico e prático, a disciplina idealmente deve ser ministrada em sala de aula e em laboratório de informática. Nas aulas práticas os alunos aprendem a analisar bancos de dados e ajustar os modelos de regressão vistos em sala.

Nesse contexto, o *software* estatístico R é uma ferramenta bastante interessante por ser de livre acesso e ter suas funcionalidades constantemente atualizadas através da implementação de novos pacotes. Reis *et al.* (2009) descrevem como utilizar *software* R no ajuste dos modelos de regressão, enquanto noções básicas de utilização do *software* podem ser encontradas em Landeiro (2013) e Ribeiro *et al.* (2012).

Assim, na Seção 2 demonstramos como empregar o método de Monte Carlo para o estudo das propriedades dos estimadores dos parâmetros envolvidos nos modelos de regressão linear. Alguns exercícios práticos e teóricos são propostos na Seção 3.

Esperamos que esse material possa motivar professores e alunos dos cursos de Análise de Regressão, além dos estusiastas do assunto.

2 Métodos de Monte Carlo (MMC)

Nesta seção abordamos o estudo das propriedades dos estimadores dos parâmetros envolvidos nos modelos de regressão linear. O objetivo é apresentar ao aluno ferramentas que permitem a exploração e visualização de tais propriedades num contexto numérico utilizando dados simulados.

Em sala de aula, a derivação dos estimadores para os parâmetros envolvidos num modelo de regressão linear é feito via método dos mínimos quadrados (MQ) e prova-se, matematicamente, propriedades como, por exemplo, a ausência de viés dos estimadores de MQ para o coeficientes do modelo $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)$ (i.e., a média do estimador coincide com o valor verdadeiro do respectivo parâmetro), sendo p o número de covariáveis no modelo. Além disso, as expressões para as variâncias dos estimadores são obtidas e discute-se o efeito do tamanho da amostra n sobre tais variâncias.

O método de Monte Carlo (MMC) é uma ferramenta que nos possibilita avaliar tais propriedades através de replicações de um contexto real previamente especificado. A ideia, conceitos e exploração do MMC são descritos nas subseções seguintes.

2.1 O que é MMC, como surgiu e como utilizar?

Uma breve revisão histórica sobre o surgimento do que conhecemos como *método de Monte Carlo* (MMC) é descrito em material do Sistema Maxwell da PUC-Rio (disponível em https://www.maxwell.vrac.puc-rio.br/19632/19632_4.PDF):

Em 1946 o matemático Stanislaw Ulam durante um jogo de paciência tentou calcular as probabilidades de sucesso de uma determinada jogada utilizando a tradicional análise combinatória. Após gastar bastante tempo fazendo cálculos percebeu que uma alternativa mais prática seria simplesmente realizar inúmeras jogadas, por exemplo, cem ou mil, e contar quantas vezes cada resultado ocorria.

Ulam sabia que técnicas de amostragem estatística, como esta, não eram muito usadas por envolverem cálculos extremamente demorados, tediosos e sujeitos a erros. Entretanto, nessa época, ficara pronto o primeiro computador eletrônico, desenvolvido durante a segunda guerra mundial, o ENIAC; antes dele eram usados dispositivos mecânicos para fazer cálculos. A versatilidade e rapidez do ENIAC, sem precedentes para a época, impressionaram Ulam, que sugeriu o uso de métodos de amostragem estatística para solucionar o problema da difusão de nêutrons em material sujeito a fissão nuclear, difundindo assim sua aplicação.

Posteriormente, esse método ficou conhecido como Método de Monte Carlo, nome inspirado em um tio de Ulam, que jogava constantemente no famoso cassino de Monte Carlo, cujo aspecto aleatório de suas roletas também está intimamente ligado ao método. O Método de Monte Carlo foi formalizado em 1949, por meio do artigo intitulado "Monte Carlo Method", publicado por John Von Neumann e Stanislav Ulam.

Então, basicamente, o método de Monte Carlo (MMC) é uma metodologia de si-

mulação estatística que se baseia em uma grande quantidade de amostragens aleatórias para se chegar em resultados próximos do que seriam os resultados reais de um determinado fenômeno. Ele permite, portanto, que se façam testes com variáveis aleatórias um número suficientemente grande de vezes para obter com mais precisão a chance de algum resultado específico acontecer.

MMC é utilizado rotineiramente em muitos campos de conhecimentos que vão desde simulação de complexos fenômenos físicos a econômicos. Alguns exemplos de aplicação deste método, em diferentes áreas, são:

- Atuária: tábua de expectativa de vida, casamento de passivos/ativos, etc.;
- Estatística: simulação de modelos teóricos e suas propriedades, etc.;
- Finanças: análise de ações, opções futuras, séries macroeconômicas, etc.;
- Gestão: análise de riscos, projeções, etc.;
- Computação gráfica: redução de artefatos, espalhamento, etc.;
- Geologia: caracterização de reservatórios;
- Análise de Projetos: opções reais;
- Jogos: geração de redes (grafos).

A utilização do MMC exige que o sistema físico ou o modelo matemático seja descrito em termos de funções de densidade de probabilidade (FDP). Uma vez conhecidas essas distribuições, as simulações podem proceder fazendo as amostragens aleatórias a partir das mesmas. Este processo é repetido inúmeras vezes, digamos M replicações, e o resultado desejado é então obtido por meio de estatísticas (média, desvio padrão, etc.) sobre um determinado número de realizações (amostras), digamos n.

Na prática, diante de um problema envolvendo incertezas/quantidades aleatórias, a utilização do MMC consiste dos seguintes passos:

Passo 1: Expressar o comportamento do fenômeno de interesse fazendo uso de alguma estrutura/modelo que envolva FDPs para representar o comportamento de cada uma das incertezas/variáveis aleatórias.

Passo 2: Gerar n valores pseudo-aleatórios aderentes à FDP de cada incerteza do modelo.

Passo 3: Calcular o resultado determinístico substituindo as incertezas pelos n valores gerados obtendo, assim, uma amostra/replicação do modelo.

Passo 4: Repetir os Passos 2 e 3 até se obter um total de M replicações de amostras de tamanho n (especificado no passo 3) do modelo.

Passo 5: Agregar e manipular os resultados da amostra de forma a obter uma estimativa da solução do problema, por exemplo, a média de alguma variável envolvida.

Note que este método apenas proporciona uma aproximação da solução, já que, está associado a simulação sucessiva de quantidades aleatórias. Em muitos casos o erro de aproximação, também chamado de *erro de Monte Carlo*, pode ser calculado de forma explícita. É evidente que quanto maior o tamanho da amostra n e também o total de replicações M, menor o erro de aproximação. Por sua vez, o esforço computacional envolvido está diretamente relacionado aos valores de n e M. Portanto, quanto menor o erro de aproximação desejado, maior o esforço computacional envolvido. Na prática, para definir o número de simulações deve-se fazer um balanço entre a qualidade desejada para os resultados em termos de um erro máximo pré-definido e as disponibilidades de *hardware* e de tempo.

2.2 MMC no contexto de Análise de Regressão

Nosso objetivo é utilizar o método de Monte Carlo (MMC) para verificar propriedades dos estimadores de mínimos quadrados (EMQ) para os parâmetros do modelo de regressão linear e avaliar sua precisão em termos de variabilidade. Para exemplicar como isto pode ser feito, consideremos o modelo de regressão linear simples

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \tag{1}$$

onde $\epsilon_i \overset{iid}{\sim} N(0, \sigma_\epsilon^2)$ e $i = 1, \ldots, n$. Lembre-se que, assim como os parâmetros β_0 e β_1, a variância σ_ϵ^2 comum dos termos do erro ϵ_i é tambem um parâmetro e, portanto, precisa ser estimada. Sob o modelo em (1), os estimadores de mínimos quadrados para os parâmetros envolvidos são dados por:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}; \tag{2}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2} = \frac{S_{xy}}{S_{xx}}; \tag{3}$$

$$\hat{\sigma}_\epsilon^2 = QME = \frac{SQE}{n-2} = \frac{S_{yy} - \hat{\beta}_1 S_{xy}}{n-2}. \tag{4}$$

Pelo teorema de Gauss-Markov temos que os estimadores de mínimos quadrados dos coeficientes do modelo, $\hat{\beta}_0$ e $\hat{\beta}_1$, dados em (2) e (3) respectivamente, têm a propriedade de serem estimadores não-viesados para os respectivos parâmetros e tem variância mínima entre todos os estimadores não-viesados que são combinações lineares dos y_i's, ou seja, dentre a classe dos estimadores não-viesados, aqueles dados em (2) e (3) são os que fornecem estimativas mais concentradas em torno dos valores reais de β_0, β_1, respectivamente. Também, o estimador $\hat{\sigma}_\epsilon^2$ dado em (4) é não-viesado para estimar σ_ϵ^2. É válido notar que a estimativa σ_ϵ^2 usando o estimador em (4) depende da soma de quadrado dos resíduos, SQE. Portanto, a qualidade da estimativa para este parâmetro é fortemente

dependente da adequação do modelo aos dados. Com isso, se a suposição de linearidade entre y e x e/ou a suposição de variância constante para todas as unidades amostrais (homocedasticidade) e/ou a suposição de independência (ausência de correlação) para os termos de erro não forem satisfeitas, então o estimador em (4) pode fornecer estimativas muito ruins (sub ou superestimar) para a variância do modelo, σ^2_ϵ. Formalmente, o que chamamos de viés (ou vício) de um estimador é definido como sendo

DEFINIÇÃO 3.1 (viés/vício de um estimador): *Seja $T \in \Theta$ um estimador de θ, onde Θ é o espaço paramétrico. O viés, do inglês bias, do estimador T, denotado por $b(T)$, é definido pela diferença entre o valor esperado deste estimador e o parâmetro que está sendo estimado, isto é,*

$$b(T) = \mathbb{E}(T) - \theta.$$

Se $\mathbb{E}(T) = \theta$, T é dito ser não-viesado para estimar θ e $b(T) = 0$.

Dizer que um estimador é não-viesado para o parâmetro ao qual este se destina estimar significa que, em média, o estimador produz estimativas iguais ao verdadeiro parâmetro. Ou seja, se pudéssemos colher todas as amostras possíveis de (y_i, x_i) de tamanho n da população de interesse e para cada uma dessas amostras calculássemos as estimativas $\hat{\beta}_0$, $\hat{\beta}_1$ e $\hat{\sigma}^2_\epsilon$, individualmente estas estimativas poderiam fornecer valores distantes dos valores reais dos respectivos parâmetros β_0, β_1 e σ^2_ϵ, mas a média aritmética de todas estas estimativas seriam exatamente iguais aos parâmetros, isto é,

$$\mathbb{E}[\hat{\beta}_0] = \beta_0, \quad \mathbb{E}[\hat{\beta}_1] = \beta_1 \quad \text{e} \quad \mathbb{E}[\hat{\sigma}^2_\epsilon] = \sigma^2_\epsilon. \tag{5}$$

Uma maneira de avaliar a qualidade das estimativas para os parâmetros do modelo é em termos de sua variabilidade e precisão. Temos que as variâncias dos estimadores de β_0 e β_1 são, respectivamente, dados por

$$\mathbb{V}\text{ar}[\hat{\beta}_0] = \sigma^2_\epsilon \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right) \quad \text{e} \quad \mathbb{V}\text{ar}[\hat{\beta}_1] = \frac{\sigma^2_\epsilon}{S_{xx}}. \tag{6}$$

Por sua vez, a precisão de um estimador é definida como sendo:

DEFINIÇÃO 3.2 (precisão de um estimador): *A precisão mede a proximidade de cada estimativa individual para o parâmetro θ obtida através de um estimador $T \in \Theta$ um estimador de θ com relação a média deste estimador, isto é, $precisao = T - \mathbb{E}(T)$.*

A Figura 1 apresenta uma ideia visual das Definições 3.1 (viés) e 3.2 (precisão).

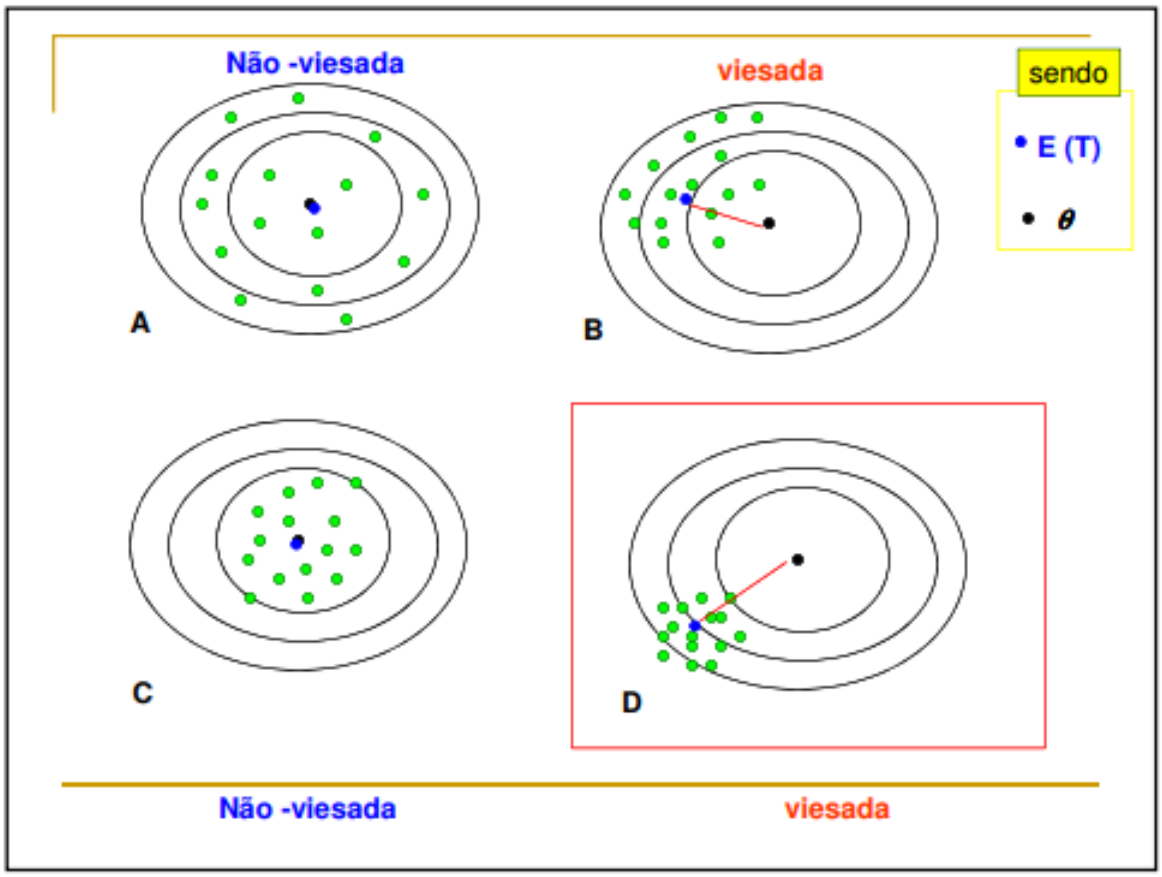

Figura 1: Viés de um estimador. No caso A, temos um exemplo de um estimador não-viesado e com baixa precisão. No caso B, temos um estimador viesado e com baixa precisão. No caso C, um estimador não-viesado e com alta precisão. No caso D, temos um exemplo de um estimador viesado e com alta precisão.

O erro quadrático médio (EQM) é uma outra medida da qualidade de estimadores e é definido como:

DEFINIÇÃO 3.3 (EQM de um estimador): *Seja $T \in \Theta$ um estimador de θ, onde Θ é o espaço paramétrico. O erro quadrático médio do estimador T, denotado por $EQM(T)$, é definido como sendo a média da diferença entre os valores do estimador e do parâmetro ao quadrado, isto é,*

$$EQM(T) = \mathbb{E}(T-\theta)^2 = \mathbb{V}\text{ar}(T) + b(T)^2,$$

onde $\mathbb{V}\text{ar}(T)$ e $b(T)$ denotam, respectivamente, a variância e o viés do estimador T. Se $\mathbb{E}(T) = \theta$, ou seja, T é não-viesado para estimar θ ($b(T) = 0$), então $EQM(T) = \mathbb{V}\text{ar}(T)$.

A Figura 2 apresenta uma ideia visual da relação entre as Definições 3.1 (viés), 3.2 (precisão) e 3.3 (EQM). Note que, baixa variância (alta precisão) não implica em baixo viés do estimador.

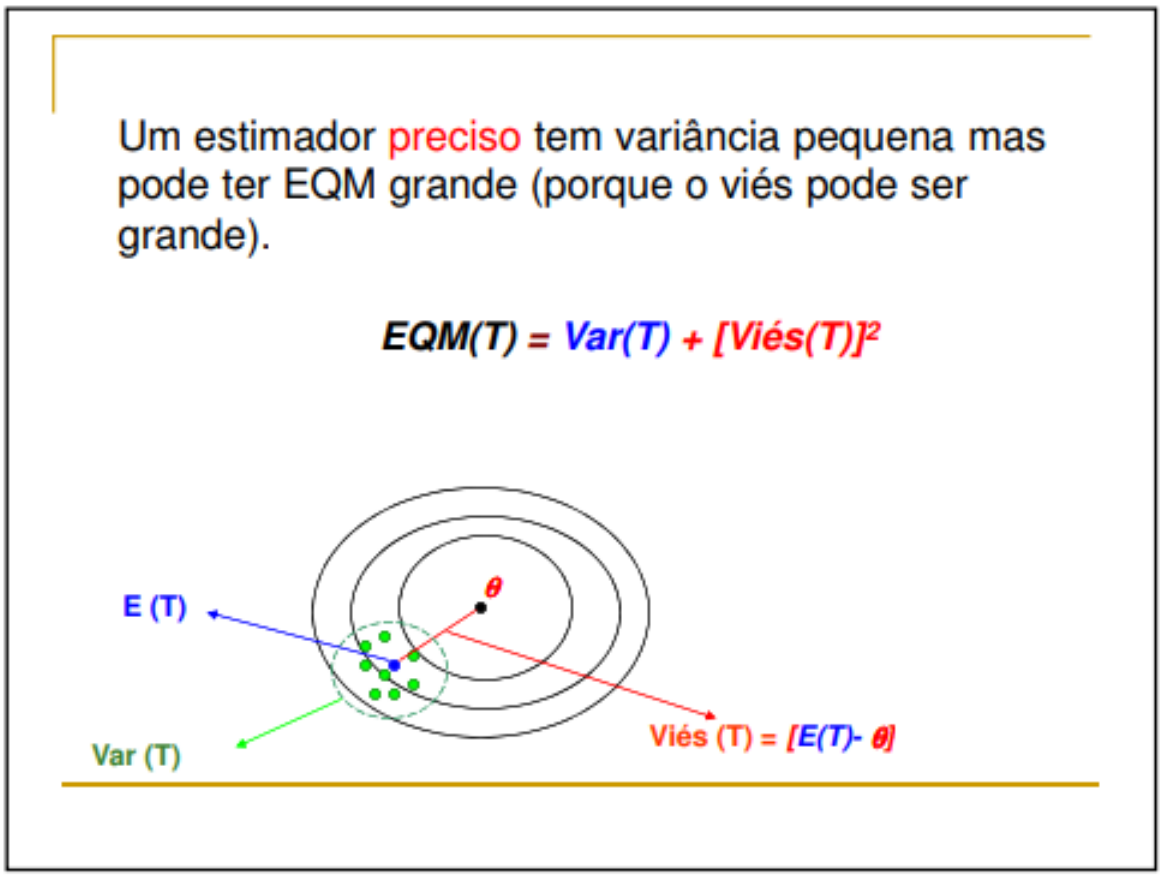

Figura 2: Relação viés, precisão e EQM de um estimador T para o parâmetro θ.

Tendo apresentado as Definições 3.1, 3.2 e 3.3, fica evidente como o método de Monte Carlo pode ser utilizado pra verificar a propriedade de ausência de viés estimadores de mínimos quadrados para os parâmetros do modelo de regressão linear e também avaliar sua qualidade (variabilidade/precisão) em função do tamanho da amostra n. Basicamente, seguindo os passos citados na Seção 2.1, devemos fazer:

Passo 1: Definir o modelo $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, com $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2_\epsilon)$ e $i = 1, \ldots, n$, escolhendo valores específicos para os parâmetros β_0, β_1 e σ^2_ϵ e para um tamanho de amostra desejado n. Os parâmetros verdadeiros serão, portanto, conhecidos. Além disso, é necessário pré-especificar um vetor de tamanho n para a covariável x, ou seja, fixe algum $\boldsymbol{x} = (x_1, \ldots, x_n)$. Uma maneira prática para isso é gerar a covariável usando uma distribuição de probabilidade qualquer.

Passo 2: Gerar n valores pseudo-aleatórios para o termo de acordo com a distribuição $N(0, \sigma_\epsilon^2)$ usando o valor de σ_ϵ^2 especificado no Passo 1.

Passo 3: Calcular os valores de $\boldsymbol{y} = (y_1, \ldots, y_n)$ substituindo os valores de β_0, β_1 e $\boldsymbol{x}$ especificados no Passo 1 e também os respectivos termos de erro gerados no Passo 2. Com isso, obteremos uma amostra $[\boldsymbol{y}, \boldsymbol{x}] = (y_1, x_1), \ldots, (y_n, x_n)$ sob o modelo de interesse.

Passo 4: Repetir os Passos 2 e 3 até se obter um total de M replicações de amostras de tamanho n do modelo.

Passo 5: Para cada uma das M amostras do tipo $[\boldsymbol{y}, \boldsymbol{x}]$, estime os parâmetros β_0, β_1 e σ_ϵ^2 usando os métodos de mínimos quadrados. Calcule o viés e EQM dos estimadores com base na amostra de estimativas de tamanho M que obteve. Compare com os valores verdadeiros especificados para tais parâmetros no Passo 1.

Estes passos podem ser seguidos para diferentes especificações do modelo e tornam fácil o estudo do efeito que o tamanho da amostra n, o número de replicações M e a magnitude da variância do erro σ_ϵ^2 têm sobre a qualidade e precisão das estimativas quando consideramos os estimadores de mínimos quadrados dados em (2), (3) e (4).

Como os Passos 1, 2 e 3 envolvem a simulação de um modelo de regressão linear para parâmetros especificados, antes de utilizar o MMC para avaliar os estimadores de MQ par os parâmetros do modelo, apresentamos na próxima seção um exemplo de como a geração do modelo pode ser feita utilizando o *software* R.

2.2.1 Simulando um modelo de regressão linear simples no *software* R

A simulação de um modelo de regressão depende basicamente da definição de valores para todos os parâmetros envolvidos, a escolha de um tamanho de amostra n, um vetor de covariáveis $\boldsymbol{x} = (x_1, \ldots, x_n)$ e uma função de distribuição para o termo de erro. No caso do modelo de regressão linear simples, para que os parâmetros do modelo possam ser estimados de forma apropriada, o tamanho amostral n deve ser no mínimo igual a 3. Nos estudos feitos neste trabalho utilizou-se sempre $n \geq 10$.

No código em R que segue é gerada uma mostra de tamanho $n = 100$ do modelo de regressão linear simples $y_i = 1 - x_i + \epsilon_i$, com $\epsilon_i \overset{iid}{\sim} N(0, 1)$ para $i = 1, \ldots, 100$. Portanto, os valores verdadeiros para os parâmetros do modelo são: $\beta_0 = 1$, $\beta_1 = -1$ e $\sigma_\epsilon^2 = 1$. Os valores da covariável x são gerados de uma distribuição exponencial com taxa 5, ou seja, assume-se que $x_i \overset{iid}{\sim} exp(5)$. Vale ressaltar que qualquer outra distribuição poderia ter sido utilizada para gerar o vetor $\boldsymbol{x} = (x_1, \ldots, x_n)$, inclusive distribuições que gerem variáveis aleatórias discretas, pois os valores estarão fixados e teoricamente não influenciarão na qualidade das estimativas (afinal, a inferência é feita condicionando/fixando na covariável!). O uso do comando *set.seed* é importante para que, ao reproduzir o código, os mesmos valores sejam gerados tanto para a covariável quanto para o termo de erro. Ao final, teremos uma amostra de tamanho n para $[\boldsymbol{y}, \boldsymbol{x}] = (y_1, x_1), \ldots, (y_n, x_n)$ sob o modelo especificado. Usando tal amostra podemos estimar os parâmetros β_0, β_1 e σ_ϵ^2 usando os métodos de mínimos quadrados e verificar a magnitude do erro com relação aos valores verdadeiros que foram especificados para estes parâmetros.

```
rm(list=ls(all=TRUE)) #limpa workspace do ambiente R
set.seed(12345)       #fixa semente para a geração de números pseudo-aleatórios

#
## Exemplo de geração dos dados para um modelo de regressão linear simples:

##Especificando valores reais dos parâmetros da reta de regressão:
beta0 <- 1
beta1 <- -1
sigma2 <- 1

##Definindo o tamanho da amostra "n"
n <- 100

##Gerando um vetor de tamanho n para covariável "x"
x <- rexp(n, rate=5)

##Gerando o erro:
erro <- rnorm(n, mean=0, sd=sqrt(sigma2)) #Média é 0 pela definição do modelo.

##Calculando a variável resposta para o modelo especificado:
y <- beta0 + beta1*x + erro

##- Note que agora temos uma amostra de tamanho n de vetores (y_i,x_i).
##- Podemos então supor que tinhamos disponíveis estes dados e
## que o interesse era ajustar o modelo de regressão linear.
##- Como sabemos os parâmetros verdadeiros que geraram os dados,
## podemos compara-los com as estimativas que serão obtidas.

#
## Ajuste do modelo:
modelo <- lm(y~x)
betas.estimados <- modelo$coefficients
sigma2.estimado <- sum(modelo$residuals^2)/modelo$df.residual

##Podemos calcular os erros de estimação...
erro.estimacao.betas <- betas.estimados - c(beta0,beta1)
erro.estimacao.sigma2 <- sigma2.estimado - sigma2

## ...e também calcular as variâncias se quisermos:
Sxx <- sum((x-mean(x))^2)
var.beta1 <- sigma2.estimado/Sxx
var.beta0 <- sigma2.estimado*(1/n+mean(x)^2/Sxx)

##- FIM
```

Usando esse procedimento para simulação de várias amostras do modelo especificado teremos então condições de calcular o viés e EQM dos estimadores de MQ no modelo de regressão linear, veja Seção 2.3.

2.3 Verificando as propriedades dos estimadores de MQ para os parâmetros do modelo de regressão linear via MMC

Nesta seção o método de Monte Carlo (MMC) é aplicado para efetivamente avaliar a propriedade de ausência de viés dos estimadores de mínimos quadrados no contexto de um modelo de regressão linear simples. Para um modelo de regressão linear múltiplo o procedimento e a função fornecida podem ser facilmente adaptados. Na Seção 2.3.1

2.3.1 Uma função em R para o uso do MMC no modelo de regressão linear

Nesta seção a função implementada para realizar o estudo das propriedades dos estimadores de mínimos quadrados no modelo de regressão linear simples via MMC é apresentada.

A função tem o nome *MMC.regressao* e depende de 5 argumentos de entrada: o tamanho da amostra n, o número de replicações M e os valores verdadeiros a serem considerados para os parâmetros do modelo: β_0, β_1 e σ^2_ϵ.

Dentro da função, gera-se a covariável e as M réplicas de $[\boldsymbol{y}, \boldsymbol{x}] = (y_1, x_1), \ldots, (y_n, x_n)$ sob o modelo especificado. Para cada réplica, ajusta-se o modelo linear e as estimativas de mínimos quadrados são guardadas juntamente com os erros de estimação e EQMs com base nos valores reais dos parâmetros que foram pré-especificados.

Como saída, a função *MMC.regressao* retorna, dentro do diretório em que se está trabalhando, gráficos *box-plot* para as estimativas dos parâmetros com base nas M amostras geradas e, também, retorna dentro do ambiente R uma tabela com os valores médios das estimativas, o viés, o EQM e as variâncias associadas.

```
rm(list=ls(all=TRUE)) #limpa workspace

## Construindo a função para o estudo Monte Carlo:

MMC.regressao <- function(n, M, beta0, beta1, sigma2){
set.seed(12345)

x <- rexp(n, rate=5) #Gerando a covariável.

betas.estimados <- matrix(,M,2) #Guarda estimativas betas, uma réplica
 #em cada linha
sigma2.estimado <- numeric(M) #Guarda estimativas sigma2
erro.estimacao = erro.quadratico <- matrix(,M,3) #Guarda erros e EQMs
```

```
for(r in 1:M){ #Gerando as M amostras do tipo (y,x) como no instruído
#no Passo 3
erro <- rnorm(n, mean=0, sd=sqrt(sigma2))
y <- beta0 + beta1*x + erro
modelo <- lm(y~x)
betas.estimados[r,] <- coefficients(modelo)
sigma2.estimado[r] <- sum(modelo$residuals^2)/modelo$df.residual

erro.estimacao[r,]<-c(betas.estimados[r,],sigma2.estimado[r]) -
c(beta0,beta1,sigma2)
erro.quadratico[r,] <- erro.estimacao[r,]^2
}

windows()
boxplot(betas.estimados,main=paste("Boxplot betas para n=",n,"e M=",M))
savePlot(filename = paste("Boxplot betas para n=",n,"e M=",M),type="pdf")
dev.off()

windows()
boxplot(sigma2.estimado,main=paste("Boxplot sigma2 para n=",n,"e M=",M))
savePlot(filename = paste("Boxplot sigma2 para n=",n,"e M=",M),type="pdf")
dev.off()

media.estimativas <- c(colMeans(betas.estimados), mean(sigma2.estimado))
eqm <- round(colMeans(erro.quadratico),5)
vies <- round(colMeans(erro.estimacao),5)
var.estimadores <- round(eqm - vies^2,5)

saida <- rbind(media.estimativas,eqm,vies,var.estimadores)
colnames(saida) = c(paste("beta0 =",beta0),paste("beta1 =",beta1),
paste("sigma2 =",sigma2))
return(saida)
}
```

Vamos aplicar esta função simulando o modelo com os parâmetros especificados no exemplo da Seção 2.2.1, ou seja, serão geradas uma amostras de tamanho $n = 100$ do modelo de regressão linear simples $y_i = 1 - x_i + \epsilon_i$, com $\epsilon_i \overset{iid}{\sim} N(0, 1)$ para $i = 1, \ldots, 100$. Portanto, os valores verdadeiros para os parâmetros do modelo são: $\beta_0 = 1$, $\beta_1 = -1$ e $\sigma^2_\epsilon = 1$. A covariável x é gerada de forma que $x_i \overset{iid}{\sim} exp(5)\ \forall\ i$. Consideremos um total de $M = 100$ réplicas para o estudo Monte Carlo.

Para obter os resultados neste contexto, após executar a função *MMC.regressao* criada acima, basta usarmos o comando MMC.regressao(n=100,M=100,beta0=1,beta1=-1,sigma2=1) ou, simplesmente, MMC.regressao(100,100,1, -1,1), lembrando que os parâmetros de entrada da função devem ser especificados na ordem em que aparecem na definição da mesma.

Ao executar tal comando, obtemos os resultados mostrados nas Figuras 3 e 4. A propriedade de ausência de viés dos estimadores já é bem aparente para amostras de tamanho $n = 100$, pois a média das estimativas já são bem próximas dos valores reais usados para simular os dados. O viés não é muito grande (na teoria sabemos que o viés é nulo) e, portanto, o EQM e a variância dos estimadores apresentam valores próximos. Também com base nos gráficos de *box-plots* vemos que as estimativas de β_0 foram as que apresentaram menor dispersão neste caso.

```
> MC.regressao(100, 100, 1, -1, 1)
                   beta0 = 1 beta1 = -1 sigma2 = 1
media.estimativas  1.004896  -1.033221  0.9933106
eqm                0.015650   0.219730  0.0200200
vies               0.004900  -0.033220 -0.0066900
var.estimadores    0.015630   0.218630  0.0199800
>
```

Figura 3: Média, viés, EQM e variância dos estimadores de mínimos quadrados via MMC para o caso $n = 100$, $M = 100$, $\beta_0 = 1$, $\beta_1 = -1$ (gráfico à esquerda, onde estão representados por 1 e 2, respectivamente) e $\sigma^2_\epsilon = 1$ (gráfico à direita).

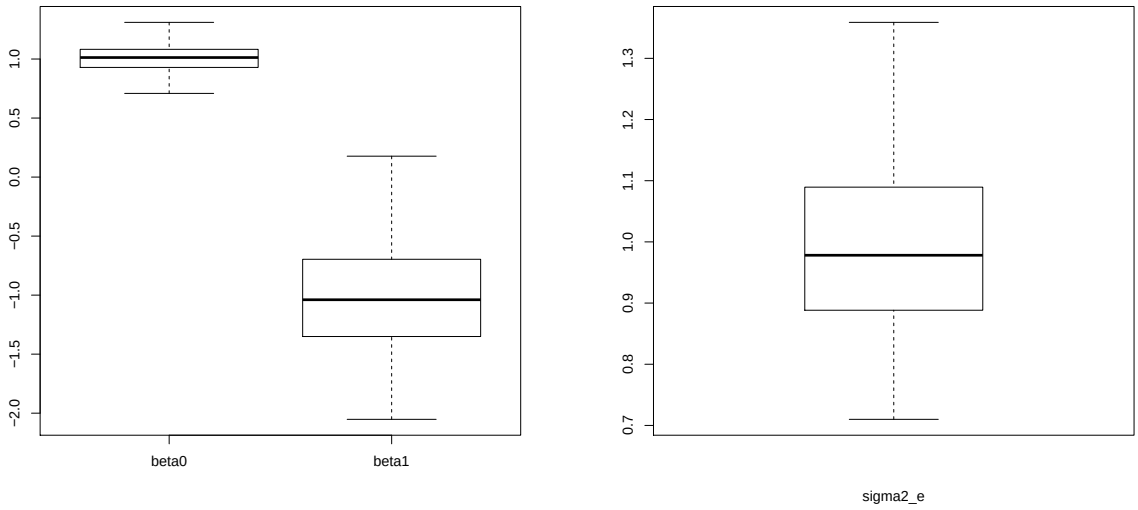

Figura 4: Gráficos *boxplot* das estimativas de mínimos quadrados para β_0, β_1 (à esquerda) e σ^2_ϵ (à direita) no caso $n = 100$, $M = 100$, $\beta_0 = 1$, $\beta_1 = -1$ e $\sigma^2_\epsilon = 1$.

2.3.2 Efeito do tamanho da amostra n na qualidade das estimativas de MQ

Nesta seção a função apresentada na Seção 2.3.1 será utilizada para implementarmos o MMC com diferentes tamanhos de amostra n. O objetivo é analisar o efeito de n na qualidade das estimativas em termos de viés, EQM e variância. Considere para tal os valores de $n = (10, 20, 50, 100, 500, 1000)$. Os parâmetros β_0, β_1 e σ^2_ϵ do modelo a ser simulado, assim como a covariável x, são fixados como aqueles na Seção 2.3.1. Serão consideradas $M = 100$ réplicas. A função *MMC.regressao* precisa ser executada e então um *loop* iterativo é estabelecido para gerar de uma só vez os resultados para todos os tamanhos de amostra especificados. Os comandos em R necessários para obter os resultados são apresentados no Apêndice A.

Os resultados são apresentados nas Figuras 5, 6 e 7. É evidente que, no geral, quando maior o tamanho da amostra, menores são a variância, o viés e EQM. Isso deveria mesmo acontecer, pois para amostras maiores se tem mais informação sobre a população de interesse. Conforme n cresce, mais evidente fica a propriedade de ausência de viés dos estimadores via método dos mínimos quadrados. Os *box-plots* permitem visualizar a maior concentração das estimativas em torno do valor real conforme se aumenta o tamanho da amostra.

```
> print(media.estimativas)
         beta0 = 1 beta1 = -1 sigma2 = 1
n = 10   0.9852544 -0.7743965  0.9356410
n = 20   1.0086224 -1.0336816  0.9454556
n = 50   1.0084162 -1.0459322  0.9765573
n = 100  1.0048962 -1.0332213  0.9933106
n = 500  0.9974640 -0.9814994  1.0025469
n = 1000 1.0084126 -1.0395713  1.0015476
>
```

A: média estimativas

```
> print(var.estimadores)
         beta0 = 1 beta1 = -1 sigma2 = 1
n = 10     0.15194    0.73166    0.20418
n = 20     0.09477    0.48323    0.11664
n = 50     0.03373    0.27434    0.03417
n = 100    0.01563    0.21863    0.01998
n = 500    0.00346    0.05841    0.00300
n = 1000   0.00238    0.03046    0.00210
```

B: variância estimativas

```
> print(eqm)
         beta0 = 1 beta1 = -1 sigma2 = 1
n = 10     0.15216    0.78256    0.20832
n = 20     0.09484    0.48436    0.11961
n = 50     0.03380    0.27645    0.03472
n = 100    0.01565    0.21973    0.02002
n = 500    0.00347    0.05875    0.00301
n = 1000   0.00245    0.03203    0.00210
>
```

C: EQM

```
> print(vies)
         beta0 = 1 beta1 = -1 sigma2 = 1
n = 10    -0.01475    0.22560   -0.06436
n = 20     0.00862   -0.03368   -0.05454
n = 50     0.00842   -0.04593   -0.02344
n = 100    0.00490   -0.03322   -0.00669
n = 500   -0.00254    0.01850    0.00255
n = 1000   0.00841   -0.03957    0.00155
>
```

D: viés

Figura 5: Média, viés, EQM e variância dos estimadores de mínimos quadrados via MMC para o caso $n = (10, 20, 50, 100, 500, 1000)$, $M = 100$, $\beta_0 = 1$, $\beta_1 = -1$ e $\sigma^2_\epsilon = 1$.

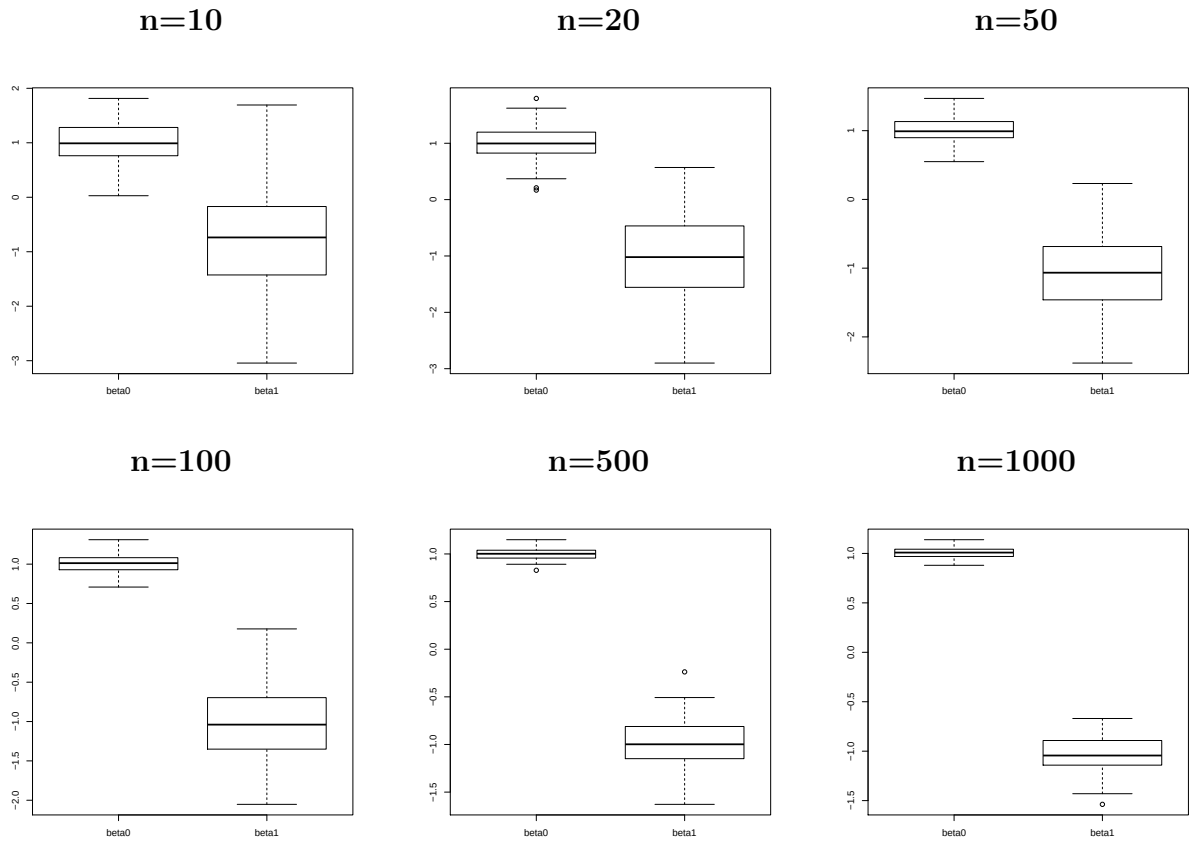

Figura 6: Gráficos *boxplot* das estimativas de mínimos quadrados para β_0, β_1 no caso $n = (10, 20, 50, 100, 500, 1000)$, $M = 100$, $\beta_0 = 1$, $\beta_1 = -1$ e $\sigma_\epsilon^2 = 1$.

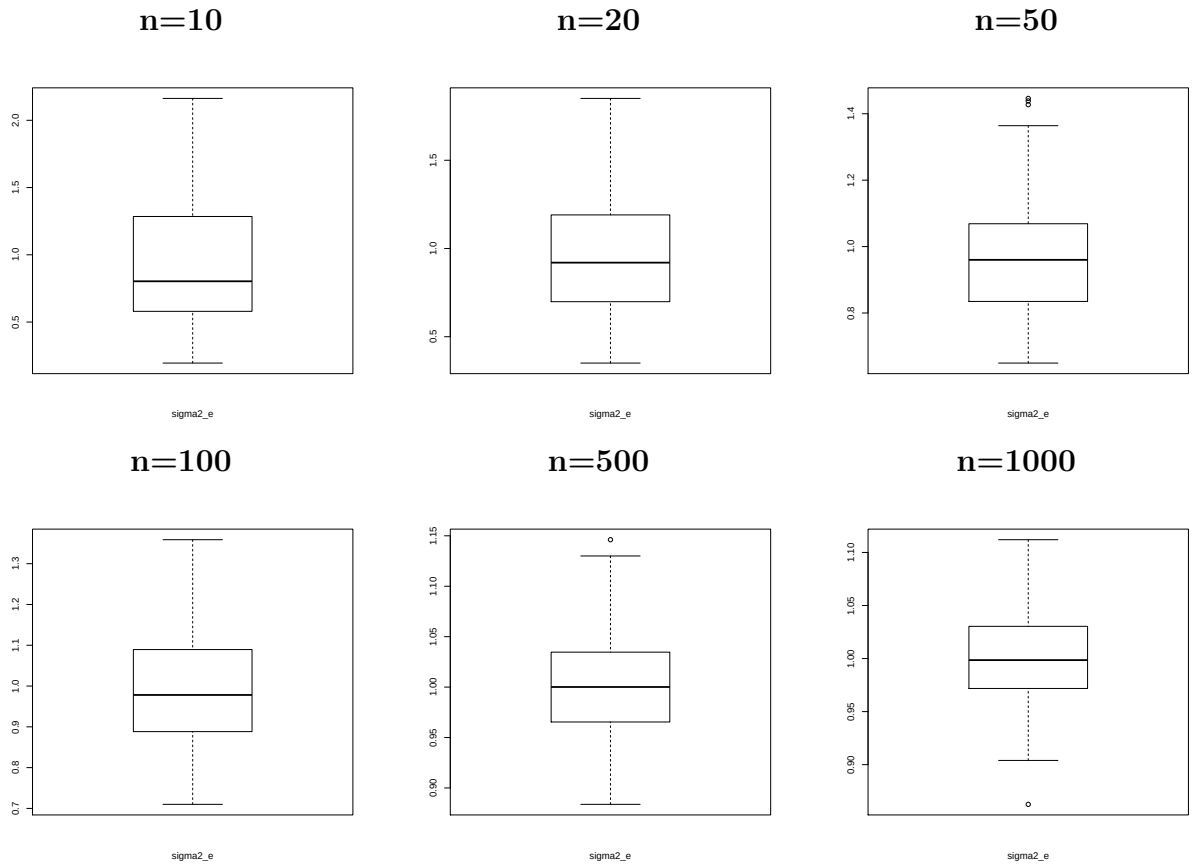

Figura 7: Gráficos *boxplot* das estimativas de mínimos quadrados para σ^2_ϵ no caso $n = (10, 20, 50, 100, 500, 1000)$, $M = 100$, $\beta_0 = 1$, $\beta_1 = -1$ e $\sigma^2_\epsilon = 1$.

2.3.3 Efeito da variância do erro σ_ϵ^2 na qualidade das estimativas de MQ

Nesta seção a função apresentada na Seção 2.3.1 será utilizada para implementarmos o MMC com diferentes valores para a variância do erro σ_ϵ^2. O objetivo é analisar o efeito do valor de σ_ϵ^2 na qualidade das estimativas em termos de viés, EQM e variância. Considere para tal os valores de $\sigma_\epsilon^2 = (1, 10, 100)$. O tamanho da amostra será fixado como sendo $n = 100$ e o total de réplicas será $M = 100$. Os parâmetros β_0, β_1 do modelo a ser simulado, assim como a covariável x são fixados como aqueles na Seção 2.3.1. A função *MMC.regressao* precisa ser executada e então um *loop* iterativo é estabelecido para gerar de uma só vez os resultados para todos os valores de variância especificados. Os comandos em R necessários para obter os resultados são apresentados no Apêndice B.

Os resultados são apresentados nas Figuras 8, 9 e 10. É evidente que, no geral, quando maior a variância do erro, maiores são a variância das estimativas, o viés e EQM. Isso deveria mesmo acontecer, pois para amostras geradas com variância maior há mais ruído nos dados e, com isso, maior a ocorrência de valores atípicos e as amostras serão mais diferentes entre si. Logo, estimativas também apresentarão mais variabilidade de uma amostra pra outra. Conforme σ_ϵ^2 cresce, menos evidente fica a propriedade de ausência de viés dos estimadores via método dos mínimos quadrados. Os *box-plots* permitem visualizar a maior concentração das estimativas em torno do valor real conforme se aumenta a variância do erro σ_ϵ^2, note que as escalas do eixo vertical dos gráficos são diferentes.

```
> print(media.estimativas)
            beta0 = 1 beta1 = -1     sigma2
sigma1= 1    1.002592 -0.9869403   1.005805
sigma1= 10   1.076221 -1.2726967  10.179965
sigma1= 100  1.109417 -1.4406416 101.220789
```

A: média das estimativas

```
> print(var.estimadores)
             beta0 = 1 beta1 = -1       sigma2
sigma1= 1   0.01438762  0.2684992   0.02233535
sigma1= 10  0.28021880  3.0048311   2.17687725
sigma1= 100 2.37326525 53.2532254 158.03334893
```

B: variância das estimativas

```
> print(eqm)
             beta0 = 1 beta1 = -1       sigma2
sigma1= 1   0.01439434  0.2686698   0.02236905
sigma1= 10  0.28602848  3.0791946   2.20926466
sigma1= 100 2.38523735 53.4473904 159.52367557
```

C: EQM das estimativas

```
> print(vies)
              beta0 = 1  beta1 = -1      sigma2
sigma1= 1   0.002592309  0.01305969 0.005805246
sigma1= 10  0.076221254 -0.27269672 0.179965029
sigma1= 100 0.109417106 -0.44064162 1.220789350
```

D: viés das estimativas

Figura 8: Média, viés, EQM e variância dos estimadores de mínimos quadrados via MMC no caso $\sigma_\epsilon^2 = (1, 10, 100)$, $n = 100$, $M = 100$, $\beta_0 = 1$ e $\beta_1 = -1$.

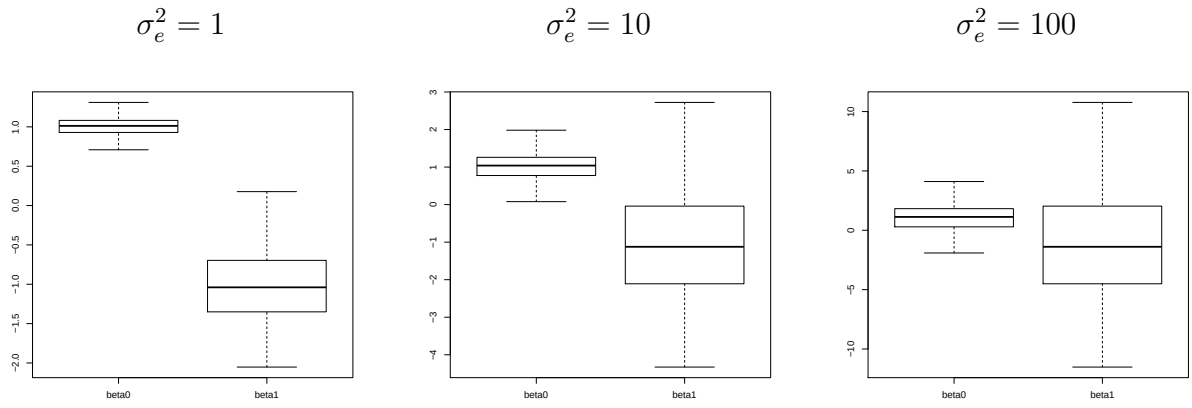

Figura 9: Gráficos *boxplot* das estimativas de mínimos quadrados para β_0 e β_1 no caso $\sigma_\epsilon^2 = (1, 10, 100)$, $n = 100$, $M = 100$, $\beta_0 = 1$ e $\beta_1 = -1$.

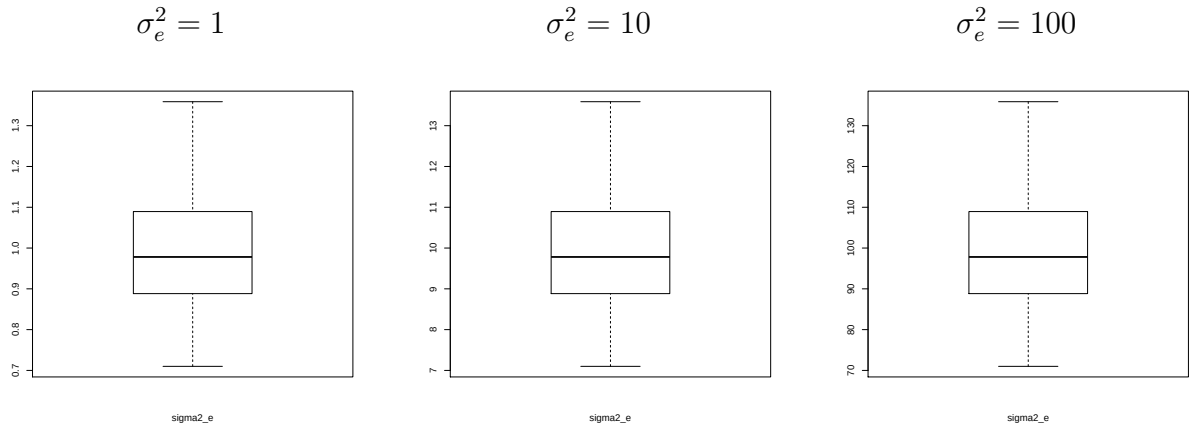

Figura 10: Gráficos *boxplot* das estimativas de mínimos quadrados para σ_ϵ^2 no caso $\sigma_\epsilon^2 = (1, 10, 100)$, $n = 100$, $M = 100$, $\beta_0 = 1$ e $\beta_1 = -1$.

3 Sugestões de Exercícios Práticos e Teóricos

Nesta seção são enumerados alguns exercícios práticos e teóricos cuja resolução pode ajudar na fixação do conteúdo visto em sala de aula e, além disso, promover maior entendimento dos conceitos envolvidos na análise de regressão linear.

3.1 Exercícios Práticos

Exercício P1. Identifique a variável resposta e a variável explicativa em cada caso:

a) As variáveis de interesse são as toneladas de adubo orgânico por ha e a produção da cultura A por ha.

RESPOSTA: Variável resposta: Nível de produção da cultura A por ha. Variável explicativa: Quantidades de adubo orgânico por ha.

b) Pretende-se estudar a relação entre pressão sanguínea sistólica e consumo de álcool.

RESPOSTA: Variável resposta: Pressão sanguínea. Variável explicativa: Níveis de consumo do álcool.

c) Deseja-se verificar se o tempo de treinamento é importante para avaliar o desempenho na execução de uma dada tarefa.

RESPOSTA: Variável resposta: Desempenho. Variável explicativa: Tempo de treinamento.

d) Suponha que uma cadeia de supermercados tenha financiado um estudos dos gastos com mercadorias para famílias de 4 pessoas. O estudo se limitou a famílias com renda líquida entre 8 e 20 salários mínimos.

RESPOSTA: Variável resposta: Gastos com mercadorias. Variável explicativa: Renda.

e) Estuda-se a relação entre o uso do fumo e a incidência de câncer pulmonar, relacionando o número de anos que uma pessoa fumou com a percentagem de incidência de câncer pulmonar em cada grupo.

RESPOSTA: Variável resposta: Porcentagem de incidência do câncer. Variável explicativa: Consumo de cigarros (Fumante).

Exercício P2. A Tabela 1 indica o valor y do aluguel e a idade x de 5 casas.

Tabela 1: preço (y) e idade (x) das casas

y	10	13	5	7	20
x	4	3	6	5	2

a) Construa o gráfico de dispersão para as valores observados das variáveis y e x.

RESPOSTA:

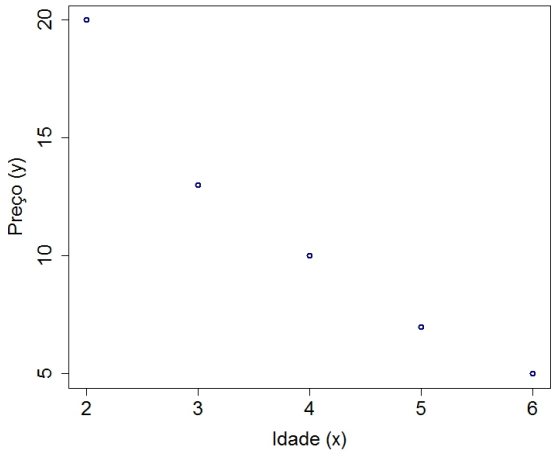

Figura 11: Diagrama de dispersão para os valores de x e y.

```
# Código do R
plot(x,y, col="darkblue",lwd=2, cex.axis=1.6, cex.lab=1.6,
 xlab="Idade (x)", ylab="Preço (y)" )
```

b) Encontre a reta de mínimos quadrados, supondo a relação $E(y|x) = \beta_0 + \beta_1 x$.

RESPOSTA:

$\sum_{i=1}^{n} x_i = 20$	$\sum_{i=1}^{n} y_i = 55$
$\sum_{i=1}^{n} x_i^2 = 90$	$\sum_{i=1}^{n} y_i^2 = 743$
$\bar{x} = 5$	$\bar{y} = 11$
$S_{xx} = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2 = 10$	
$S_{xy} = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y} = -36$	$n = 5$

A partir do resumo da tabela acima obtemos os estimadores de mínimos quadrados (EMQ):

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{-36}{10} = -3,600 \quad \text{e} \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 25,4$$

Portanto, a reta de mínimos quadrados é dada por:

$$E(y_i|x_i) = \hat{\beta}_0 + \hat{\beta}_1 x_i = 25,4 - 3,600 x_i.$$

No R, o ajuste do modelo de regressão linear é feito por meio da função *lm()*, ver detalhes em Reis *et. al* (2009) [i].

c) Retorne ao gráfico construído no item (a) e adicione a reta ajustada no item (b). Você acha que o modelo adotado é razoável?

RESPOSTA:

Com base na Figura 12 observamos que o modelo é razoável para descrever a relação existente entre a idade das casas e o preço

```
# Código do R
fit=lm(y~x);
abline(fit, col="red")
```

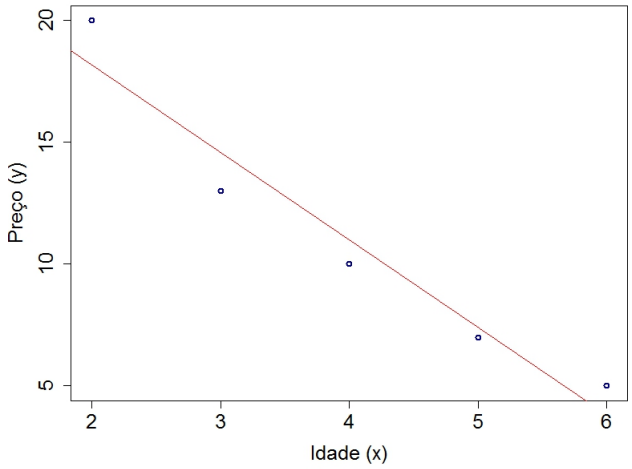

Figura 12: Ajuste da reta de regressão.

d) Qual o significado do coeficiente angular e do coeficiente linear para este caso.

RESPOSTA:

(β_0) Supondo que a idade das casas esteja codificada em 0, 1, 2 etc, o coeficiente linear β_0 representa o preço médio das casas antes de completarem o primeiro ano, isto é, quando a idade codificada assume o valor 0.

(β_1) O coeficiente angular da reta β_1 indica o quanto varia o preço médio, para cada variação unitária da idade das casas.

Exercício P3. Um laboratório está interessado em medir o efeito da temperatura sobre a potência de um antibiótico. Dez amostras de 50 gramas cada foram guardadas a diferentes temperaturas e após 15 dias mediu-se a potência. Os resultados são mostrados na Tabela 2.

Tabela 2: temperatura e potência

temperatura	30°		50°			70°			90°	
potência	38	43	32	26	33	19	27	23	14	21

a) Faça a representação gráfica dos dados. Calcule o coeficiente de correlação linear e comente sobre a validade da suposição de linearidade entre as variáveis.

RESPOSTA:

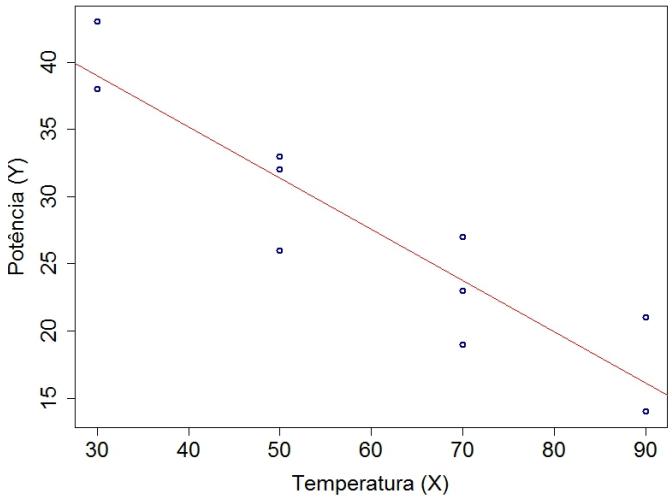

Figura 13: Ajuste da reta de regressão.

X: Temperatura	Y: Potência
$\sum_{i=1}^{n} x_i = 600$	$\sum_{i=1}^{n} y_i = 276$
$\sum_{i=1}^{n} x_i^2 = 40200$	$\sum_{i=1}^{n} y_i^2 = 8338$
$\bar{x} = 60$	$\bar{y} = 27.6$
$S_{xx} = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2 = 4200$	
$S_{xy} = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y} = -1600$	$n = 10$

O coeficiente de correlação entre a temperatura e a potência é dado por:

$$r = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sqrt{\left(\sum_{i=1}^{n} x_{i=1}^2 - n\bar{x}^2\right)\left(\sum_{i=1}^{n} y_{i=1}^2 - n\bar{y}^2\right)}} = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \frac{-1600}{\sqrt{4200 * 720}} \approx -0,920.$$

O coeficiente de correlação e o gráfico da Figura 13 evidenciam a existência de uma correlação forte e negativa entre a temperatura e a potência dos antibióticos. No R, este coeficiente pode ser obtido utilizando a função *cor()*, use o comando *help(cor)* para obter informações sobre a utilização desta função.

b) Ajuste o modelo de regressão linear apropriado, apresentando sua equação.

RESPOSTA:

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{-1600}{4200} = -0,3809524 \quad \text{e} \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} = 50,45714.$$

E portanto,

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i = 50,45714 - 0,3809524 x_i$$

c) Interprete, de acordo com o problema, os coeficientes do modelo obtido em (b).

RESPOSTA:

$\hat{\beta}_0 = 50.45714$ representa a média de $\hat{y}_i$ quando a temperatura é de 0^o.

$\hat{\beta}_1 = -0.38095$ representa o grau de mudança na potência média dos antibióticos, para cada variação unitária da temperatura. Isto é, a potência média dos antibióticos diminui em 0.38095 unidades para cada variação unitária da temperatura.

d) Usando o modelo ajustado em (b) responda:

i) Qual seria a potência do antibiótico se amostras de 50 gramas fossem guardadas a uma temperatura de 65°?

RESPOSTA: $E(y_i|x_i = 65^o) = 50.45714 - 0.3809524 * 65 = 25.69539$.

ii) Qual seria a potência do antibiótico se amostras de 50 gramas fossem guardadas a uma temperatura de 25°?

RESPOSTA: O modelo não é apropriado para prever a temperatura de 25^o, pois estaríamos cometendo uma extrapolação: o valor 25^o para a temperatura não pertence à faixa de valores observados.

Exercício P4. (Simulando dados de um modelo de regressão linear simples) Acredita-se que a resistência elétrica (em ohms/cm) de fios de aço está relacionada com o carbono contido (em porcentagem) no aço através do modelo $Y_i = 13,4 + 13,3X_i + \epsilon_i$, em que Y é a resistência, X é o percentual de carbono contido e ϵ é um erro aleatório com distribuição normal de média 0 e variância 0,62.

a) Interprete o coeficiente da percentagem de carbono neste modelo.

RESPOSTA:

$\beta_0 = 13,4$ é o valor da resistência elétrica quando não existe nenhuma porcentagem de carbono contido.

$\beta_1 = 13,3$ representa a mudança na resistência elétrica média para cada variação unitária da porcentagem de carbono contido.

b) Qual a resistência esperada de um fio de aço com 0.5% de carbono contido?

RESPOSTA:

$E(y_i|x_i = 0,5\%) = 13,4 + 13,3 * 0,5 = 20,5$ohms/cm.

c) (Pode ser computacional) Simule a resistência de 5 fios de aço para cada uma das seguintes porcentagens de carbono contido: 0,2%, 0,4%, 0,6%, 0,8%.

RESPOSTA: Os dados podem ser obtidos usando o comando do R:

```
x=c(rep(0.2, 5), rep(0.4, 5),
rep(0.6, 5), rep(0.8, 5))

y=13.4+13.3*x+rnorm(20,0, sqrt(0.62))
```

d) (Pode ser computacional) Represente graficamente os dados obtidos em (c). A suposição de linearidade parece ser verdadeira?

RESPOSTA: Com base no gráfico observamos que é valida a suposição de linearidade.

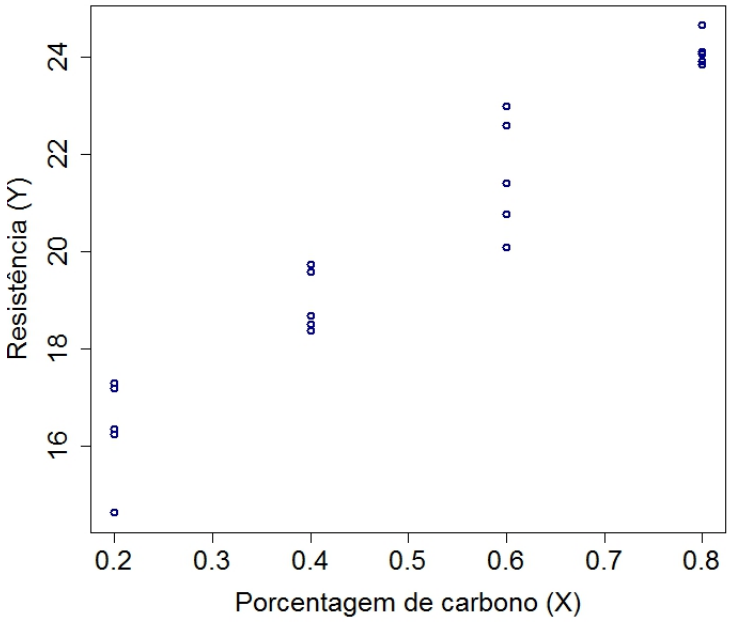

Figura 14: Diagrama de dispersão.

```
plot(x,y, col="darkblue", lwd=2, cex.axis=1.6,
cex.lab=1.6, xlab="Porcentagem de carbono (X)",
ylab="Resistência (Y)")
```

Exercício P5. Na Tabela 3 a seguir estão os dados observados de 5 carros, em que y indica o número de acidentes sofridos por carros viajando à velocidade de x km/h.

x_i	80	90	100	110	120
y_i	79	83	90	95	99

Tabela 3: Número de acidentes (y) e velocidade (x)

Use o R para ajustar o modelo, $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ e:

Resposta:

	Coef.	Erro padrão	T	Valor-p
Intercepto	37.,200	2,85657	13.02	0,000978
x	0,5200	0,02828	18.39	0,000351

O modelo de regressão é dado por:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i = 37,2 + 0,52 x_i$$

Pelos resultados da tabela, conclui-se que tanto o intercepto quanto o coeficiente angular ambos são significativos ao nível de 5%.

a) Verifique a suposição de normalidade dos resíduos através do gráfico de probabilidade Normal.

Resposta:

Teste de Normalidade:

H_0 : Os resíduos seguem uma distribuição normal

H_1 : Os resíduos não seguem uma distribuição normal

Estatística do teste	valor-p
0,92006	0,5303

Tabela 4: Teste de normalidade de Shapiro-Wilk

O gráfico de probabilidade normal (*QQ-Plot*), Figura 15, mostrou que não existem grandes desvios dos pontos em relação a reta. As mesmas conclusões podem ser tiradas com base no teste de Shapiro-Wilks ($p - valor > 0,05$).

b) Faça uma análise apropriada para os gráficos dos resíduos (resíduos vs $\hat{y}_i$ e resíduos studentizado vs $\hat{y}_i$).

Resposta:

O gráfico dos resíduos vs valores estimados (ou variável explicativa), Figura 16, não mostrou nenhum padrão na distribuição dos pontos, isto é, os resíduos estão distribuídos de forma aleatória em torno de zero.

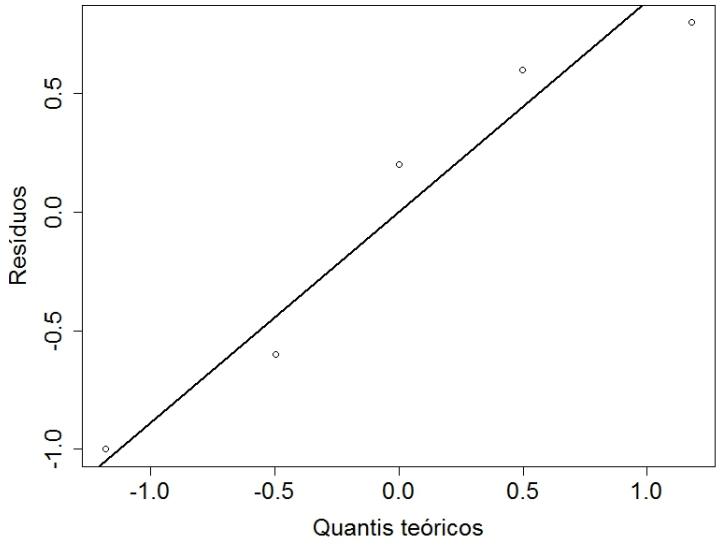

Figura 15: QQ Plot.

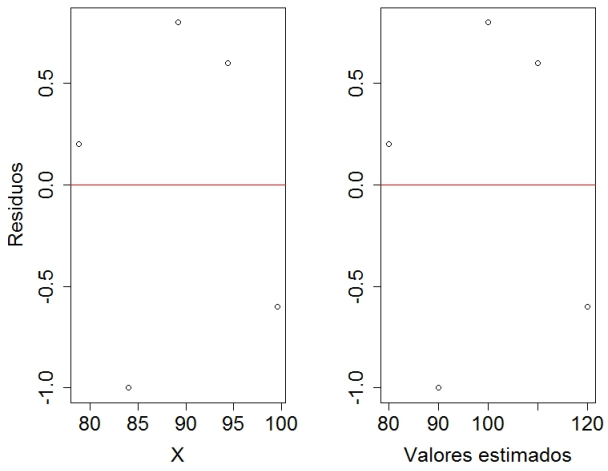

Figura 16: Resíduos vs x e Resíduos vs $\hat{y}_i$.

c) Apresente o histograma dos resíduos e, confronte-os com as conclusões do item (a).

Resposta:

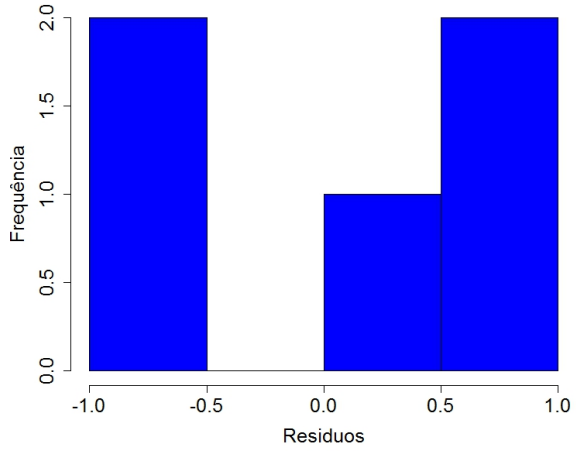

Figura 17: Histograma dos resíduos.

O histograma dos resíduos não permite conclusão clara quanto à validade da suposição de normalidade que fora verificada no item (a), o que pode estar relacionado ao fato de o tamanho amostral ser pequeno a ponto do gráfico não ser conclusivo neste sentido.

d) Reporte os problemas encontrados na análise dos resíduos.

Resposta:

Em geral, a análise de resíduos aqui feita, não mostrou nenhuma violação das suposições.

e) A partir das análise feitas nos itens anteriores, conclua sobre a relação entre y e x.

Resposta:

De acordo com resultados dos itens anteriores a velocidade (x) influencia significativamente na ocorrência dos acidentes (y), então o número de acidentes pode ser explicada com a velocidade com o qual os carros viajam na rodovia.

Exercício P6. Observando os resultados de cinco barcos numa corrida, vemos que existem dois valores para o tempo: tempo real (y) e tempo corrigido (x). Estamos supondo que o modelo ideal para relacionar os dois tempos, é aquele linear passando pela origem, $y_i = \beta_1 x_i + \epsilon_i$.

x_i	2,1	2,2	3,4	6,4	5,2
y_i	2,3	2,5	3,6	7,1	5,8

Tabela 5: Tempo real (y) e Tempo corrigido (x)

$\sum_{i=1}^{n} x_i = 19,2$	$\sum_{i=1}^{n} y_i = 21,3$	$\sum_{i=1}^{n} x_i y_i = 97,92$
$\sum_{i=1}^{n} x_i^2 = 88,38$	$\sum_{i=1}^{n} y_i^2 = 108,55$	$n = 5$

a) Calcule o estimador de mínimos quadrados (EMQ) para β_1 e ajuste o modelo de regressão.

Resposta:

O estimadores de mínimos quadrados (EMQ) para β_1 é dado por:

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum_{i=1}^{5} x_i y_i}{\sum_{i=1}^{5} x_i^2} = \frac{97,92}{88,38} = 1,108,$$

entretanto, o modelo ajustado será: $\hat{y}_i = \hat{\beta}_1 x_i = 1,108 x_i$.

b) Faça o gráfico de dispersão dos dados e represente graficamente o modelo ajustado no item anterior.

Resposta:

c) Qual é a interpretação que você daria para o parâmetro β_1?

Resposta: $\hat{\beta}_1 = 1,108$, representa o aumento esperado no tempo real para cada variação unitária do tempo corrigido.

d) Qual é o tempo real esperado para um barco cujo tempo corrigido foi igual a 5.0?

Resposta:

Para um tempo corrigido $x_0 = 5,0$ unidades, o tempo real esperado é de:
$\hat{y}_i|_{x_0=5,0} = 1,108 x_0 = 1,108 * 5,0 = 5,53972$.

e) Forneça um intervalo a 95% de confiança para a variância dos erros e conclua.

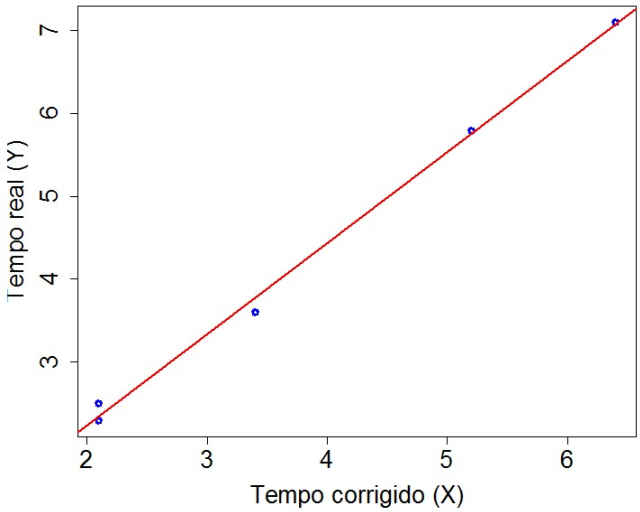

Figura 18: Diagrama de dispersão para os valores de x e y.

Resposta: Para obter o $IC_{95\%}$ para σ^2, vamos calcular sua estimativa pontual $\hat{\sigma}^2_\epsilon$ dada por:

$$\hat{\sigma}^2_\epsilon = \frac{SQE}{n-1} = \frac{\sum_{i=1}^{5} y_i^2 - \hat{\beta}_1^2 \sum_{i=1}^{5} x_i^2}{n-1} = \frac{108,55 - 1,108^2 * 88,38}{5-1} = 0,01505,$$

$$\begin{aligned} IC_{95\%}(\sigma^2) \quad : \quad & \frac{(n-1)\hat{\sigma}^2_\epsilon}{\chi^2_{(\alpha/2,n-1)}} \leq \sigma^2 \leq \frac{(n-1)\hat{\sigma}^2_\epsilon}{\chi^2_{(1-\alpha/2,n-1)}} \\ & \frac{(5-1)\hat{\sigma}^2_\epsilon}{\chi^2_{(0,025,4)}} \leq \sigma^2 \leq \frac{(5-1)\hat{\sigma}^2_\epsilon}{\chi^2_{(0,975,4)}} \\ & \frac{4 * 0,01505}{11,14329} \leq \sigma^2 \leq \frac{4 * 0,01505}{0,48442} \\ & 0,0054 \leq \sigma^2 \leq 0,12427. \end{aligned}$$

f) Forneça um intervalo a 95% de confiança para o tempo real médio de barcos cujo tempo corrigido é de 2.5.

Resposta:

Para um tempo corrigido $x_0 = 2,5$ unidades então, $\hat{\mu}_{y_i|x_0=2,5} = 1,108x_0 = 1,108 * 2,5 = 2,77$ unidades. O IC para a resposta média com variância

$$\sigma^2_{\hat{\mu}_{y_i|x_0}} = \hat{\sigma}^2_\epsilon \frac{x_0^2}{\sum_{i=1}^{5} x_i^2} = \frac{2,5^2 * 0,01505}{88,38} = 0,001064$$

será:

$$\begin{aligned} IC_{1-\alpha}(\mu_{y_i|x_0}) \quad : \quad & \hat{\mu}_{y_i|x_0} \pm t_{(1-\alpha/2,n-1)}\sigma_{\hat{\mu}_{y_i|x_0}} \\ & \hat{\mu}_{y_i|x_0} - t_{(1-\alpha/2,n-1)}\sigma_{\hat{\mu}_{y_i|x_0}} \leq \mu_{y_i|x_0} \leq \hat{\mu}_{y_i|x_0} + t_{(1-\alpha/2,n-1)}\sigma_{\hat{\mu}_{y_i|x_0}} \\ & 2,77 - 2,776\sqrt{0,001064} \leq \mu_{y_i|x_0} \leq 2,77 + 2,776\sqrt{0,001064} \\ & 2,6795 \leq \mu_{y_i|x_0} \leq 2,8806. \end{aligned}$$

g) Qual é a predição ($\hat{y}_0$) para o tempo real de um novo barco (y_0) cujo tempo corrigido é de 2,5.

Resposta:

Assuma que $\hat{\phi} = y_0 - \hat{y}_0$. Tal que,

$$\begin{aligned} \sigma^2_{\hat{\phi}} &= Var(y_0 - \hat{y}_0) = \hat{\sigma}^2_\epsilon + \hat{\sigma}^2_{y_i}, \quad \text{Cov}(y_0, \hat{y}_0) = 0 \\ &= 0,01505 + 0,001064 \\ &= 0,0161143. \end{aligned}$$

O intervalo de predição (IP) associado é:

$$\begin{aligned} IP_{1-\alpha}(y_0) \quad : \quad & \hat{y}_0 - t_{(1-\alpha/2,n-1)}\sigma_{\hat{\phi}} \leq y_0 \leq \hat{y}_0 + t_{(1-\alpha/2,n-1)}\sigma_{\hat{\phi}} \\ & 2,77 - 2,776 * \sqrt{0,0161143} \leq y_0 \leq 2,77 + 2.776 * \sqrt{0,0161143} \\ & 2,4176 \leq y_0 \leq 3,1224. \end{aligned}$$

h) Construa a tabela ANOVA e teste a significância da regressão para um nível de 5% de significância.

Resposta:

Formulação das hipóteses:

$$H_0 : \beta_1 = 0 \quad \text{vs} \quad H_1 : \beta_1 \neq 0$$

Estatística de teste:

$$F^* = \frac{QMReg}{QME} = \frac{\frac{SQReg}{1}}{\frac{SQE}{n-1}} \approx F_{(1,n-1)}$$

Regra de decisão:

$$\begin{aligned} &\text{Se} \quad F^* \leq F_{(1-\alpha,1,n-1)} \Rightarrow \text{não rejeita } H_0 \\ &\text{Se} \quad F^* > F_{(1-\alpha,1,n-1)} \Rightarrow \text{rejeita } H_0 \end{aligned}$$

Temos que:

Fontes de variação	gl	SQ	QM	F^*
Regressão	1	108.493	SQReg/1=108.493	7205.991
Erro	n-1=4	0.06022	SQE/4=0.01505	
Total	n=5	108.554		

Tabela 6: Tabela Anova: gl(graus de liberdade), SQ (soma de quadrados) e QM (quadrado médio).

$$SQReg = \hat{\beta}_1 \sum_{i=1}^{5} x_i y_i = 1.108 * 97.92 = 108.4898 \quad \text{e} \quad SQT = \sum_{i=1}^{5} y_i^2 = 108.55$$

$$SQE = SQT - SQReg = 108.55 - 108.4898$$

$$F^* = \frac{QMReg}{QME} = \frac{108.493}{0.01505} = 7205.991$$

Como $F^* > 7.7087 = F_{(0.95,1,4)}$ então, RH_0. E portanto, o modelo é significativo ao nível de 5% de significância.

Exercício P7. Na Tabela 7 a seguir estão os dados observados de 5 carros, em que y indica o número de acidentes sofridos por carros viajando à velocidade de x km/h.

x_i	80	90	100	110	120
y_i	79	83	90	95	99

Tabela 7: Número de acidentes (y) e velocidade (x)

$\sum_{i=1}^{n} x_i = 500$	$\sum_{i=1}^{n} y_i = 446$	$\sum_{i=1}^{n} x_i y_i = 45120$
$\sum_{i=1}^{n} x_i^2 = 51000$	$\sum_{i=1}^{n} y_i^2 = 40056$	$n = 5$

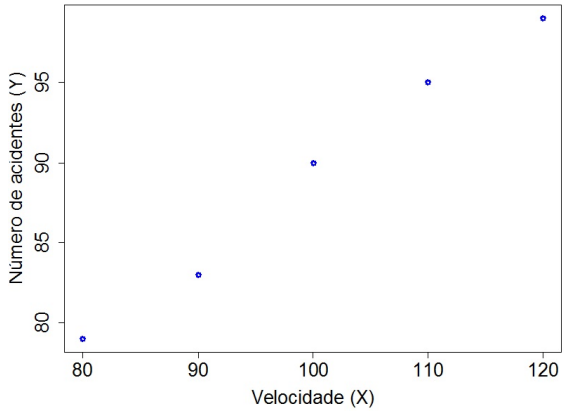

Figura 19: Diagrama de dispersão para os valores de x e y.

a) Faça o gráfico de dispersão para os dados, e averigue se um modelo linear sem intercepto é uma escolha razoável para modelar o comportamento de x e y.

Resposta:

b) Ajuste o modelo 1: $y_i = \beta_1 x_i + \epsilon_{1i}$ e interprete a estimativa de β_1.

O estimadores de mínimos quadrados (EMQ) para β_1 é dado por:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{5} x_i y_i}{\sum_{i=1}^{5} x_i^2} = \frac{45120}{51000} = 0,8847,$$

entretanto, o modelo ajustado é dado por: $\hat{y}_i = \hat{\beta}_1 x_i = 0,8847 x_i$.

$\hat{\beta}_1$ representa o aumento esperado no número de acidentes quando a velocidade aumenta em 1km/h.

c) Construa a respectiva tabela ANOVA e teste a significância do modelo para 10% de significância.

Resposta:

Formulação das hipóteses:

$$H_0 : \beta_1 = 0 \quad \text{vs} \quad H_1 : \beta_1 \neq 0$$

Estatística de teste:

$$F^* = \frac{QMReg}{QME} = \frac{\frac{SQReg}{1}}{\frac{SQE}{n-1}} \approx F_{(1,n-1)}$$

Temos que:

Fontes de variação	gl	SQ	QM	F^*
Regressão	1	39917,93	SQReg/1=39917,93	1156,455
Erro	n-1=4	138,07	SQE/4=34,5175	
Total	n=5	40056		

Tabela 8: Tabela Anova: gl(graus de liberdade), SQ (soma de quadrados) e QM (quadrado médio).

$$SQReg = \hat{\beta}_1 \sum_{i=1}^{5} x_i y_i = 0.8847 * 45120 = 39917.93,$$

$$SQT = \sum_{i=1}^{5} y_i^2 = 40056,$$

$$SQE = SQT - SQReg = 40056 - 39917,93 = 138,07,$$

$$F^* = \frac{QMReg}{QME} = \frac{39917,93}{34,5175} = 1156,455$$

Como $F^* > 4,545 = F_{(0,90;1,4)}$ então, RH_0. E portanto, o modelo é significativo ao nível de 10% de significância.

d) Conclua sobre a significância do modelo a 10% usando o teste t, construa a região crítica e conclua. É possível usarmos um IC para concluir sobre a significância do modelo? Se sim, construa-o, analise-o e use-o como ferramenta para tomada de decisão.

Resposta:

Com base nas hipóteses formuladas no item (c), a significância do modelo pode ser testada usando a o teste t cuja sua estatística é dada por:

$$t^* = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\sigma^2_{\hat{\beta}_1}}}$$

$$\begin{aligned} &\text{Se} \quad |t^*| \leq t_{(1-\alpha,n-1)} \Rightarrow NRH_0 \\ &\text{Se} \quad |t^*| > t_{(1-\alpha,n-1)} \Rightarrow RH_0 \end{aligned}$$

A região crítica será:

$$\text{RC} = \left\{t^* \in R : |t^*| > t_{(0,95;4)}\right\} = \left\{t^* \in R : |t^*| > 2,1318\right\}.$$

Onde $\hat{\sigma}^2_\epsilon = 34,5175$ e $\hat{\sigma}^2_{\hat{\beta}_1} = 34,5175/51000 = 0,0006768$ e portanto,

$$t^* = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\sigma^2_{\hat{\beta}_1}}} = \frac{0,8847 - 0}{\sqrt{0,0006768}} = 34,007.$$

Como $t^* \in$ RC então, há evidencias suficientes para afirmar que o modelo é significativo a 10% de significância.

Evidentemente podemos usar o IC para testar a significância do modelo verificando se o IC contém ou não o 0. Neste caso modelo não é significativo se $0 \in$ IC e significativo no caso contrário.

$$\begin{aligned}
IC_{90\%}(\beta_1) \quad &: \quad \hat{\beta}_1 \pm t_{(1-\alpha/2,n-1)}\sigma_{\hat{\beta}_1} \\
&\hat{\beta}_1 - t_{(0,95,4)}\sigma_{\hat{\beta}_1} \leq \beta_1 \leq \hat{\beta}_1 + t_{(0,95;4)}\sigma_{\hat{\beta}_1} \\
&0,8847 - 2,1318\sqrt{0,0006768} \leq \beta_1 \leq 0,8847 + 2,1318\sqrt{0,0006768} \\
&0,8847 \leq \beta_1 \leq 0,940181.
\end{aligned}$$

Entretanto, como o IC não contém o 0, podemos concluir a um nível de 10% de significância que o modelo é significativo ($\beta_1 \neq 0$).

e) Forneça a previsão para o número de acidentes sofridos por um carro que usualmente viaja a uma velocidade igual a 1 km/h. Use 5% de significância.

Resposta:

Observação: O valor fornecido no exercício ($x_0 = 1km/h$) esta fora do padrão dados, assim os IC e IP obtidos podem não ser informativos. Para efeito de ilustração, assuma (por exemplo) que $x_0 = 115km/h$, então,

$\hat{\mu}_{y_i|x_0=115} = 0,8847x_0 = 0,8847 * 115 = 101,7405$ unidades. O IC para a resposta média com variância

$$\sigma_{\hat{\mu}_{y_i|x_0}} = \hat{\sigma}^2_\epsilon \frac{x_0^2}{\sum_{i=1}^{5} x_i^2} = \frac{115^2 * 34,5175}{51000} = 8,9507$$

será:

$$
\begin{aligned}
IC_{1-\alpha}(\mu_{y_i|x_0}) \quad &: \quad \hat{\mu}_{y_i|x_0} \pm t_{(1-\alpha/2,n-1)}\sigma_{\hat{\mu}_{y_i|x_0}} \\
& \hat{\mu}_{y_i|x_0} - t_{(1-\alpha/2,n-1)}\sigma_{\hat{\mu}_{y_i|x_0}} \leq \mu_{y_i|x_0} \leq \hat{\mu}_{y_i|x_0} + t_{(1-\alpha/2,n-1)}\sigma_{\hat{\mu}_{y_i|x_0}} \\
& 101,7405 - 2,776\sqrt{8,9507} \leq \mu_{y_i|x_0} \leq 101,7405 + 2,776\sqrt{8,9507} \\
& 93,435 \leq \mu_{y_i|x_0} \leq 110,048.
\end{aligned}
$$

f) Estime o número médio de acidentes por carros que viajam a velocidade de 1 km/h. Use $\alpha = 5\%$. Qual é a diferença entre as inferências feitas em (e) e (f)?

Resposta:

Assuma que $\hat{\phi} = y_0 - \hat{y}_0$. Tal que,

$$
\begin{aligned}
\sigma^2_{\hat{\phi}} &= Var(y_0 - \hat{y}_0) = \hat{\sigma}^2_{\epsilon} + \hat{\sigma}^2_{\hat{y}_0}, \quad \text{Cov}(y_0, \hat{y}_0) = 0 \\
&= 34,5175 + 8,9507 \\
&= 43,4682.
\end{aligned}
$$

$$
\begin{aligned}
IP_{1-\alpha}(y_0) \quad &: \quad \hat{y}_0 - t_{(1-\alpha/2,n-1)}\sigma_{\hat{\phi}} \leq y_0 \leq \hat{y}_0 + t_{(1-\alpha/2,n-1)}\sigma_{\hat{\phi}} \\
& 101,7405 - 2,776\sqrt{43,4682} \leq y_0 \leq 101,7405 + 2,776\sqrt{43,4682} \\
& 83,4382 \leq y_0 \leq 120,0428.
\end{aligned}
$$

Comparando os dois intervalos, observamos que o intervalo de predição é sempre mais amplo que o intervalo de confiança para a resposta média. Essa diferença surge pelo fato do IP depender não só da variância do erro mas também, do erro associado a observação futura.

Exercício P8. Refaça o Exercício P7 ajustando agora o modelo 2: $y_i = \beta_0 + \beta_1 x_i + \epsilon_{2i}$.
Resposta: Os EMQ de β_0 e β_1 são dados por:

$$
\hat{\beta}_1 = \frac{\sum_{i=1}^{5}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{5}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{5} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{5} x_i^2 - n\bar{x}^2} = \frac{45120 - 5 * 89,2 * 100}{51000 - 5 * 100^2} = 0,520,
$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 89,2 - 0,520 * 100 = 37,200.$$

O modelo ajustado será: $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i = 37,200 + 0,520 x_i$. Cuja SQE é dada por:

$$SQE = \sum_{i=1}^{5} (y_i - \hat{y}_i)^2 = 2,40$$

a) Os resultados obtidos são comparáveis? Justifique.

Resposta: Sim os resultados pode ser comparados. Para tal, uma quantidade importante para comparar os modelos com e sem intercepto é a SQE. Como pode-se observar, a SQE é maior no modelo sem intercepto $SQE_{\epsilon_1} = 138,03$ que no modelo com intercepto, cuja $SQE_{\epsilon_2} = 2,400$, o que dá uma ideia de que o modelo com intercepto apresenta um bom ajuste em relação ao modelo sem intercepto.

b) Verifique se a omissão do intercepto no modelo pode ter influenciado na qualidade do ajuste do mesmo (dica: verifique se o intercepto é importante no modelo).

Resposta: Uma forma de complementar a conclusão do item anterior seria testar a importância do β_0 no modelo. Desta forma, considere as seguintes hipóteses:

$$H_0 : \beta_0 = 0 \quad \text{vs} \quad H_1 : \beta_0 \neq 0$$

Cuja estatística de teste é dada por:

$$t^* = \frac{\hat{\beta}_0 - \beta_0}{\sqrt{\sigma^2_{\hat{\beta}_0}}} \approx t_{n-2}$$

Região critica:

$$\text{RC} = \left\{ t^* \in R : |t^*| > t_{(1-\alpha/2, n-2)} \right\} = \{ t^* \in R : |t^*| > 3,183 \}$$

$$t^* = \frac{37,2 - 0}{\sqrt{\hat{\sigma}^2_\epsilon \left(1/5 + \frac{100^2}{(51000 - 5 * 100^2)} \right)}} = 13,0226.$$

Como podemos observar, o valor calculado pertence a região crítica, mostrando desta forma que o intercepto é importante no modelo ao nível de 5% de significância. Desta forma, concluímos mais uma vez que o modelo com intercepto se ajusta melhor aos dados, como pode-se observar a disposição dos pontos em torno das retas de regressão da Figura 20.

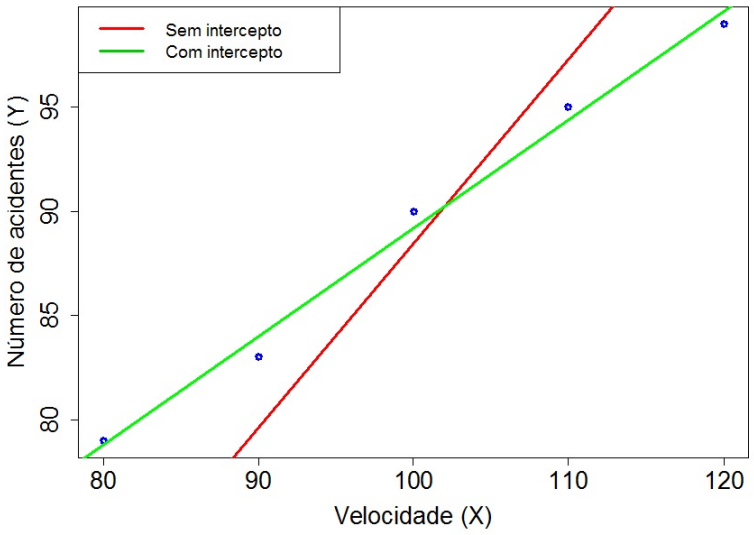

Figura 20: Ajuste dos modelos com e sem intercepto.

Exercício P9. Os dados da Tabela 9 foram gerados assumindo-se que $x_i \sim \mathrm{N}(1, 4)$ e $\epsilon_i \sim \mathrm{N}(0, 1)$ em que as variáveis respostas foram obtidas considerando-se os seguintes modelos:

Dados 1: $y_{1i} = 1 + 2,5x_i + \epsilon_i$

Dados 2: $y_{2i} = 0,1 + 2,5x_i + \epsilon_i$

Dados 3: $y_{3i} = 0,01 + 2,5x_i + \epsilon_i$

Dados 4: $y_{4i} = 2,5x_i + \epsilon_i$

Usando o R, ajuste os modelos lineares com e sem intercepto relacionando y_{ji} vs x_i, $j = 1, 2, 3, 4$. Isto é, para os quatro bancos de dado gerados e:

(a) Forneça os EMQ para os parâmetros do modelo.

(b) Forneça:

(b1) A variância de $\hat{\beta}_1$;

(b2) A variância dos erros;

(b3) O coeficiente de determinação;

i	x_i	y_{1i}	y_{2i}	y_{3i}	y_{4i}
1	2,45342122	7,6384254	6,7384254	6,6484254	6,6384254
2	3,13409457	9,2311122	8,3311122	8,2411122	8,2311122
3	-2,62917047	-4,1573884	-5,0573884	-5,1473884	-5,1573884
4	1,01759384	2,8216603	1,9216603	1,8316603	1,8216603
5	0,34600260	1,2466495	0,3466495	0,2566495	0,2466495
6	2,15625859	4,8280261	3,9280261	3,8380261	3,8280261
7	3,43908692	9,7256761	8,8256761	8,7356761	8,7256761
8	0,42029116	1,8937758	0,9937758	0,9037758	0,8937758
9	-1,55606391	-4,4054960	-5,3054960	-5,3954960	-5,4054960
10	0,02624818	2,2272221	1,3272221	1,2372221	1,2272221
11	1,33304706	3,2709542	2,3709542	2,2809542	2,2709542
12	-2,61474916	-4,4843024	-5,3843024	-5,4743024	-5,4843024
13	-0,91861567	-2,3868563	-3,2868563	-3,3768563	-3,3868563
14	3,14065688	7,9012058	7,0012058	6,9112058	6,9012058
15	1,98149483	6,1426313	5,2426313	5,1526313	5,1426313
16	1,36726578	3,1112517	2,2112517	2,1212517	2,1112517
17	-0,06207828	-0,2481617	-1,1481617	-1,2381617	-1,2481617
18	2,48490842	8,3796986	7,4796986	7,3896986	7,3796986
19	-3,55904155	-6,6994742	-7,5994742	-7,6894742	-7,6994742
20	1,36259588	2,4438902	1,5438902	1,4538902	1,4438902

Tabela 9: Dados simulados

(c) Teste a significância dos coeficientes de regressão.

Resposta:

Formulação das hipóteses:

$$H_0 : \beta_{1i} = 0 \quad \text{vs} \quad H_1 : \beta_{1i} \neq 0 \quad \text{com i=1,2,3,4}$$

$$H_0 : \beta_{0i} = 0 \quad \text{vs} \quad H_1 : \beta_{0i} \neq 0 \quad \text{com i=1,2,3,4}$$

Modelo sem intercepto: $\text{RC} = \left\{t^* \in R : |t^*| > t_{(0,975;19)}\right\} = \left\{t^* \in R : |t^*| > 2,09302\right\}$.

Modelo com intercepto: $\text{RC} = \left\{t^* \in R : |t^*| > t_{(0,975;18)}\right\} = \left\{t^* \in R : |t^*| > 2,1009\right\}$.

Em geral, β_{1i} é significativamente diferente de zero tanto no modelo com e sem intercepto, para $i = 1, 2, 3$ e 4. No caso do modelo com intercepto, observamos que, com exceção do **Modelo 1**, nos restantes três modelos o intercepto não é importante.

	Modelo sem intercepto					Modelo com intercepto						
	$\hat{\beta}_1$	$\hat{\sigma}^2_{\hat{\beta}_1}$	$\hat{\sigma}^2_{\epsilon}$	$\%R^2$	$\lvert t^*_{\beta_1}\rvert$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\sigma}^2_{\hat{\beta}_1}$	$\hat{\sigma}^2_{\epsilon}$	$\%R^2$	$\lvert t^*_{\beta_0}\rvert$	$\lvert t^*_{\beta_1}\rvert$
Modelo_1	2,49	0,0199	**1,77**	94,23	17,61	0,85	2,36	0,0144	**1,15**	95,57	**3,37**	19,71
Modelo_2	2,35	0,0123	1,09	95,96	21,25	-0,05	2,36	0,01	1,15	95,57	0,19	-
Modelo_3	2,34	0,013	1,15	95,86	20,98	-0,14	2,36	0,05	1,15	95,57	0,550	-
Modelo_4	2,34	0,013	1,11	95,85	20,94	-0,15	2,37	0,01	1,15	95,57	0,59	-

Tabela 10: Comparando estimativas dos parâmetros em modelos com e sem intercepto.

(d) Para cada situação qual modelo escolheria, modelo com ou sem intercepto? Justifique a resposta.

Resposta:

Modelo 1: O intercepto é importante, isto é, $\beta_0 \neq 0$. O erro quadrático médio é menor no modelo com intercepto, apresentando igualmente um R^2 maior. Assim, o modelo com intercepto apresenta um bom ajuste que no caso sem intercepto.

Nos Modelo 2,3 e 4: O intercepto não é importante. Entretanto, o erro quadrático médio é menor no modelo sem intercepto que no outro caso. De igual forma, o modelo sem intercepto apresenta um R^2 maior. E como conclusão, o modelo sem intercepto apresenta o melhor ajuste.

Exercício P10. Um laboratório está interessado em medir o efeito da temperatura sobre a potência de um antibiótico. Dez amostras de 50 gramas cada foram guardadas a diferentes temperaturas e após 15 dias mediu-se a potência. Os resultados são mostrados na Tabela 11.

Temperatura	30°		50°			70°			90°	
Potência	38	43	32	26	33	19	27	23	14	21

Tabela 11: Temperatura e potência

Considere $\alpha = 5\%$. Sempre que necessário, enuncie as hipóteses, regra de decisão e apresente as conclusões.

a) Com base nos dados da Tabela 11, é correto afirmar que o intercepto é importante no modelo?

b) Pode-se afirmar que o fator temperatura sobre a potência é significativo? Qual é a estatística de teste apropriada e qual é a sua distribuição? Apresente uma justificativa para a distribuição desta estatística de teste.

c) Supondo que o efeito da temperatura sobre a potência seja significativo para o modelo (item (b)), pode-se afirmar que a temperatura influencia positivamente na potência dos antibióticos?

d) Obtenha uma estimativa intervalar para o intercepto, o coeficiente angular da reta de regressão e σ^2. Interprete-os.

e) São os resultados de (a), (c) e (d) consistentes quanto a significância dos parâmetros? Justifique a resposta.

f) Forneça uma previsão para a potência assumindo uma temperatura de 60^oC. Obtenha um intervalo com $(1-\alpha)100\%$ de confiança para a resposta média e para a resposta individual assumindo o mesmo valor para a temperatura. Explique a diferença entre estes dois intervalos.

g) Produza a tabela ANOVA para este modelo. Usando o teste F determine se existe ou não associação linear entre a temperatura e a potência.

h) Para os itens (a), (b) e (g) calcule o valor-p e decida.

i) Com base nos resultados da tabela ANOVA, que porcentagem da variabilidade na variável resposta é atribuída a causas aleatórias?

Exercício P11. Um pesquisador de marketing estudou as vendas anuais de um produto que havia sido introduzido a 10 anos. Na tabela a seguir apresentamos os anos (codificados) e o número de vendas em milhares de unidades.

i	1	2	3	4	5	6	7	8	9	10
X_i	0	1	2	3	4	5	6	7	8	9
Y_i	98	135	162	178	221	232	283	300	374	395

Tabela 12: Número de vendas (Y) e anos (X)

a) É correto afirmar que o modelo sem intercepto se ajusta melhor aos dados?

b) Podemos considerar que os anos influenciam significativamente nas vendas do produto?

c) Supondo que o efeito dos anos sobre o número de vendas seja significativo (item (b)), é correto afirmar que o número de vendas diminui em função dos anos?

d) Obtenha uma estimativa intervalar a $(1-\alpha)100\%$ de confiança para o intercepto, o coeficiente angular da reta de regressão e σ^2. Interprete-os.

e) São os resultados de (a), (c) e (d) consistentes quanto a significância dos parâmetro? Justifique a resposta.

f) Forneça uma previsão para a variável resposta assumindo que X_i=7.5. Obtenha um intervalo com $(1-\alpha)100\%$ de confiança para a resposta média e para a resposta individual assumindo este valor para a variável preditora. Explique a diferença entre estes dois intervalos.

g) Produza a tabela ANOVA para este modelo. Usando o teste F determine se existe ou não associação linear entre os anos e o número de vendas.

h) Com base nos resultados da tabela ANOVA, qual é a porcentagem da variabilidade na da variável resposta que é explicada pela variável preditora?

Exercício P12. Com objetivo de avaliar a qualidade dos estimadores de mínimos quadrados no modelo de regressão linear, faça um estudo de simulação Monte Carlo nos moldes discutidos a seguir. Assumindo um modelo de regressão linear simples, considere um total de M=1000 réplicas para o estudo Monte Carlo e, avalie o comportamento das seguintes quantidades: Erro Quadrático Médio (EQM), o viés e a variância dos estimadores de β_0, β_1 e σ^2 para $n = \{10, 20, 50, 100, 500, 1000\}$. Para o processo de geração dos dados, assuma que tanto a variável preditora como o termo de erro são provenientes de uma distribuição normal padrão. Assuma também que $\beta_0 = 2$, $\beta_1 = -2$. A partir dos resultados de simulação:

a) Apresente uma tabela contendo as quantidades que estamos interessados em avaliar.

b) O que se pode dizer quanto as estimativas dos parâmetros a medida em que o tamanho de amostra aumenta.

c) O que acontece com a variância das estimativas a medida em que o tamanho de amostra aumenta.

d) Apresente os gráficos de *box-plot* e os histogramas das estimativas de cada parâmetro.

e) Apresente um gráfico que relacione o tamanho de amostra e os valores do EQM, viés e a variância dos estimadores de β_0, β_1 e σ^2.

f) Pode-se afirmar que as estimativas menos viesadas (com viés próximo do zero) apresentam um EQM aproximadamente igual a variância das estimativas? Justifique a resposta.

3.2 Exercícios Teóricos

Exercício T1. Em cada item identifique se o modelo é linear ou não-linear:

a) $y_i = \beta_0 + \beta_1(x_i - \bar{x}) + \epsilon$.

b) $y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \epsilon$.

c) $y_i = \beta_0 + e^{\beta_1} x_i + \epsilon$.

d) $y_i = \beta_0 + \frac{1}{\beta_1} x_i + \epsilon$.

e) $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon$.

f) $y_i = \beta_0 + \beta_1 \frac{1}{x_i} + \epsilon$.

Exercício T2. Os estimadores de mínimos quadrados foram obtidos minimizando a função $S = \sum_{i=1}^{n} e_i^2$. Por que não minimizamos simplesmente $S^* = \sum_{i=1}^{n} e_i$? Dica: encontre os estimadores se utilizássemos S^*.

Exercício T3. Considere o modelo de regressão linear simples $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$ em que $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$. Demonstre os resultados das variâncias dos estimadores de mínimos quadrados:

a) $Var(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{\sigma^2}{S_{xx}}$.

b) $Var(\hat{\beta}_0) = \frac{\sigma^2}{n} + \frac{\sigma^2 \bar{X}^2}{S_{xx}}$.

Exercício T4. Suponha que estamos interessados em ajustar um modelo de regressão linear simples $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$ em que o intercepto β_0 seja conhecido.

a) Encontre os estimadores de mínimos quadrados para β_1.

b) Qual é a variância do estimador encontrado em (a)?

Exercício T5. Considere o modelo de regressão linear simples $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$ em que $E(\epsilon_i) = 0$, $Var(\epsilon_i) = \sigma_\epsilon^2$ e $Cov(\epsilon_i, \epsilon_j) = 0, \ \forall \ i \neq j$. Prove que:

a) $E(\bar{Y}) = \beta_0 + \beta_1 \bar{X}$ e $Var(\bar{Y}) = \sigma_\epsilon^2 / n$.

b) $E(SS_T) = (n-1)\sigma_\epsilon^2 + \beta_1^2 S_{XX}$.

Exercício T6. Suponha que estamos interessados em ajustar o modelo de regressão linear simples $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$ porém decidiu-se reescalonar a variável regressora tomando-a como o desvio com relação a sua média, isto é, $X_i^* = X_i - \bar{X}, \quad i = 1, ..., n$.

a) Prove que o modelo que relaciona Y_i e a nova variável regressora X_i^* é $Y_i = \beta_0^* + \beta_1 X_i + \epsilon_i$, onde $\beta_0^* = \beta_0 + \beta_1 \bar{X}$.

b) Prove que o EMQ para β_0^* é igual a $\bar{Y}$ o que o estimador de β_1 permanece igual ao do modelo original.

c) Prove que ambos os modelos fornecem os mesmos valores preditos para Y.

Exercício T7. Seja o modelo $y_i = \beta_1 x_i + \epsilon_i$ e assuma que $E(\epsilon_i) = 0$, $\text{Var}(\epsilon_i) = \sigma^2_\epsilon$ e, $\text{Cov}(\epsilon_i, \epsilon_j) = 0$, para todo $i \neq j$.

(a) Prove que o EMQ para β_1 é dado por: $\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$.

(b) Verifique se $\hat{\beta}_1$ é um estimador não-viesado para β_1.

(c) Prove que $\text{Var}(\hat{\beta}_1) = \frac{\sigma^2_\epsilon}{\sum_{i=1}^{n} x_i^2}$.

(d) Prove que $Var(\hat{y}_i) = \frac{x_i^2}{\sum_{i=1}^{n} x_i^2}\sigma^2_\epsilon$, em que $\hat{y}_i$ é o valor predito pelo modelo.

(e) Verifique se $SS_\epsilon = \sum_{i=1}^{n} y_i^2 - \hat{\beta}_1^{\,2} \sum_{i=1}^{n} x_i^2$ é um estimador viesado para σ^2_ϵ. (dica: *use o fato que* $Var(Z) = E(Z^2) - E^2(Z)$)

Agradecimentos

Agradecemos à Pró-reitoria de Graduação da UFMG (PROGRAD) pela complementação de bolsa concedida ao aluno bolsista durante a vigência do projeto PIFD2017-64. Agradecimentos também a Frederico Machado Almeida, aluno de doutorado em Estatística da UFMG, pelo auxílio na seção de exercícios.

Referências

Landeiro, V. L. (2013) *Introdução ao uso do programa R*, Instituto de Biologia, Departamento de Botânica e Ecologia, Universidade Federal de Mato Grosso.

Reis, E. A.; Amaral, G. D. e Silva, V. L. (2009) *Análise de Regressão Linear no Pacote R*, Relatório Técnico - Série Ensino - RTE-001/2009, Departamento de Estatística da UFMG.

Ribeiro, A. J. F.; Ferreira, E. F.; Reis, I. A. e Montenegro, L. C. C. (2012) *Bioestatística básica usando o ambiente computacional*, Relatório Técnico - Série Ensino - RTE-01/2012, Departamento de Estatística da UFMG.

Apêndices

A Comandos em R para avaliar o efeito do tamanho n da amostra na qualidade das estimativas de MQ - Seção 2.3.2

```
n <- c(10,20,50,100,500,1000)
M <- 100
beta0 <- 1
beta1 <- -1
sigma2 <- 1

media.estimativas = eqm = vies = var.estimadores <- matrix(,length(n),3)
nomes.linhas <- numeric(length(n))
for(i in 1:length(n)){
saida <- MC.regressao(n[i], M, beta0, beta1, sigma2)
media.estimativas[i,] <- saida[1,]
eqm[i,] <- saida[2,]
vies[i,] <- saida[3,]
var.estimadores[i,] <- saida[4,]

nomes.linhas[i] <- paste("n =",n[i])
}

nomes.colunas = c(paste("beta0 =",beta0),paste("beta1 =",beta1),
                  paste("sigma2 =",sigma2))

rownames(media.estimativas) <- nomes.linhas
colnames(media.estimativas) <- nomes.colunas
print(media.estimativas)

rownames(eqm) <- nomes.linhas
colnames(eqm) <- nomes.colunas
print(eqm)

rownames(vies) <- nomes.linhas
colnames(vies) <- nomes.colunas
print(vies)

rownames(var.estimadores) <- nomes.linhas
colnames(var.estimadores) <- nomes.colunas
print(var.estimadores)
```

```
windows()
par(mfrow=c(1,3))
for(i in 1:3) plot(n,media.estimativas[,i],main=nomes.colunas[i],
ylab="media.estimativas")
savePlot(filename = "n_vs_media_estimativas",type="pdf")
dev.off()

windows()
par(mfrow=c(1,3))
for(i in 1:3) plot(n,eqm[,i],main=nomes.colunas[i],ylab="eqm")
savePlot(filename = "n_vs_eqm",type="pdf")
dev.off()

windows()
par(mfrow=c(1,3))
for(i in 1:3) plot(n,vies[,i],main=nomes.colunas[i],ylab="vies")
savePlot(filename = "n_vs_vies",type="pdf")
dev.off()

windows()
par(mfrow=c(1,3))
for(i in 1:3) plot(n,var.estimadores[,i],main=nomes.colunas[i],
ylab="var.estimadores")
savePlot(filename = "n_vs_var_estimadores",type="pdf")
dev.off()
```

B Comandos em R para avaliar o efeito da variância do erro na qualidade das estimativas de MQ - Seção 2.3.3

```
n <- 100
M <- 100
beta0 <- 1
beta1 <- -1
sigma2 <- c(1,10,100)

media.estimativas = eqm = vies = var.estimadores <- matrix(,length(sigma2),3)
nomes.linhas <- numeric(length(sigma2))
for(i in 1:length(sigma2)){
saida <- MC.regressao(n, M, beta0, beta1, sigma2[i])
media.estimativas[i,] <- saida[1,]
eqm[i,] <- saida[2,]
vies[i,] <- saida[3,]
```

```
var.estimadores[i,] <- saida[4,]

nomes.linhas[i] <- paste("sigma1=",sigma2[i])
}

nomes.colunas = c(paste("beta0 =",beta0),paste("beta1 =",beta1),
                  paste("sigma2 =",sigma2))

rownames(media.estimativas) <- nomes.linhas
colnames(media.estimativas) <- nomes.colunas
print(media.estimativas)

rownames(eqm) <- nomes.linhas
colnames(eqm) <- nomes.colunas
print(eqm)

rownames(vies) <- nomes.linhas
colnames(vies) <- nomes.colunas
print(vies)

rownames(var.estimadores) <- nomes.linhas
colnames(var.estimadores) <- nomes.colunas
print(var.estimadores)

windows()
par(mfrow=c(1,3))
for(i in 1:3) plot(sigma2,media.estimativas[,i],main=nomes.colunas[i],
ylab="media.estimativas")
savePlot(filename = "sigma2_vs_media_estimativas",type="pdf")
dev.off()

windows()
par(mfrow=c(1,3))
for(i in 1:3) plot(sigma2,eqm[,i],main=nomes.colunas[i],ylab="eqm")
savePlot(filename = "sigma2_vs_eqm",type="pdf")
dev.off()

windows()
par(mfrow=c(1,3))
for(i in 1:3) plot(sigma2,vies[,i],main=nomes.colunas[i],ylab="vies")
savePlot(filename = "sigma2_vs_vies",type="pdf")
dev.off()

windows()
par(mfrow=c(1,3))
```

```
for(i in 1:3) plot(sigma2,var.estimadores[,i],main=nomes.colunas[i],
ylab="var.estimadores")
savePlot(filename = "sigma2_vs_var_estimadores",type="pdf")
dev.off()
```