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

Análise de Regressão Linear no Pacote R

Gabriela Domingues do Amaral
Vanessa Loureiro Silva
Edna Afonso Reis

RELATÓRIO TÉCNICO
SÉRIE ENSINO
RTE 001/2009

Análise de Regressão Linear no Pacote R

Gabriela Domingues do Amaral*
Vanessa Loureiro Silva*
Edna Afonso Reis**

* Graduandas em Estatística - UFMG
* Profa. Adjunta do Departamento de Estatística - UFMG

Sumário

1 Introdução

Análise de regressão é uma técnica estatística utilizada para investigar *a relação existente entre variáveis* através da construção de uma equação (*um modelo*). De maneira geral, 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.

A grande utilização da análise de regressão em diversas áreas, aliada à disseminação do software estatístico gratuito *R*, motivaram a criação desta apostila, que tem o objetivo de apresentar, por meio de exemplos, as principais tarefas implementadas no software *R* sobre regressão. Os recursos computacionais serão apresentados supondo que o usuário já possui conhecimentos teóricos suficientes de análise de regressão.

Iniciemos com um exemplo. Um investigador deseja estudar a possível relação entre o salário (em anos completos) e o tempo de experiência (em mil reais) no cargo de gerente de agências bancárias de uma grande empresa. Os dados coletados são mostrados abaixo:

Tabela 1: Salário e tempo de experiência dos gerentes da agência bancária em estudo

Salário	Experiência
1.9307	0
3.1769	17
2.2769	8
3.1307	15
2.7769	9
3.0923	15
2.6538	8
2.2230	5
2.8538	13
3.2307	20
2.8230	11
1.9076	1
2.5384	6
2.5692	7
4.2230	23
4.0923	20
3.6000	18
4.7076	27
3.1461	11
2.9923	10
4.7461	29
4.1153	23
2.3615	4
4.0923	22
4.5076	25
2.9076	9
4.4846	25

Note que são considerados 27 pares de observações correspondentes à variável resposta Salário e à variável explicativa Experiência, para cada um dos gerentes da empresa.

1.1 Leitura dos Dados

Inicialmente, os dados devem ser organizados como *objetos de dados R*, nesse caso como um *data frame* (planilha). Para isso, é necessário que a tabela acima se encontre numa estrutura tabular, na qual as colunas representam as variáveis e as linhas representam os indivíduos. Nesses termos, seja o arquivo de texto *gerentes.txt*, utiliza-se a função `read.table` para que o arquivo seja lido pelo *R*:

```
> gerentes = read.table("gerentes.txt",header=T)
> attach(gerentes)
> gerentes
   Experiencia Salario
1            0  1.9307
2           17  3.1769
3            5  2.2769
4           15  3.1307
.
.
.
```

Observe que o argumento `header=T` indica que a primeira linha do arquivo contém os rótulos da planilha e que a função `attach` anexa o objeto *gerentes* no caminho de procura do software.

2 Análise Exploratória

2.1 Estatísticas Descritivas

Uma maneira fácil de obter algumas estatísticas descritivas das variáveis em estudo é através do comando `summary()`, que retorna as estatísticas *mínimo*, *quartis*, *média* e *máximo*. Para medir a variabilidade, utilize as funções `var()` e `sd()` para obter a *variância* e o *desvio padrão*.

```
> summary(Salario)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.908   2.611   3.092   3.228   4.092   4.746
> var(Salario)
[1] 0.7366968
> sd(Salario)
[1] 0.8583104
```

2.2 Diagrama de Dispersão

Para verificar a existência de alguma relação entre *Salário* e *Experiência*, deve-se construir um *Diagrama de Dispersão* para as duas variáveis:

```
> plot(Experiencia,Salario)
```

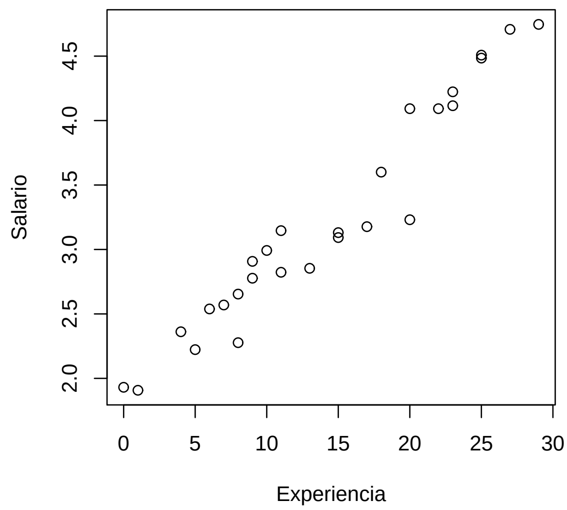


Figura 1: Diagrama de Dispersão de Salário versus Experiência

2.3 Correlação Linear

Para calcular o Coeficiente de Correlação Linear de Pearson entre as variáveis, utilize a função `cor`:

```
> cor(Experiencia,Salario)
[1] 0.9735413
```

Observe que o *R* retornou o valor `0.9735413` o que evidencia uma forte relação linear entre as variáveis em estudo. Para avaliar se esse resultado é significativo, pode-se realizar um *Teste de Hipóteses* para a o Coeficiente de Correlação (supondo que as suposições do teste sejam satisfeitas):

```
> cor.test(Experiencia,Salario)

        Pearson's product-moment correlation

data:  Experiencia and Salario
t = 21.3018, df = 25, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.942046 0.988026
sample estimates:
      cor
0.9735413
```

Como o *Valor P* do teste (`p-value < 2.2e-16`) é bem pequeno, conclui-se que o valor do Coeficiente de Correlação Linear de Pearson tem significância Estatística.

3 Regressão Linear Simples

3.1 Ajuste do Modelo Linear

Sejam X e Y, respectivamente, as variáveis *Experiência* (explicativa) e *Salário* (resposta). Propõe-se um modelo de regressão linear de primeira ordem, dado pela equação: $Y = \beta_0 + \beta_1 X + \epsilon$, onde β_0 e β_1 são parâmetros desconhecidos e ϵ é o erro aleatório.

Para ajustar um modelo de regressão linear no R utiliza-se a função `lm`:

```
> ajuste=lm(Salario ~ Experiencia)
>
> ajuste

Call:
lm(formula = Salario ~ Experiencia)

Coefficients:
(Intercept)  Experiencia
    1.83070      0.09982
```

Note que função `lm()` é chamada com o formato lm(y ~ x), ou seja, a variável resposta é y e a preditora é x, sempre nessa ordem.

O R retorna o valor dos coeficientes de $\hat{\beta}_0$ e $\hat{\beta}_1$ estimados via Método de Mínimos Quadrados. Logo, a equação da reta ajustada é dada por $\hat{Y} = 1,83 + 0,0998X_i$.
Com a função `summary`, diversas medidas descritivas úteis para a análise do ajuste podem ser obtidas:

```
> summary(ajuste)

Call:
lm(formula = Salario ~ Experiencia)

Residuals:
     Min       1Q   Median       3Q      Max
-0.59637 -0.07929  0.03977  0.14499  0.26523

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.830696   0.076062   24.07   <2e-16
Experiencia 0.099819   0.004686   21.30   <2e-16
---
Residual standard error: 0.2 on 25 degrees of freedom
Multiple R-squared: 0.9478,     Adjusted R-squared: 0.9457
F-statistic: 453.8 on 1 and 25 DF,  p-value: < 2.2e-16
```

Da execução desse comando, pode-se obter, por exemplo, os erros-padrão (`Std. Error`) das estimativas dos coeficientes de regressão: EP($\hat{\beta}_0$) = 0,0761 e EP($\hat{\beta}_1$) = 0,0047. Além disso, obtém-se o valor do

Coeficiente de Determinação (`Multiple R-Squared`), $R^2 = 0,9478$.

Com a função `anova`, pode-se construir a Tabela da Análise de Variância:

```
> anova(ajuste)
Analysis of Variance Table

Response: Salario
            Df  Sum Sq Mean Sq F value    Pr(>F)
Experiencia  1 18.1539 18.1539  453.77 < 2.2e-16
Residuals   25  1.0002  0.0400
---
```

Da tabela ANOVA, obtém-se o Quadrado Médio (`Mean Sq`) Residual, que é uma estimativa para a variância dos erros (σ^2), ou seja, $s^2 = 0,04$.

Para esboçar a reta ajustada no diagrama de dispersão, utilize a função abline:

```
> windows()
> plot(Experiencia,Salario)
> abline(lm(Salario ~ Experiencia))
```

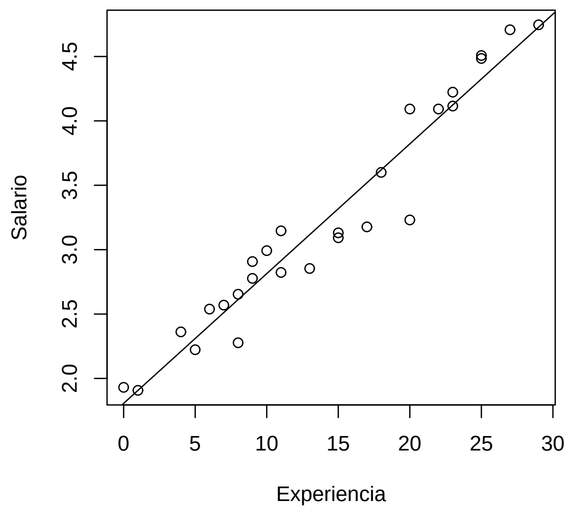


Figura 2: Diagrama de Dispersão de Salário versus Experiência com a reta ajustada

3.2 Intervalos de Confiança para β_0 e β_1

Para construir os Intervalos de Confiança (95%) para os coeficientes da regressão, utiliza-se o seguinte comando:

```
> confint(ajuste)
                 2.5 %    97.5 %
```

```
(Intercept) 1.67404277 1.9873489
Experiencia 0.09016774 0.1094694
```

3.3 Testes de Hipóteses

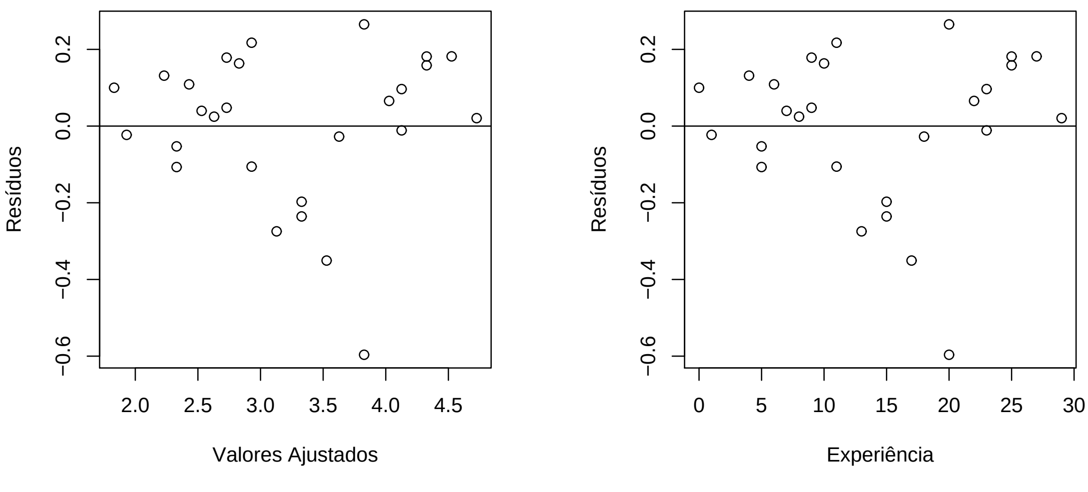
Para proceder o *Teste F da Significâcnia da Regressão* e os *Testes t individuais*, verifique o *Valor P* para cada caso através da saída da função `summary`:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.830696   0.076062   24.07   <2e-16
Experiencia 0.099819   0.004686   21.30   <2e-16
---
(...)
F-statistic: 453.8 on 1 and 25 DF,  p-value: < 2.2e-16
```

3.4 Análise dos Resíduos

Para avaliar as suposições de que os erros possuem variância constante e são não correlacionados entre si, construa os gráficos de "Resíduos versus Valores Ajustados da Variável Resposta" e "Resíduos versus Valores da Variável Explicativa":

```
> windows()
> plot(fitted(ajuste),residuals(ajuste),xlab="Valores Ajustados",ylab="Resíduos")
> abline(h=0)
> windows()
> plot(Experiencia,residuals(ajuste),xlab="Experiência",ylab="Resíduos")
> abline(h=0)
```



(a) Resíduos vs Valores Ajustados

(b) Resíduos vs Variável Explicativa

Figura 3: Gráficos para Análise dos Resíduos

Para exibir os Valores Ajustados e os Resíduos do ajuste, digite os comandos:

```
> ajuste$residuals
> ajuste$fitted.values
```

Na Figura 3, observa-se a violação da suposição de homocedasticidade dos erros. Para corroborar esse resultado, pode-se dividir o conjunto de dados em duas partes, utilizando a *mediana* por exemplo, e realizar um teste para comparar as variâncias de cada subconjunto:

```
> median(Experiencia)
[1] 13

> var.test(residuals(ajuste)[gerentes$Experiencia>13],residuals(ajuste)
[gerentes$Experiencia<13])
F test to compare two variances
F = 5.4334, num df = 12, denom df = 12, p-value = 0.006408
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
  1.657891 17.806654
sample estimates:
ratio of variances
          5.433369
```

Observe que o *Valor P* do teste (`p-value = 0.006408`) é menor que os níveis de significância mais usuais (0,01; 0,05; 0,10). Portanto, conclui-se que a variância dos dois subconjuntos não é igual, o que implica a heterocedasticidade dos erros.

Outra maneira de avaliar a heterocedasticidade dos erros é realizar algum teste de homocedasticidade. Na biblioteca *lmtest* do *R*, a função `bptest` realiza o teste de Breusch-Pagan. Ressalta-se, entretanto, que tal teste não é muito poderoso e pode levar à

Para avaliar a suposição de normalidade dos erros, deve-se construir o gráfico da "Probabilidade Normal dos Resíduos":

```
> windows()
> qqnorm(residuals(ajuste), ylab="Resíduos",xlab="Quantis teóricos",main="")
> qqline(residuals(ajuste))
```

Pela Figura 4 (página seguinte), observa-se a violação da suposição de que os erros aleatórios têm distribuição Normal. Considere, também o *Teste de Normalidade de Shapiro Wilk*:

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

        Shapiro-Wilk normality test

data:  residuals(ajuste)
W = 0.9025, p-value = 0.01531
```

Portanto, como o *Valor P* do teste é pequeno, rejeita-se a hipótese de normalidade dos resíduos e, por consequência, conclui-se que os erros não são normalmente distribuídos.

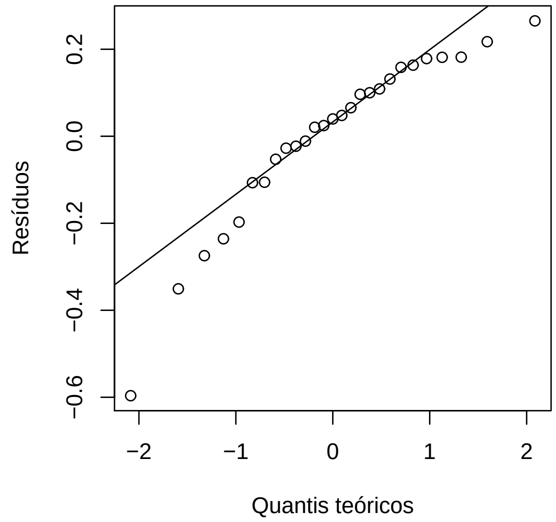


Figura 4: Gráfico de Probabilidade Normal dos Resíduos

3.5 Intervalos de Confiança para Resposta Média e Individual

Dado um novo conjunto de preditoras, $X = X_0$, a fim de fazer inferência sobre os valores preditos das respostas média e individual de Y, utilize a função `predict()`. Essa função requer que o segundo argumento seja um data frame com as covariáveis nomeadas do mesmo modo que o banco de dados original:

```
> x0 = data.frame(Experiencia=3)
```

Dado $X_0 = 3$, as correspondentes estimativas pontual e intervalar (95% de confiança) para a resposta média são obtidas por meio do comando:

```
>  predict(ajuste,x0,interval="confidence")
        fit      lwr      upr
  2.130151 1.997657 2.262646
```

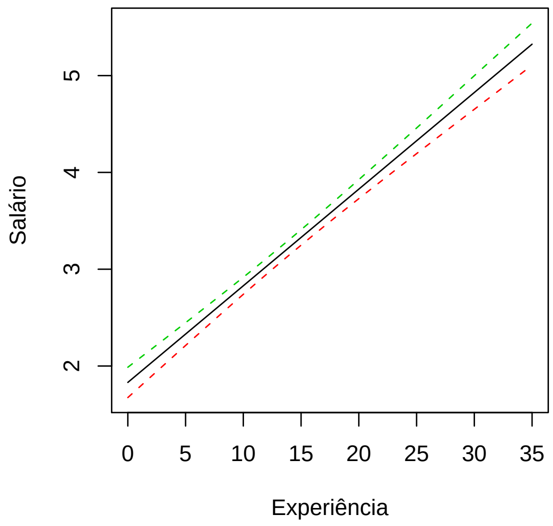
Já para a resposta individual, dado $X_0 = 3$, as correspondentes estimativas pontual e intervalar (95% de confiança) para são obtidas por meio do comando:

```
> predict(ajuste,x0,interval="prediction")
        fit      lwr      upr
  2.130151 1.697425 2.562878
```

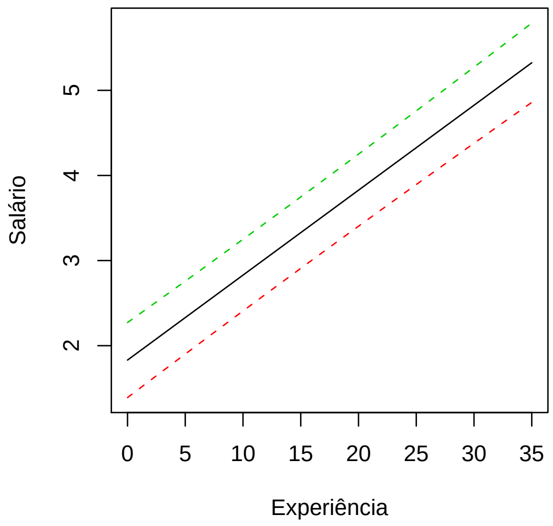
Se X_0 é um vetor, por exemplo, $X_0 = (0, 1, ..., 34, 35)$, pode-se contruir gráficos com as estimativas pontuais e intervalares das respostas médias e individuais, por meio dos respectivos comandos:

```
> windows()
> par(mfrow=c(1,2))
> x0 = data.frame(Experiencia=seq(0,35,1))
> p1 = predict(ajuste,x0,interval="confidence",se=T)
> matplot(x0,p1$fit,lty=c(1,2,2),type="l",xlab="Experiência",ylab="Salário")
```

```
> p2 = predict(ajuste,x0,interval="prediction",se=T)
> matplot(x0,p2$fit,lty=c(1,2,2),type="l",xlab="Experiência",ylab="Salário")
```



(a) Respostas médias



(b) Respostas individuais

Figura 5: Estimativas pontuais e Intervalos de Confiança de 95% para as respostas médias e individuais

Destaca-se, no entanto, que os intervalos de confiança aqui construídos foram feitos apenas em caráter ilustrativo, uma vez que, como as suposições para este modelo não foram satisfeitas, qualquer inferência feita a partir dele é inválida.

4 Regressão Linear Simples - Valores Replicados da Variável Explicativa

Em um estudo sobre a relação entre a massa muscular de uma pessoa e sua idade, foram coletadas informações sobre estas duas variáveis em 26 indivíduos.

Tabela 2: Massa Muscular (em unidades de massa muscular) e Idade (em anos) de 26 indivíduos

Indivíduo	Massa	Idade
1	100	43
2	95	43
3	116	45
4	97	45
5	100	45
6	95	16
7	105	49
8	100	53
9	105	53
10	97	53
11	87	56
12	80	56
13	76	58
14	91	64
15	84	65
16	90	65
17	80	65
18	68	67
19	78	68
20	82	71
21	76	71
22	84	71
23	73	73
24	65	76
25	68	76
26	77	78

4.1 Ajuste do Modelo Linear

Como na Seção 01, a Tabela 2 deve ser convertida em um arquivo de texto (*massa.txt*, por exemplo) para que seja reconhecida pelo R como um objeto do tipo data frame:

```
> idade_massa = read.table("massa.txt",header=T)
> attach(idade_massa)
> idade_massa
   Massa Idade
1    100    43
2     95    43
3    116    45
.
.
.
.
```

Sejam X e Y, respectivamente, as variáveis Idade (preditora) e Massa Muscular (resposta). A função `lm` fornece o ajuste de um modelo de regressão linear de primeira ordem: $Y = \beta_0 + \beta_1 X + \epsilon$, onde β_0 e β_1 são parâmetros desconhecidos e ϵ é o erro aleatório.

```
> ajuste = lm(Massa ~ Idade)
> summary(ajuste)

Call:
lm(formula = Massa ~ Idade)

Residuals:
    Min      1Q  Median      3Q     Max
-12.973  -4.173  -1.387   6.739  14.771

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 143.6514     7.9441  18.083 1.74e-15
Idade        -0.9427     0.1305  -7.223 1.83e-07
---
Residual standard error: 7.525 on 24 degrees of freedom
Multiple R-squared: 0.6849,     Adjusted R-squared: 0.6718
F-statistic: 52.17 on 1 and 24 DF,  p-value: 1.832e-07
```

Logo, a equação da reta ajustada é dada por $\hat{Y} = 143,6514 - 0,9427X_i$ e o *Diagrama de Dispersão* para as variáveis em estudo é construído com a execução do comando:

```
> plot(Idade,Massa)
> abline(lm(Massa ~ Idade))
```

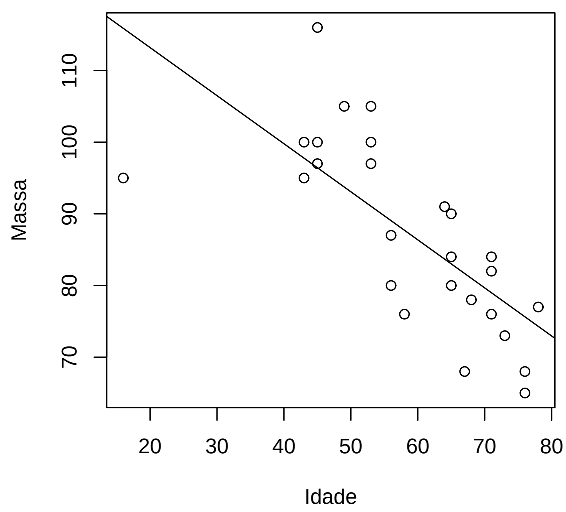


Figura 6: Diagrama de Dispersão de Massa Muscular versus Idade com a reta ajustada

Pela Figura 6, pode-se observar que existem valores diferentes de Y para valores fixos de X. Sendo assim, é necessário calcular o *Coeficiente de Determinação Real*, e as medidas do *Erro Puro* e da *Falta de Ajuste*.

Para completar a tabela ANOVA com os valores *Erro Puro* e a *Falta de Ajuste*, utilize a seguinte estratégia: ajuste um novo modelo que reserve um parâmetro em comum para agrupar os dados que possuem os mesmos valores de X. Para isso, basta declarar a variável preditora como `factor` (função que representa dados categóricos no R):

```
> erro_puro = lm(Massa ~ factor(Idade))
> anova(ajuste,erro_puro)
Analysis of Variance Table

Model 1: Massa ~ Idade
Model 2: Massa ~ factor(Idade)
  Res.Df     RSS Df Sum of Sq      F  Pr(>F)
1     24 1358.94
2     11  368.17 13    990.78 2.2771 0.08987 .
---
```

Note que a *Soma de Quadrados dos Resíduos* é igual `1358.94`, a *Soma de Quadrados do Erro Puro* é igual `368.17` e a *Soma de Quadrados da Falta de Ajuste* é igual `990.78`, com `24`, `11` e `13` graus de liberdade, respectivamente.

Ainda, devido aos valores replicados de X, o *Coeficiente de Determinação Real* para o modelo proposto é o quociente entre o *Coeficiente de Determinação* encontrado no ajuste do modelo e o *Coeficiente de Determinação Máximo* (obtido no ajuste em que X foi categorizado):

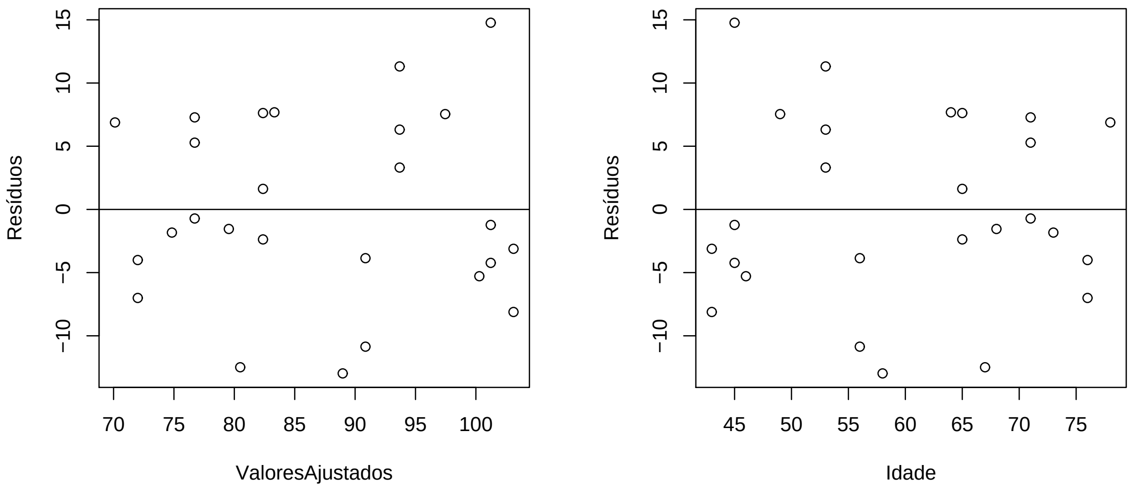
```
> ajuste_s = summary(ajuste)
> ajuste_s$r.squared
[1] 0.684928
>
> erro_puro_s = summary(erro_puro)
> erro_puro_s$r.squared
[1] 0.9146402
```

Sendo assim, $R^2 real = R^2/R^2_{max} = 0,684928/0,9146402 = 0,7488$.

4.2 Análise dos Resíduos

Considere os gráficos “Resíduos versus Valores Ajustados da Variável Resposta” e “Resíduos versus Valores da Variável Explicativa“:

```
> windows()
> plot(fitted(ajuste),residuals(ajuste),xlab="ValoresAjustados",ylab="Resíduos")
> abline(h=0)
>
> windows()
> plot(Idade,residuals(ajuste),xlab="Idade",ylab="Resíduos")
> abline(h=0)
```



(a) Resíduos vs Valores Ajustados

(b) Resíduos vs Variável Explicativa

Figura 7: Gráficos para Análise dos Resíduos

Observe da Figura 7, que não ocorre violação das suposições homocedasticidade e não correlação dos erros.

Para verificar a suposição sobre normalidade dos erros, construa o gráfico:

```
> windows()
> qqnorm(residuals(ajuste), ylab="Resíduos",main="")
> qqline(residuals(ajuste))
```

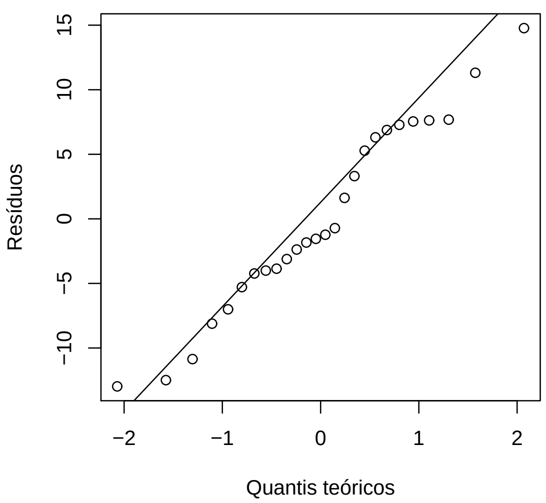


Figura 8: Gráfico de Probabilidade Normal dos Resíduos

Note que a Figura 8 não mostra evidências de que os erros não sigam a distribuição Normal. Essa suposição pode ser comprovada, também, pela realização do *Teste de Normalidade de Shapiro-Wilk* cujo

Valor P é `0.564`:

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

        Shapiro-Wilk normality test

data:  residuals(ajuste)
W = 0.9677, p-value = 0.564
```

4.3 Erro Puro e Teste da Falta de Ajuste

Para realizar o *Teste da Falta de Ajuste*, verifique o *valor p* na tabela ANOVA dos dois modelos ajustados:

```
> erro_puro = lm(Massa ~ factor(Idade))
> anova(ajuste,erro_puro)
Analysis of Variance Table

Model 1: Massa ~ Idade
Model 2: Massa ~ factor(Idade)
  Res.Df     RSS Df Sum of Sq      F  Pr(>F)
1     24 1358.94
2     11  368.17 13    990.78 2.2771 0.08987 .
---
```

A estatística F é igual a `2.2771` e o *valor p* do teste é `0.08987`.

4.4 Teste da Significância da Regressão

Verififque o *valor p* do *teste da significância da regressão* na tabela ANOVA do ajuste inicial:

```
F-statistic: 52.17 on 1 and 24 DF,  p-value: 1.832e-07
```

5 Regressão Linear Múltipla - Variável Indicadora

Deseja-se estudar a relação entre os salários e o tempo de experiência no cargo de gerente de agências bancárias de uma grande empresa e avaliar se há diferenças entre os sexos. Para isso, novamente, os dados devem ser inseridos no `R` como um `data-frame` (arquivo de texto *gerentes_ sexo.txt*, por exemplo):

```
> gerentes = read.table('gerentes_sexo.txt',header=T)
> attach(gerentes)
> Sexo = factor(Sexo)
> gerentes
   Salario Experiencia Sexo
1   1.9307           0    0
2   3.1769          17    0
3   2.2769           5    0
4   3.1307          15    0
5   2.7769           9    0
6   3.0923          15    0
7   2.6538           8    0
8   2.2230           5    0
9   2.8538          13    0
10  3.2307          20    0
11  2.8230          11    0
12  1.9076           1    0
13  2.5384           6    0
14  2.5692           7    1
15  4.2230          23    1
16  4.0923          20    1
17  3.6000          18    1
18  4.7076          27    1
19  3.1461          11    1
20  2.9923          10    1
21  4.7461          29    1
22  4.1153          23    1
23  2.3615           4    1
24  4.0923          22    1
25  4.5076          25    1
26  2.9076           9    1
27  4.4846          25    1
>
```

Observe que a variável *Sexo* é categórica, por isso foi definida acima como `factor`. Dessa forma, `0` e `1` representam os sexos feminino e masculino, respectivamente.

Para construir um diagrama de dispersão com marcadores diferentes para a variável *Sexo*:

```
> par(mfrow=c(1,1))
> plot(Experiencia[Sexo==1],Salario[Sexo==1],xlab="Experiência",ylab="Salário")
> points(Experiencia[Sexo==0],Salario[Sexo==0], pch = 19)
```

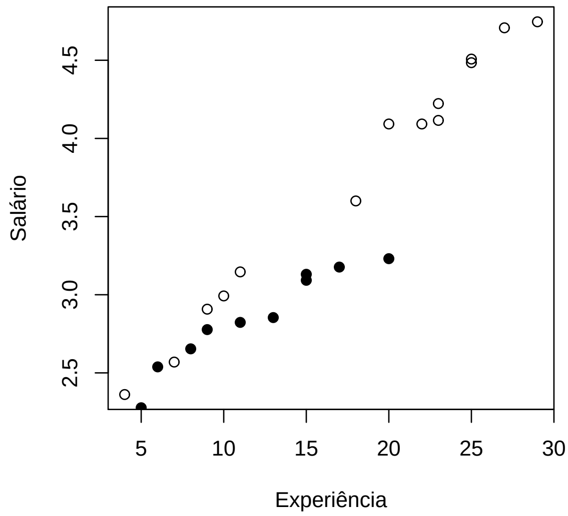


Figura 9: Gráfico de Dispersão de Salário versus Experiência: • feminino, ○ masculino.

5.1 Ajuste do Modelo Linear

Sejam as variáveis explicativas *Experiência* (X, em anos), *Sexo* (Z: 0, feminino; 1, masculino) e a variável resposta *Salário* (Y, em mil reais).

Propõe-se um modelo de regressão linear de primeira ordem, dado pela equação:

$$Y = \beta_0 + \beta_1 X + \alpha_1 Z + \gamma_1 XZ + \epsilon,$$

onde β_0, β_1, α_1 e γ_1 são parâmetros desconhecidos e ϵ é o erro aleatório.

```
> ajuste = lm(Salario ~ Experiencia + Sexo + Experiencia*Sexo)
> ajuste

Call:
lm(formula = Salario ~ Experiencia + Sexo + Experiencia * Sexo)

Coefficients:
     (Intercept)        Experiencia              Sexo1  Experiencia:Sexo1
        1.968445           0.072199           0.009082           0.026062
```

Logo, a equação da reta ajustada é dada por $\hat{Y} = 1,97 + 0,722X_i + 0,0091Z_i + 0,0261X_iZ_i$.

A tabela ANOVA é obtida com a execução do comando:

```
> anova(ajuste)

Analysis of Variance Table

Response: Salario
                 Df  Sum Sq Mean Sq  F value    Pr(>F)
Experiencia       1 18.1539 18.1539 1752.696 < 2.2e-16
Sexo              1  0.5570  0.5570   53.772 1.847e-07
Experiencia:Sexo  1  0.2050  0.2050   19.791 0.0001840
Residuals        23  0.2382  0.0104
---
```

Para exibir o valor do *Coeficiente de Determinação Ajustado*:

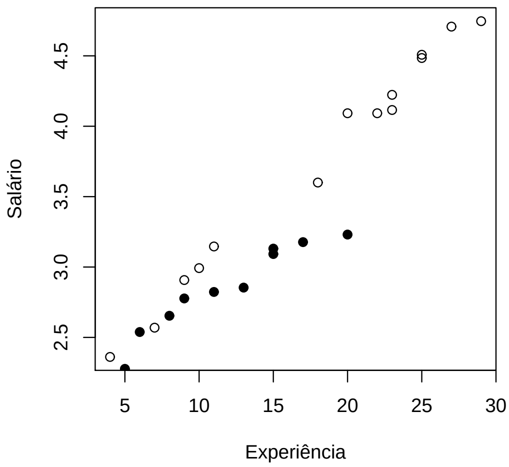
```
> ajuste_s = summary(ajuste)
> ajuste_s$adj.r.squared
[1] 0.9859403
```

5.2 Análise de Resíduos

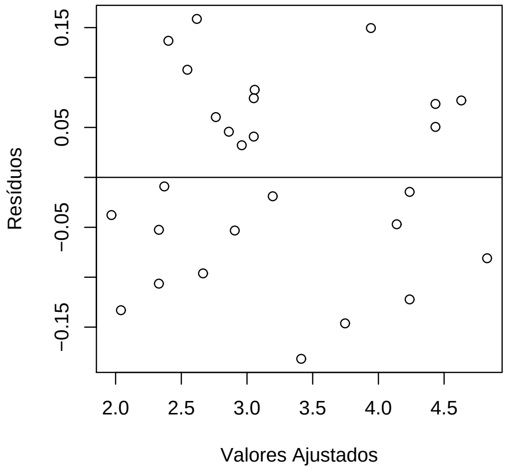
Considere os seguintes gráficos para a *Análise dos Resíduos*:

```
> windows()
> plot(fitted(ajuste),residuals(ajuste),xlab="ValoresAjustados",ylab="Resíduos")
> abline(h=0)
>
> windows()
> plot(Experiencia,residuals(ajuste),xlab="Experiencia",ylab="Resíduos")
> abline(h=0)
>
> windows()
> boxplot(residuals(ajuste)~ Sexo)
>
> windows()
> qqnorm(residuals(ajuste), ylab="Resíduos")
> qqline(residuals(ajuste))
```

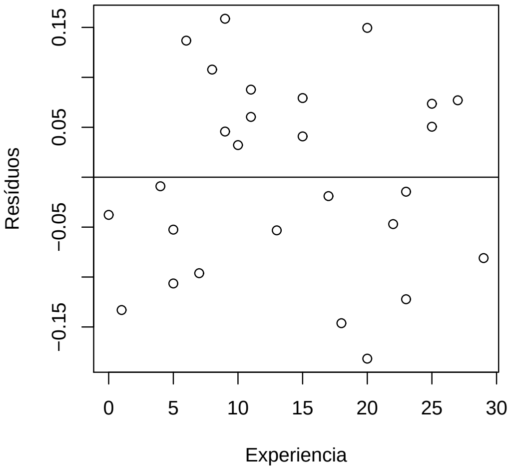
Observe, nestes gráficos (Figura 10), que há indicativos de que a variância dos erros é constante, não há evidências de que os erros não sigam a distribuição Normal e percebe-se de que os erros não são correlacionados entre si.



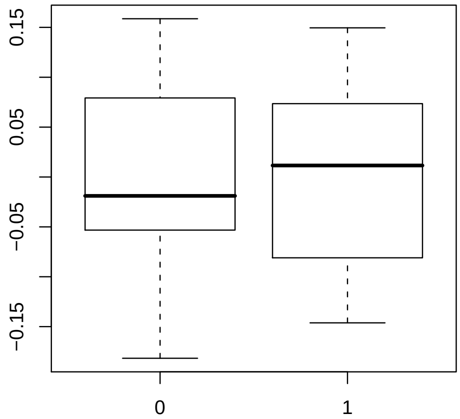
(a) Salário vs Experiência



(b) Resíduos vs Valores Ajustados



(c) Resíduos vs Experiência



(d) Boxplot de Resíduos, por sexo

Figura 10: Gráficos para Análise dos Resíduos

5.3 Testes de Hipóteses

Para proceder o *Teste F da Significância da Regressão* e os *Testes t individuais*, verifique os *Valor P* para cada caso através da saída da função `summary`:

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       1.968445   0.053691  36.662  < 2e-16
Experiencia       0.072199   0.004750  15.200 1.73e-13
Sexo              0.009082   0.086385   0.105 0.917183
Experiencia:Sexo  0.026062   0.005858   4.449 0.000184
---
F-statistic: 608.8 on 3 and 23 DF,  p-value: < 2.2e-16
```

Note que o coeficiente da variável *Sexo* não é significativo (*valor p* = 0.917183), o que indica que as duas retas, para homens e mulheres, de *Salario* em função da *Experiencia*, são paralelas. Desse modo, o modelo pode ser reajustado sem o termo linear aditivo de *Sexo*:

```
ajuste2 = lm(Salario ~ Experiencia + Experiencia*Sexo - Sexo)
> ajuste2

Call:
lm(formula = Salario ~ Experiencia + Experiencia * Sexo - Sexo)

Coefficients:
     (Intercept)       Experiencia  Experiencia:Sexo
         1.97195           0.07194           0.02658

> summary(ajuste2)

Call:
lm(formula = Salario ~ Experiencia + Experiencia * Sexo - Sexo)

Residuals:
      Min        1Q    Median        3Q       Max
-0.179959 -0.068831 -0.004533  0.074128  0.157529

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      1.971953   0.041185  47.880  < 2e-16
Experiencia      0.071935   0.003948  18.221 1.47e-15
Experiencia:Sexo 0.026585   0.003035   8.759 6.12e-09
---
Residual standard error: 0.09965 on 24 degrees of freedom
Multiple R-squared: 0.9876,     Adjusted R-squared: 0.9865
F-statistic: 952.4 on 2 and 24 DF,  p-value: < 2.2e-16
```

6 Modelos Polinomiais

Uma indústria está iniciando a produção de uma nova substância química. A meta é a produção da substância com um valor mínimo estabelecido para a variável *rendimento da reação química* (Y) a partir da qual a substância é produzida. As variáveis candidatas a preditoras são o *tempo de reação* (X_1) e a *temperatura do reator* (X_2).

Tabela 3: Tempo, Temperatura e Rendimento de uma reação química

Tempo	Temperatura	Rendimento
81.0	173	60.95
85.5	168	57.35
83.0	185	60.99
94.0	188	54.96
98.0	183	51.89
97.1	175	51.44
82.8	173	61.79
89.0	183	60.78
92.3	168	52.48
80.0	175	59.8
90.0	188	58.74
95.0	179	56.2
90.0	181	60.49
84.2	177	62.78
88.0	171	59.71
87.0	182	62.75
99.0	184	49.41
96.4	187	53.63
100.0	180	48.19
86.1	172	60.92

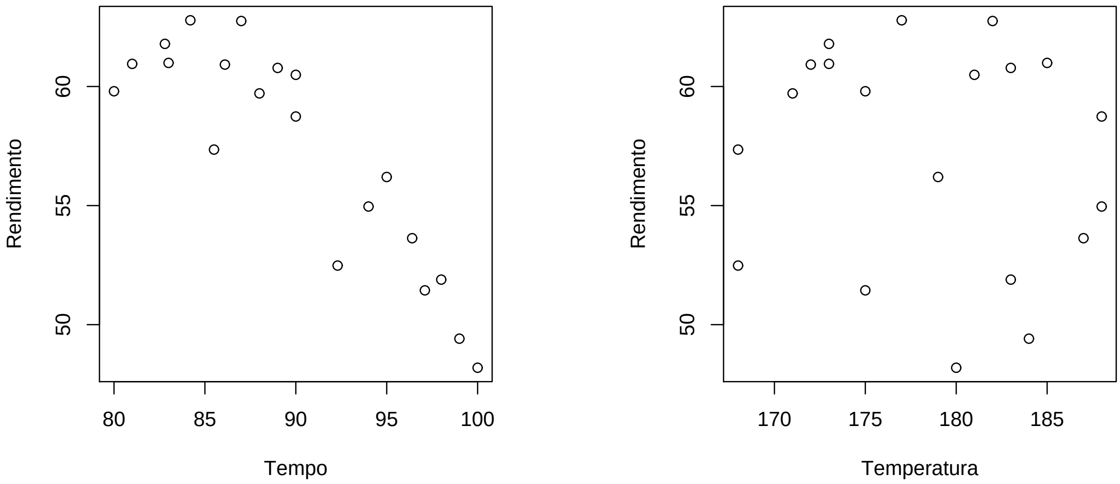
Para a análise dos dados no `R` suponha que a tabela acima esteja contida no seguinte arquivo de texto *quimica.txt*, desse modo:

```
> quimica = read.table("quimica.txt",header=T)
> attach(quimica)
> quimica
   Tempo Temperatura Rendimento
1   81.0          173      60.95
2   85.5          168      57.35
3   83.0          185      60.99
.
.
```

Veja a relação entre as variáveis explicativas e a variável resposta:

```
> par(mfrow=c(1,2))
> plot(Rendimento~Tempo)
> plot(Rendimento~Temperatura)
```

O diagrama de dispersão da Figura 11 (a) mostra uma relação curvilínea entre as variáveis Tempo e Rendimento. Já entre as variáveis Temperatura e Rendimento (Figura 11 (b)), não há uma relação bem definida.



(a) Tempo

(b) Temperatura

Figura 11: Diagramas de dispersão das Variáveis explicativas vs Variável resposta

Ajustaremos um modelo polinomial de grau 2 aos dados. Lembrando que as variáveis explicativas devem estar centradas em suas médias para o ajuste de um modelo polinomial, primeiramente realize tais centralizações pelos seguintes comandos:

```
> mean(Tempo)
[1] 89.92
> mean(Temperatura)
[1] 178.6
> Tempo1 = Tempo-mean(Tempo)
> Temperatura1 = Temperatura-mean(Temperatura)
```

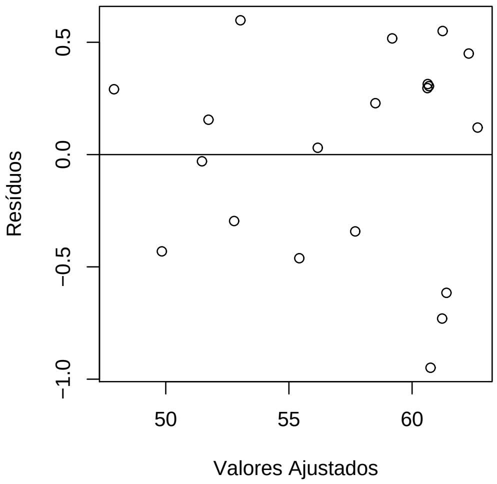
O modelo ajustado abaixo, do *Rendimento* em *Tempo−Média* e *Temperatura−Média* é:

$$\begin{aligned}\hat{Y} = 61,24 \quad &- \quad 0,66(X_1 - 89,92) - 0.07(X_1 - 89,92)^2 \\ &+ \quad 0,12(X_2 - 178,60) - 0.04(X_2 - 178,60)^2 \\ &+ \quad 0,02(X_1 - 89,92)(X_2 - 178,60)\end{aligned}$$

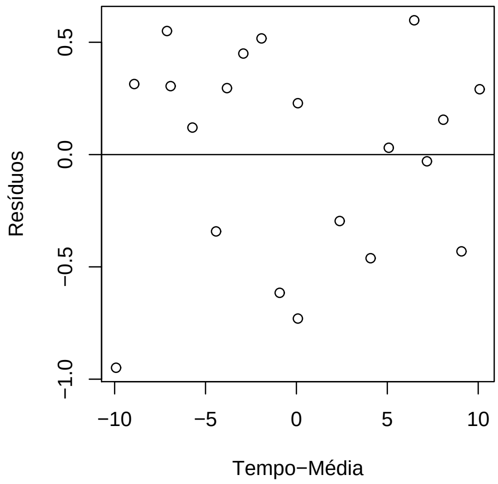
com $R^2 ajustado = 0,99$. Para obter $\hat{Y}$ em função de X_1 e X_2, basta desenvolver a equação acima.

```
> ajuste = lm(Rendimento~Tempo1+I(Tempo1^2)+Temperatura1+
I(Temperatura1^2)+Tempo1*Temperatura1)
> summary(ajuste)
Call:
lm(formula = Rendimento ~ Tempo1 + I(Tempo1^2) + Temperatura1 +
    I(Temperatura1^2) + Tempo1 * Temperatura1)

Residuals:
    Min      1Q  Median      3Q     Max
-0.9491 -0.3643  0.1377  0.3071  0.5979
```

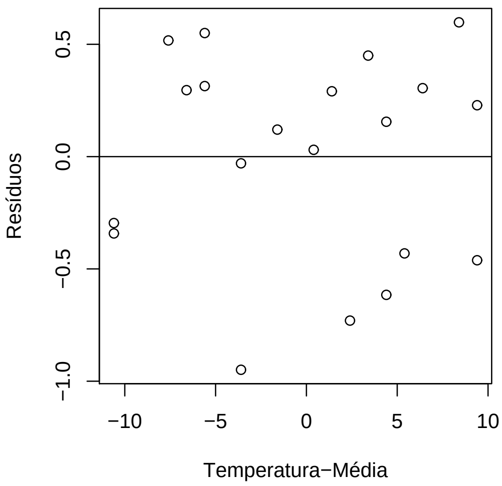
```
Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         61.239297   0.284065 215.582  < 2e-16
Tempo1              -0.663385   0.021557 -30.773 2.94e-14
I(Tempo1^2)         -0.068866   0.004575 -15.053 4.86e-10
Temperatura1         0.115252   0.021065   5.471 8.24e-05
I(Temperatura1^2)   -0.042690   0.003956 -10.791 3.61e-08
Tempo1:Temperatura1  0.018875   0.004918   3.838  0.00181
---
Residual standard error: 0.5347 on 14 degrees of freedom
Multiple R-squared: 0.9902,     Adjusted R-squared: 0.9867
F-statistic:   283 on 5 and 14 DF,  p-value: 1.521e-13


> shapiro.test(residuals(ajuste))

        Shapiro-Wilk normality test

data:  residuals(ajuste)
W = 0.9283, p-value = 0.1429


> windows()
> par(mfrow = c(2, 2))
> plot(fitted(ajuste), residuals(ajuste),xlab="Valores Ajustados",ylab="Resíduos")
> abline(h=0)
> plot(Tempo1, residuals(ajuste),xlab="Tempo-Média",ylab="Resíduos")
> abline(h=0)
> plot(Temperatura1, residuals(ajuste),xlab="Temperatura-Média",ylab="Resíduos")
> abline(h=0)
> qqnorm(residuals(ajuste), ylab="Resíduos",xlab="Quantis teóricos",main="")
> qqline(residuals(ajuste))
```



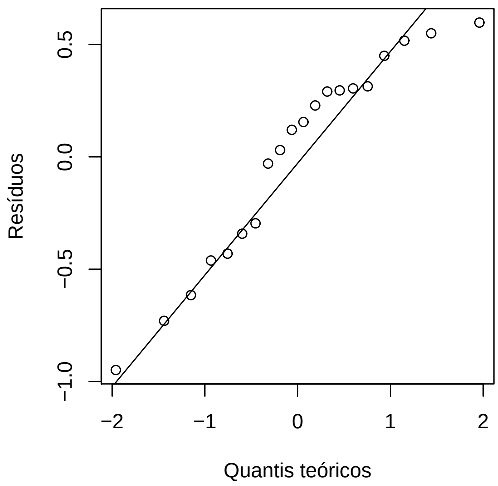
(a) Resíduos vs Valores ajustados



(b) Resíduos vs Tempo-Média



(c) Resíduos vs Temperatura-Média



(d) Gráfico de Probabilidade Normal dos Resíduos

Figura 12: Gráficos para Análise dos Resíduos

A análise dos resíduos mostrada na Figura 12 indica que as condições de homocedasticidade, normalidade e ausência de correlação entre erros foram satisfeitas. A normalidade ainda é confirmada pelo Teste de normalidade de *Shapiro-Wilk*, cujo *P-valor* é 0,1429.

7 Método dos Mínimos Quadrados Generalizados

Um analista de vendas gostaria de encontrar a relação entre a *renda mensal média de vendas de refeições* (Y) e o *gasto mensal com propaganda* (X). Os dados de 30 restaurantes encontram-se na tabela abaixo:

Tabela 4: Renda Mensal Média de Vendas de Refeições e Gasto Mensal com Propaganda em 30 restaurantes

Gasto(X)	Venda(Y)
3	81
3	73
3	72
5	91
5	99
5	97
9	127
9	114
9	116
9	123
9	131
12	141
12	151
12	147
12	131
12	145
12	147
15	179
15	166
15	181
15	178
15	185
15	156
17	176
17	189
17	192
19	203
19	193
19	219
19	214

Para a análise dos dados no `R` suponha que a tabela acima esteja contida no seguinte arquivo de texto *restaurantes.txt*, desse modo:

```
> gasto_venda = read.table("restaurantes.txt",header=T)
> attach(gasto_venda)
> gasto_venda
> plot(Venda ~ Gasto)
```

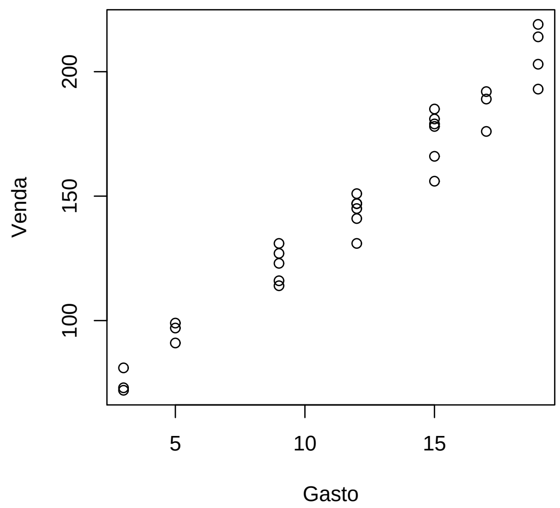


Figura 13: Gráfico de Dispersão entre Vendas de Refeições e Gasto com Propaganda para os 30 restaurantes.

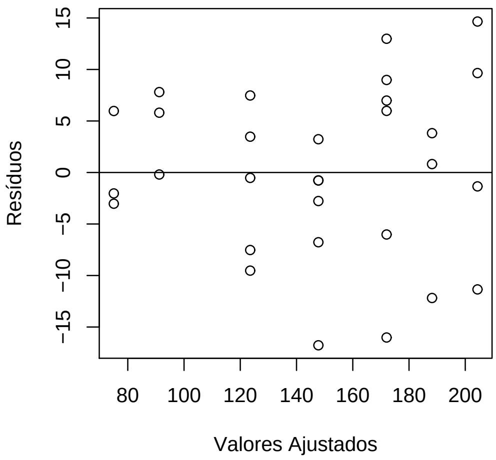
O Diagrama acima evidencia uma forte correlação linear crescente entre as variáveis em estudo. Sendo assim, propõe-se um modelo de regressão linear de primeira ordem, dado pela equação: $Y = \beta_0 + \beta_1 X + \epsilon$, onde β_0 e β_1 são parâmetros desconhecidos e ϵ é o erro aleatório.

7.1 Ajuste do Modelo via Método de Mínimos Quadrados Ordinários

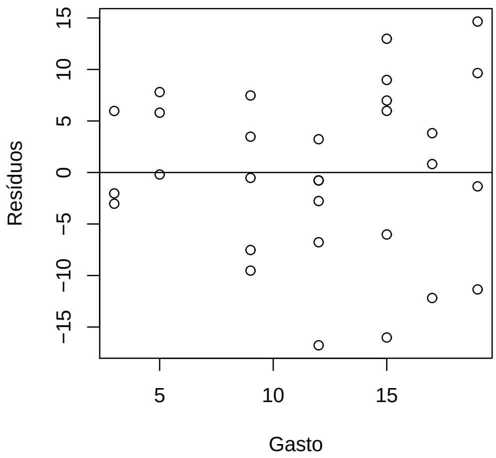
Para ajustar o modelo via Mínimos Quadrados Ordinários, execute os seguintes comandos:

```
> ajuste = lm(Venda ~ Gasto)
> summary(ajuste)

Call:
lm(formula = Venda ~ Gasto)

Residuals:
     Min       1Q   Median       3Q      Max
-16.7722  -5.2722  -0.3607   5.9778  14.6518

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  50.7846     3.9049   13.01 2.18e-13
Gasto         8.0823     0.3015   26.81  < 2e-16
---
Residual standard error: 8.311 on 28 degrees of freedom
Multiple R-squared: 0.9625,     Adjusted R-squared: 0.9612
F-statistic: 718.6 on 1 and 28 DF,  p-value: < 2.2e-16

> windows()
> par(mfrow=c(2,2))
```

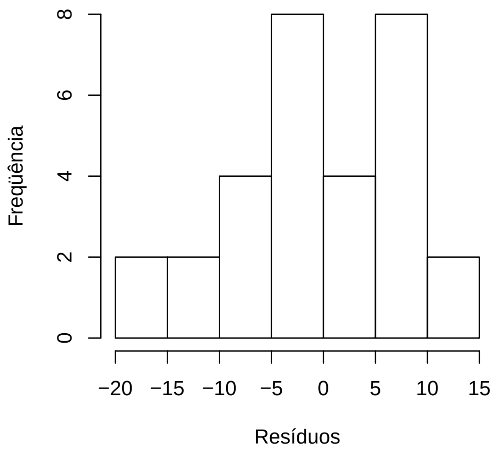
```
> plot(fitted(ajuste),residuals(ajuste),xlab="ValoresAjustados",ylab="Resíduos")
> abline(h=0)
> plot(Gasto,residuals(ajuste),xlab="Gasto",ylab="Resíduos")
> abline(h=0)
> hist(residuals(ajuste), xlab="Resíduos",ylab="Freqüência",main="")
> qqnorm(residuals(ajuste), ylab="Resíduos",main="")
> qqline(residuals(ajuste))
```



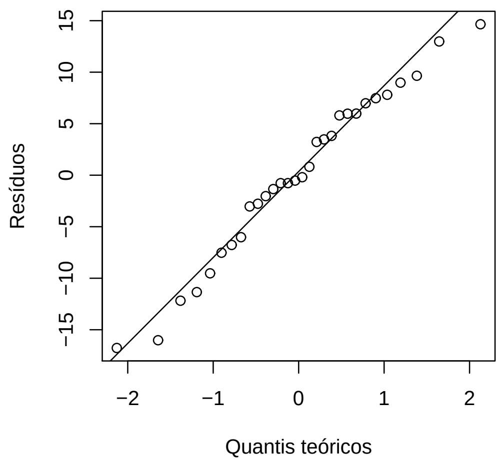
(a) Resíduos vs Valores Ajustados



(b) Resíduos vs Variável explicativa



(c) Histograma dos Resíduos



(d) Gráfico de Probabilidade Normal dos Resíduos

Figura 14: Gráficos para Análise dos Resíduos

Logo, a equação da reta ajustada é dada por $\hat{Y} = 50,785 + 8,082X_i$.

Observe pelos gráficos acima, que os resíduos não se distribuem homogeneamente em torno de zero ao longo dos valores ajustados de *vendas* e dos valores do *gasto*, nos dois casos a variação dos resíduos cresce, o que indica que a variância dos erros não é constante.

7.2 Ajuste do Modelo via Método de Mínimos Quadrados Ponderados

Para corrigir o problema da heterocedasticidade dos erros, detectado na seção anterior, deve-se realizar o ajuste do modelo utilizando o Método de Mínimos Quadrados Ponderados. Para isso deve-se observar as estimativas do *Erro Puro* para cada nível de X, ou seja, os valores de Var(Y | X). Observe a função para o *R* abaixo:

```
> tapply(gastovenda[,2],as.factor(gastovenda[,1]),var)
        3         5         9        12        15        17        19
 24.33333  17.33333  51.70000  49.06667 119.76667  72.33333 134.91667
>
> plot ( Gasto,v,xlab="Gasto",ylab="Variância (Venda|Gasto)" )
```

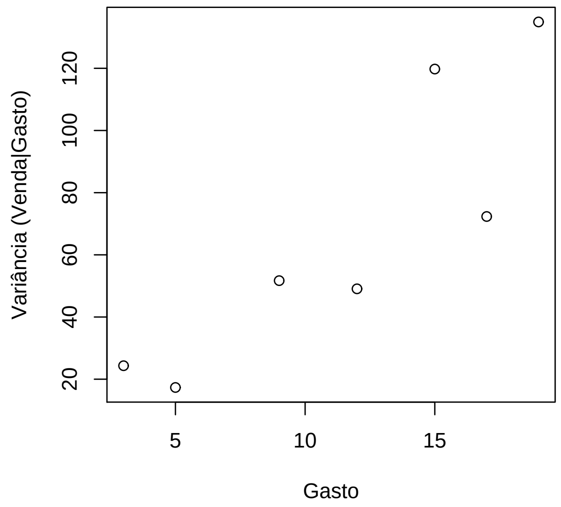


Figura 15: Gráfico de Dispersão de Gasto por Var(Venda | Gasto)

A Figura 15 mostra que Var(Venda | Gasto) é proporcional ao Gasto. Sendo assim, o peso W_i deve ser inversamente proporcional ao X_i.

Tabela 5: Valores para os pesos W_i

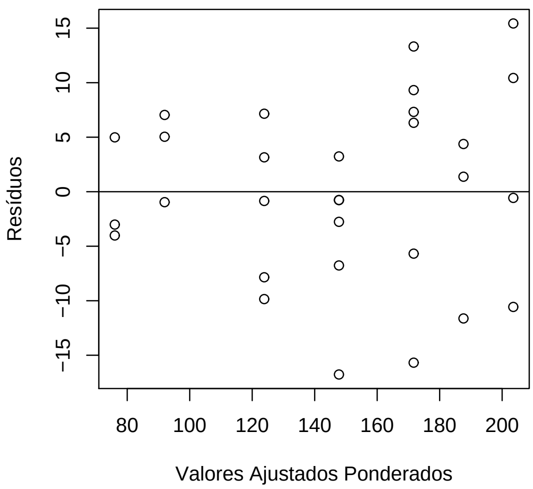
X	3	5	9	12	15	17	19
$w_i = 1/X_i$	0.3333	0.2000	0.1111	0.0833	0.0667	0.0588	0.0526

Abaixo encontra-se o comando do *R* para o ajuste do modelo via Método de Mínimos Quadrados Ponderados e a respectiva saída do software com os coeficientes ajustados.

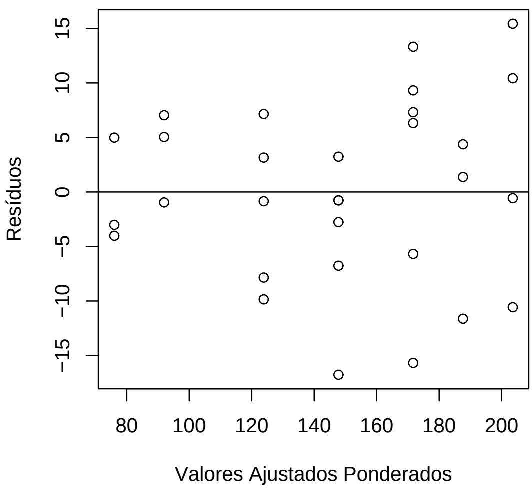
```
> ajuste_ponderado = lm(Venda ~ Gasto, weights = 1/Gasto)
> summary(ajuste_ponderado)

Call:
lm(formula = Venda ~ Gasto, weights = 1/Gasto)
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-4.8396 -1.8999 -0.1757  2.1629  3.5400

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   52.099      2.488   20.94   <2e-16
Gasto          7.972      0.244   32.68   <2e-16
---
Residual standard error: 2.397 on 28 degrees of freedom
Multiple R-squared: 0.9744,     Adjusted R-squared: 0.9735
F-statistic:  1068 on 1 and 28 DF,  p-value: < 2.2e-16

> anova(ajuste_ponderado)

Analysis of Variance Table

Response: Venda
          Df Sum Sq Mean Sq F value    Pr(>F)
Gasto      1 6137.2  6137.2  1067.8 < 2.2e-16
Residuals 28  160.9     5.7
---
```

A Figura 16 evidencia que o problema da heterocedasticidade dos erros foi solucionado, pois nos dois gráficos os resíduos ponderados estão dispostos homogeneamente em torno de zero.



(a) Resíduos vs Valores ajustados



(b) Resíduos vs Gasto

Figura 16: Gráficos para Análise dos Resíduos

8 Transformações

8.1 Transformação na Variável Resposta

No arquivo *dadosozonio.txt* estão as variáveis *ozônio* e *temperatura* de 330 medições. Estas variáveis são parte do banco de dados *ozone* da biblioteca *faraway* do *R*.

```
> ozdata = read.table("dadosozonio.txt", header = T)
> attach(ozdata)
> ozdata
    ozonio temperatura
1        3          40
2        5          45
3        5          54
4        6          35
.
.
.
```

Após a leitura dos dados, veja a relação entre as duas variáveis:

```
> plot(ozonio ~ temperatura)
```

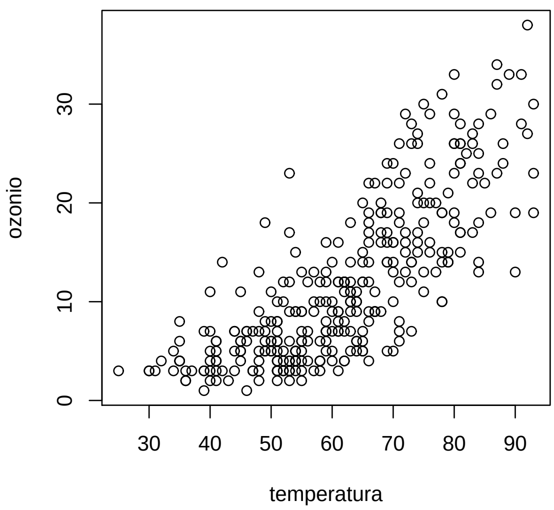
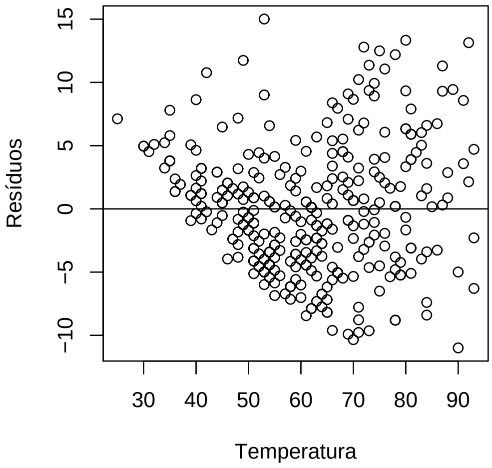


Figura 17: Gráfico de dispersão de Temperatura vs Ozônio

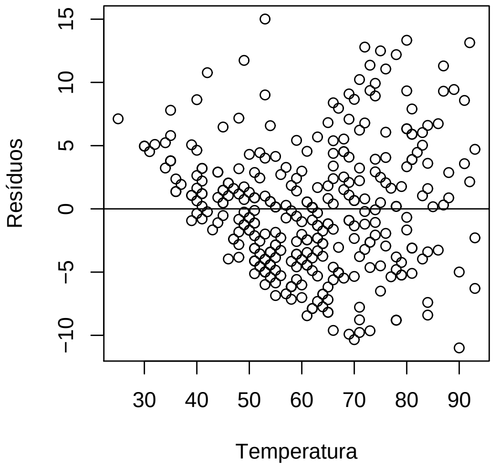
Observe que o diagrama de dispersão da Figura 17 mostra uma forte relação crescente não linear entre as medidas de *ozônio* e *temperatura*. As medidas de *ozônio* apresentam aumento de variabilidade para valores crescentes de *temperatura*.

Diante disso, é fácil concluir que o ajuste do modelo de regressão linear simples com as variáveis na sua forma original é inadequado neste caso. No entanto, tal ajuste é realizado a seguir, a fim de evidenciar sua inadequação na Análise de resíduos. O modelo linear ajustado abaixo, do *ozônio* em *temperatura* é $\hat{Y} = -14,9 + 0,43X_i$, com $R^2 ajustado = 0,61$.

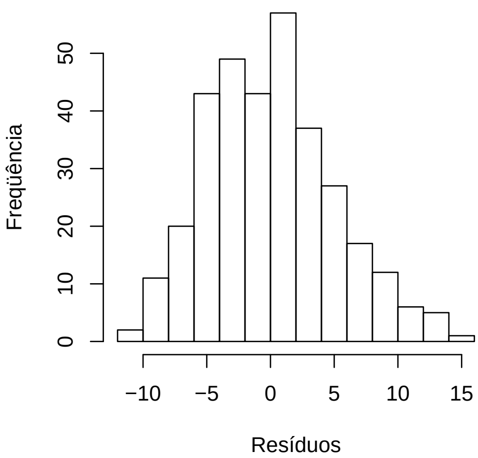
```
> oz1 = lm(ozonio ~ temperatura, ozdata)
man
> summary(oz1)

Call:
lm(formula = ozonio ~ temperatura, data = ozdata)

Residuals:
     Min       1Q   Median       3Q      Max
-10.9939  -3.8202  -0.1796   3.1951  15.0112

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -14.93745    1.21247  -12.32   <2e-16
temperatura   0.43257    0.01912   22.63   <2e-16
---

Residual standard error: 5.014 on 328 degrees of freedom
Multiple R-squared: 0.6095,     Adjusted R-squared: 0.6083
F-statistic: 511.9 on 1 and 328 DF,  p-value: < 2.2e-16

> shapiro.test(residuals(oz1))

        Shapiro-Wilk normality test

data:  residuals(oz1)
W = 0.9856, p-value = 0.002237

> windows()
> par(mfrow = c(2, 2))
> plot(fitted(oz1), residuals(oz1),xlab="Valores Ajustados",ylab="Resíduos")
> abline(h=0)
> plot(temperatura, residuals(oz1),xlab="Temperatura",ylab="Resíduos")
> abline(h=0)
> hist(residuals(oz1),main="",xlab="Resíduos",ylab="Freqüência")
> qqnorm(residuals(oz1),main="",xlab="Quantis teóricos",ylab="Resíduos")
> qqline(residuals(oz1))
```



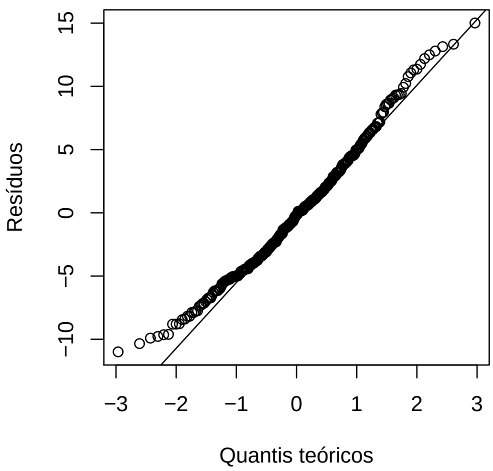
(a) Resíduos vs Valores ajustados



(b) Resíduos vs Variável explicativa



(c) Histograma dos Resíduos



(d) Gráfico de Probabilidade Normal dos Resíduos

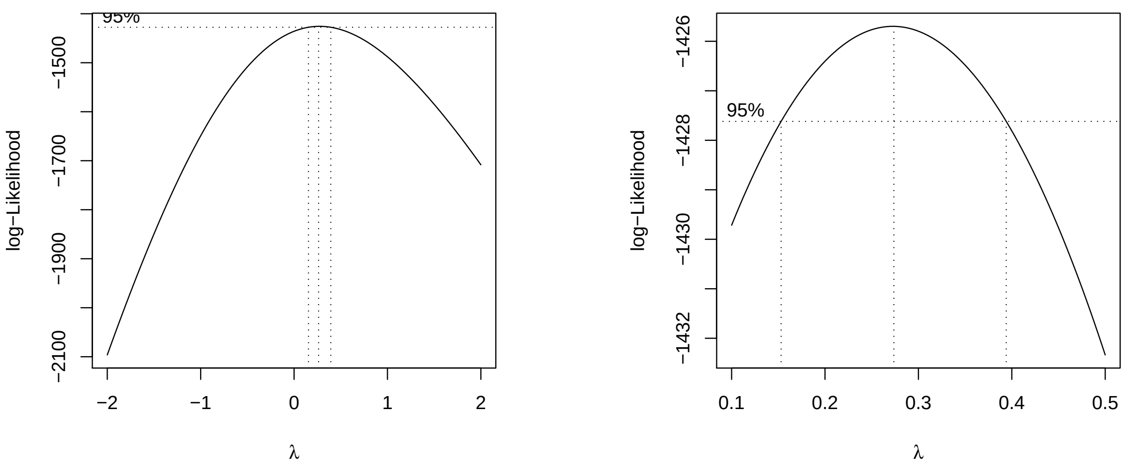
Figura 18: Gráficos para Análise dos Resíduos

Observe da Figura 18 que a variância dos erros não é constante e que a normalidade dos erros é violada, suposição que também é rejeitada pelo Teste de normalidade de *Shapiro-Wilk*, cujo *P-valor* é 0,0022.

A fim de solucionar os problemas de variância não-constante e não-normalidade dos erros, deve-se tentar realizar uma transformação na variável resposta. Apesar de ser possível, em muitos casos, selecionar empiricamente a transformação adequada, apresentaremos aqui apenas a técnica mais formal e objetiva. Uma transformação adequada para a variável resposta via *Procedimento de Box Cox* é obtida da seguinte forma:

```
> require(MASS)
> windows()
> par(mfrow = c(1, 2))
```

```
> boxcox(oz1)
> boxcox(oz1, lambda = seq(0.1, 0.5, by = 0.01))
```



(a) Resíduos vs Valores Ajustados

(b) Resíduos vs Variável explicativa

Figura 19: Log-verossimilhança e Intervalo de Confiança de 95% para os valores de λ da transformação de Box Cox

Note que a Figura 19 mostra os valores da log-verossimilhança para um intervalo de valores do parâmetro de transformação λ. O máximo da verossimilhança foi atingido com aproximadamente $\lambda = 0,27$, com intervalo de confiança de 95% igual a [0,15;0,39]. Como esse intevalo não inclui o valor 1, há forte evidência da necessidade de transformação na variável resposta *ozônio*, dado por: $ozônio^* = (ozônio^{0,27} - 1)/0,27$. Sendo assim a nova variável transformada *ozoniotrans* deve ser inserida no banco de dados, para que o novo modelo de regressão linear simples seja ajustado.

```
> ozoniotrans = (ozonio^0.27-1)/0.27
> ozdatatrans = data.frame(cbind(ozdata,ozoniotrans))
> ozdatatrans
    ozonio temperatura ozoniotrans
1        3          40   1.2789302
2        5          45   2.0157973
3        5          54   2.0157973
.
.
.
> plot(temperatura, ozoniotrans,ylab="ozonio transformado")
```

A Figura 24 (página seguinte) mostra uma forte relação linear crescente entre as medidas de *ozônio* transformadas, via método de *Box Cox*, versus *temperatura*, com variabilidade aproximadamente constante.

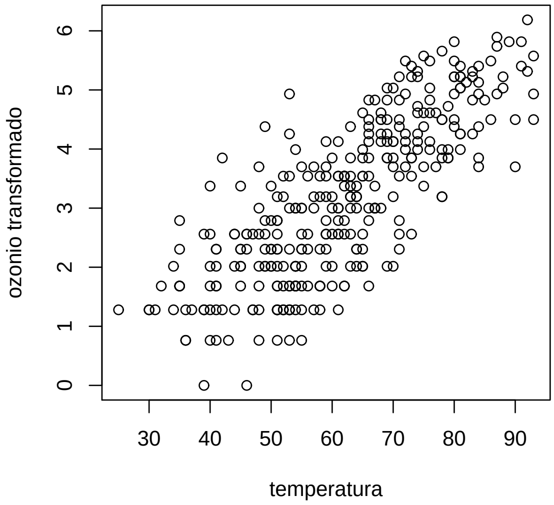
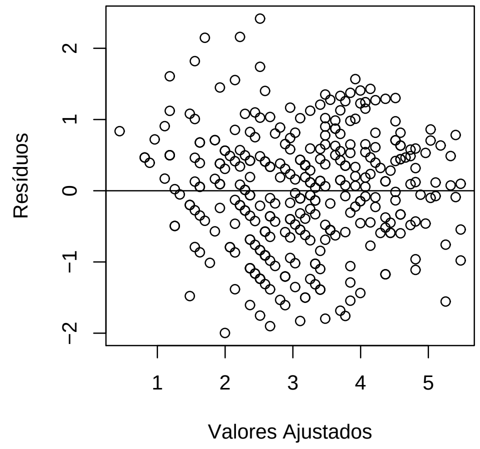
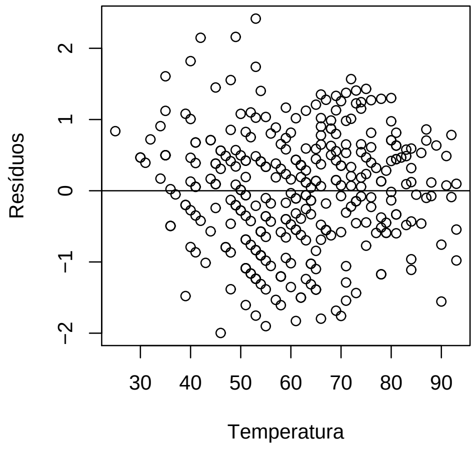


Figura 20: Diagrama de dispesão da transformação de *ozônio* versus *temperatura*

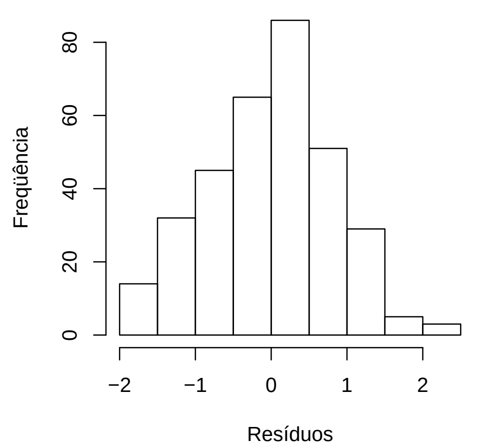
```
> oz2 = lm(ozoniotrans ~ temperatura, ozdatatrans)
> summary(oz2)

Call:
lm(formula = ozoniotrans ~ temperatura, data = ozdatatrans)

Residuals:
     Min       1Q   Median       3Q      Max
-1.99712 -0.56569  0.07148  0.56078  2.41671

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.408685   0.199642  -7.056 1.02e-11
temperatura  0.074039   0.003148  23.520  < 2e-16
---
Residual standard error: 0.8256 on 328 degrees of freedom
Multiple R-squared: 0.6278,     Adjusted R-squared: 0.6266
F-statistic: 553.2 on 1 and 328 DF,  p-value: < 2.2e-16

> windows()
> par(mfrow = c(2, 2))
> plot(fitted(oz2), residuals(oz2),xlab="Valores Ajustados",ylab="Resíduos")
> abline(h=0)
> plot(temperatura, residuals(oz2),xlab="Temperatura",ylab="Resíduos")
> abline(h=0)
> hist(residuals(oz2),main="",xlab="Resíduos",ylab="Freqüência")
> qqnorm(residuals(oz2),main="",xlab="Quantis teóricos",ylab="Resíduos")
> qqline(residuals(oz2))
```



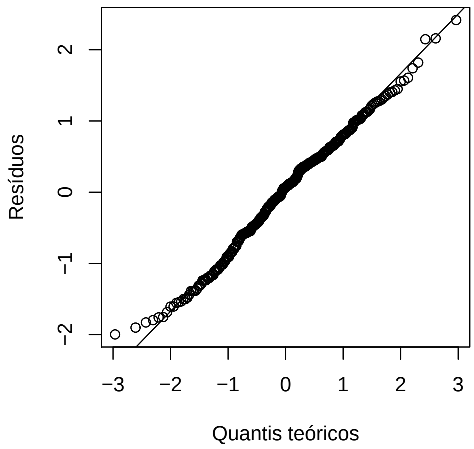
(a) Resíduos vs Valores Ajustados



(b) Resíduos vs variável explicativa



(c) Histograma dos Resíduos



(d) Gráfico de Probabilidade Normal dos Resíduos

Figura 21: Gráficos para Análise dos Resíduos

Logo, a equação da reta ajustada é dada por $\hat{Y} = -1,41 + 0,075X_i$, com $R^2 ajustado = 0,63$ e pela Figura 21 observa-se que a suposição de normalidade é aceitável, embora a transformação não tenha solucionado o problema da heterocedasticidade dos erros.

8.2 Transformação na Variável Explicativa

O gerente de Recursos Humanos de uma loja deseja estudar o efeito do *número de dias de treinamento (X)* no *desempenho em um teste simulado de vendas (Y)* aplicado em seus vendedores. Os dados coletados estão na Tabela 6.

Para a análise dos dados no `R` suponha que a tabela acima esteja contida no seguinte arquivo de texto *vendedores.txt*, desse modo:

```
> treino_venda = read.table("vendedores.txt",header=T)
```

Tabela 6: Tempo de treinamento (em dias) e Desempenho de venda de 10 vendedores

Tempo de treinamento	Desempenho
0.5	42.5
0.5	50.6
1	68.5
1	80.7
1.5	89
1.5	99.6
2	105.3
2	111.8
2.5	112.3
2.5	125.7

```
> attach(treino_venda)
> treino_venda
  Tempo Desempenho
1  0.5       42.5
2  0.5       50.6
3  1.0       68.5
.
.
.
```

No diagrama de dispersão veja a relação entre as duas variáveis:

```
> plot(Desempenho~Tempo,xlab="Tempo de treinamento",ylab="Desempenho de venda")
```

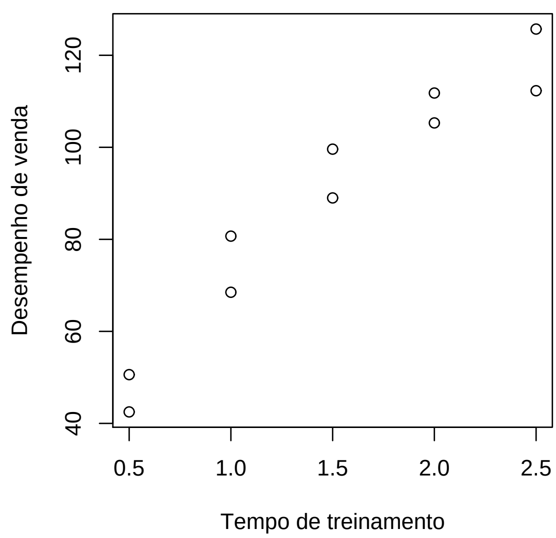
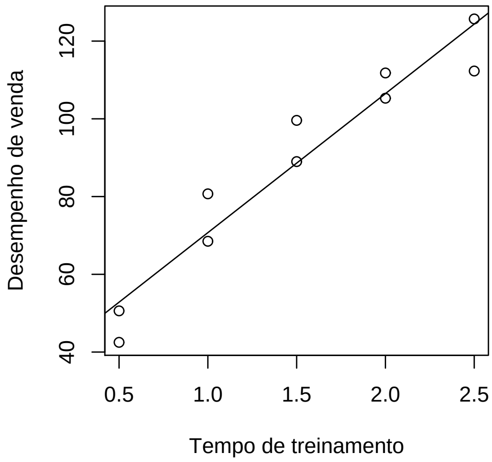


Figura 22: Gráfico de dispersão de Dias de treinamento vs Desempenho de venda

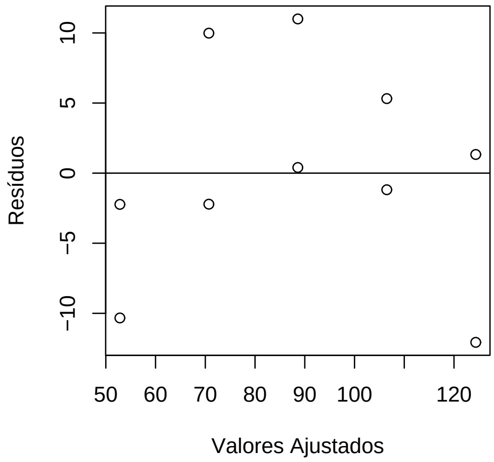
O diagrama de dispersão da Figura 22 mostra uma relação curvilínea entre X e Y, com variabilidade aproximadamente constante nos níveis de X. Desse modo, pode-se tentar realizar uma transformação apenas em X. A escolha da transformação adequada aqui é feita apenas de forma empírica. Baseando-se em padrões já conhecidos, foi escolhida a função raiz quadrada, ou seja, $X' = \sqrt{X}$.

Entretanto, para fins comparativos, antes de realizar a regressão linear simples com a variável X', realiza-se a seguir a regressão com a variável X original. O modelo linear ajustado abaixo, do *Desempenho* em *Tempo* é $\hat{Y} = 34,95 + 35,77X_i$, com $R^2 ajustado = 0,92$.

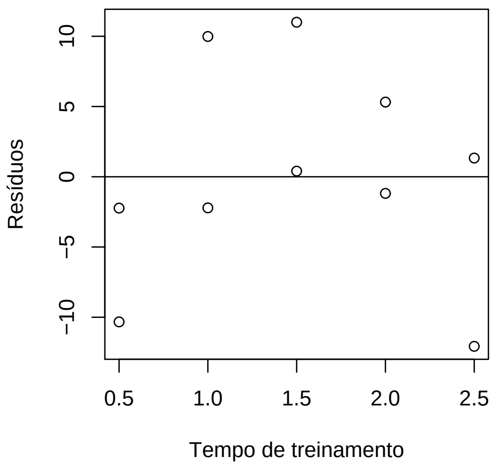
```
> ajuste1 = lm(Desempenho~Tempo,data=treino_venda)
> summary(ajuste1)

Call:
lm(formula = Desempenho ~ Tempo, data = treino_venda)

Residuals:
     Min       1Q   Median       3Q      Max
-12.0700  -2.2263  -0.3925   4.3188  11.0000

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   34.945      5.948   5.875 0.000372
Tempo         35.770      3.587   9.973 8.66e-06
---
Residual standard error: 8.02 on 8 degrees of freedom
Multiple R-squared: 0.9256,     Adjusted R-squared: 0.9163
F-statistic: 99.46 on 1 and 8 DF,  p-value: 8.66e-06

> shapiro.test(residuals(ajuste1))

        Shapiro-Wilk normality test

data:  residuals(ajuste1)
W = 0.9436, p-value = 0.5936

> windows()
> par(mfrow = c(2, 2))
> plot(Desempenho~Tempo,xlab="Tempo de treinamento",ylab="Desempenho de venda")
> abline(ajuste1)
> plot(fitted(ajuste1),residuals(ajuste1),xlab="Valores Ajustados",ylab="Resíduos")
> abline(h=0)
> plot(Tempo, residuals(ajuste1),xlab="Tempo de treinamento",ylab="Resíduos")
> abline(h=0)
> hist(residuals(ajuste1),main="",xlab="Resíduos",ylab="Freqüência")
```



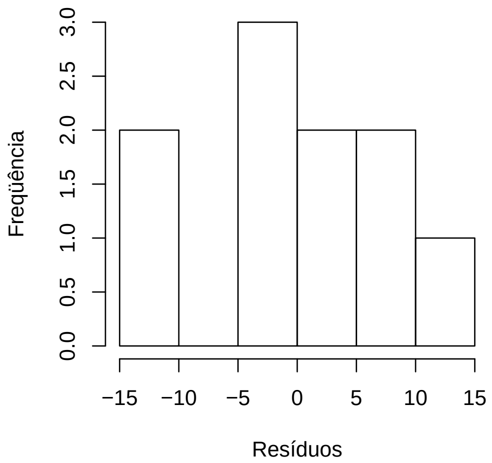
(a) Resposta vs explicativa com reta de regressão



(b) Resíduos vs Valores ajustados



(c) Resíduos vs Variável explicativa



(d) Histograma dos Resíduos

Figura 23: Gráficos para Análise dos Resíduos

Observe da Figura 23(a) que a reta de regressão não acompanha bem todos os pontos, evidenciando o aspecto curvilíneo entre X e Y. A análise dos resíduos mostrada nas Figuras 23(b), (c) e (d) indica um bom ajuste do modelo, assim como o Teste de normalidade de *Shapiro-Wilk*, cujo *P-valor* é 0,5936.

A fim de linearizar o modelo acima, sem modificar as condições de normalidade e homocedasticidade para os erros, a transformação adequada para a variável explicativa (raiz quadrada) é feita a seguir:

```
> Tempotrans = sqrt(Tempo)
> treino_venda_trans = data.frame(cbind(treino_venda,Tempotrans))
> plot(Desempenho~Tempotrans,xlab="Tempo de treinamento transformado",ylab="Desempenho de venda")
```

A Figura 24 (página seguinte) mostra agora um relacionamento linear crescente entre X e Y, ainda com variabilidade aproximadamente constante.

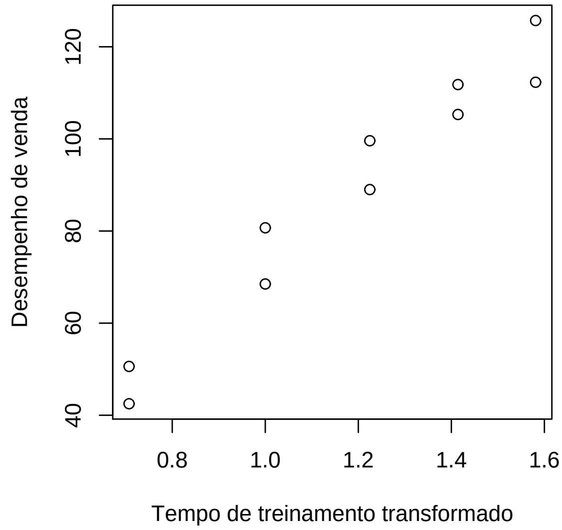
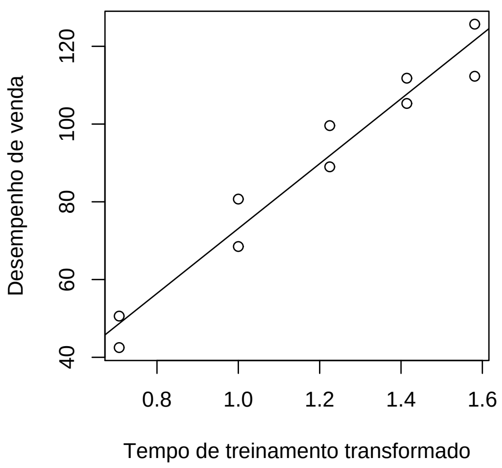


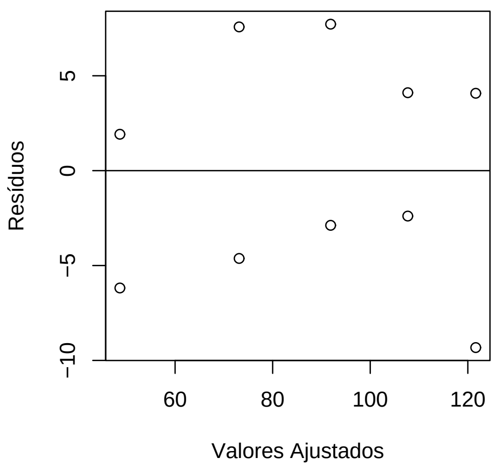
Figura 24: Diagrama de dispersão do Desempenho de venda vs Tempo de treinamento transformado

```
> ajuste2 = lm(Desempenho~Tempotrans,data=treino_venda_trans)
> summary(ajuste2)

Call:
lm(formula = Desempenho ~ Tempotrans, data = treino_venda_trans)

Residuals:
    Min      1Q  Median      3Q     Max
-9.3221 -4.1884 -0.2367  4.1007  7.7200

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -10.328      7.892  -1.309    0.227
Tempotrans    83.453      6.444  12.951 1.20e-06
---
Residual standard error: 6.272 on 8 degrees of freedom
Multiple R-squared: 0.9545,     Adjusted R-squared: 0.9488
F-statistic: 167.7 on 1 and 8 DF,  p-value: 1.197e-06

> shapiro.test(residuals(ajuste2))

        Shapiro-Wilk normality test

data:  residuals(ajuste2)
W = 0.9403, p-value = 0.5566
```

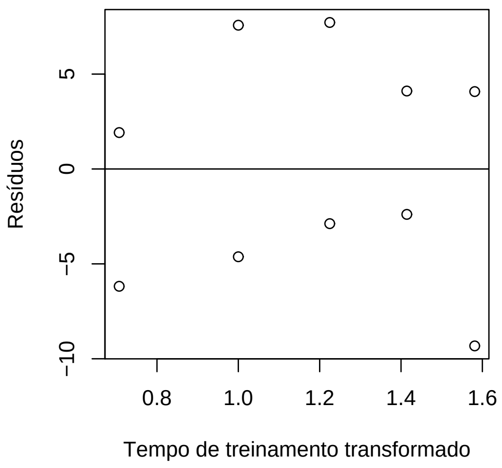
```
> windows()
> par(mfrow = c(2, 2))
> plot(Desempenho~Tempotrans,xlab="Tempo transformado",ylab="Desempenho de venda")
> abline(ajuste2)
> plot(fitted(ajuste2), residuals(ajuste2),xlab="Valores Ajustados",ylab="Resíduos")
> abline(h=0)
> plot(Tempotrans, residuals(ajuste2),xlab="Tempo transformado",ylab="Resíduos")
> abline(h=0)
> hist(residuals(ajuste2),main="",xlab="Resíduos",ylab="Freqüência")
```



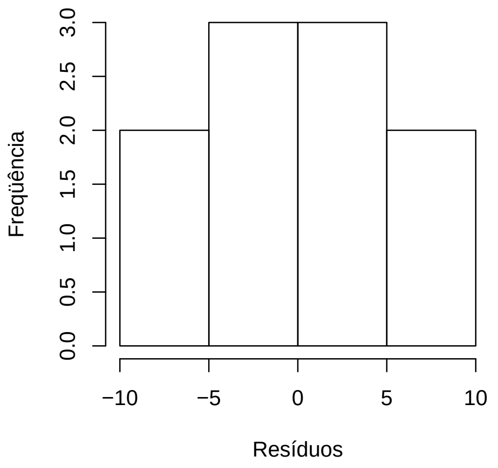
(a) Resposta vs explicativa transformada com reta de regressão



(b) Resíduos vs Valores ajustados



(c) Resíduos vs Variável explicativa transformada



(d) Histograma dos Resíduos

Figura 25: Gráficos para Análise dos Resíduos

A equação da reta ajustada é dada por $\hat{Y} = -10.328 + 83.453X_i'$, com $R^2 ajustado = 0,95$ (maior que o anterior). Observe da Figura 25(a) que a reta de regressão agora acompanha bem todos os pontos, indicando que a linearidade entre X e Y foi alcançada. A análise dos resíduos mostrada nas Figuras 23(b), (c) e (d) indica um bom ajuste do modelo, assim como o Teste de normalidade de *Shapiro-Wilk*, cujo *P-valor* é 0,5566.

9 Seleção de Variáveis Explicativas

No arquivo *prostata.txt*, estão nove variáveis relacionadas a um estudo com 97 homens com câncer de próstata. Estas variáveis constituem o banco de dados *prostate* da biblioteca *faraway* do *R*. A variável resposta é o *Logaritmo do antigénio específico da próstata*, denotada genericamente por Y, e as demais são candidatas à preditoras.

```
> prostata = read.table("prostata.txt",h=T)
> attach(prostata)
> prostata
         Y         X1     X2 X3        X4 X5       X6 X7 X8
1 -0.43078 -0.5798185 2.7695 50 -1.386294  0 -1.38629  6  0
2 -0.16252 -0.9942523 3.3196 58 -1.386294  0 -1.38629  6  0
3 -0.16252 -0.5108256 2.6912 74 -1.386294  0 -1.38629  7 20
.
.
.
> prostata[,6]= factor(prostata[,6])
```

A variável X_5 é dicotômica. Assim, foi definida como um fator.

Como há muitas variáveis candidatas a preditoras, pode-se recorrer a procedimentos automáticos de seleção de variáveis explicativas para auxiliar na escolha do melhor subconjunto delas.

9.1 *Stepwise* via AIC

Uma maneira de escolher entre as 2^8 regressões possíveis é utilizar os procedimentos *Backward* (Inclusão passo atrás), *Forward*(Inclusão passo a frente) ou uma combinação de ambos - o *Stepwise* (Seleção passo-a-passo).

Comumente, estes procedimentos automáticos avaliam em cada passo os p-valores das preditoras em comparação a um $\alpha_{critico}$. No entanto, no `R`, não há nenhuma função que considere o critério do p-valor. A função `stepwise` considera em cada passo os critérios AIC (*Akaike Information Criterion*) ou BIC (*Bayes Information Criterion*). A seguir estão os comandos das funções na direção *backward*, *forward* ou em ambas, considerando o critério AIC. Note que é necessário definir os modelos *nulo* e *cheio*, pois são argumentos dessas funções.

```
> nulo = lm(Y~1,data=prostata)
> completo = lm( Y~.,data=prostata)

> step(completo, data=prostata, direction="backward",trace=FALSE)
Call:
lm(formula = Y ~ X1 + X2 + X3 + X4 + X5, data = prostata)
Coefficients:
(Intercept)          X1          X2          X3          X4         X51
    0.95100     0.56561     0.42369    -0.01489     0.11184     0.72095
```

```
> step(nulo, scope=list(lower=nulo, upper=completo),data=prostata,
+ direction="forward",trace=FALSE)
Call:
lm(formula = Y ~ X1 + X2 + X5 + X4 + X3, data = prostata)
Coefficients:
(Intercept)           X1           X2          X51           X4           X3
    0.95100      0.56561      0.42369      0.72095      0.11184     -0.01489

> step(completo, data=prostata, direction="both",trace=FALSE)
Call:
lm(formula = Y ~ X1 + X2 + X3 + X4 + X5, data = prostata)
Coefficients:
(Intercept)           X1           X2           X3           X4          X51
    0.95100      0.56561      0.42369     -0.01489      0.11184      0.72095
```

Se colocarmos o argumento `trace=TRUE`, os passos intermediários dos procedimentos serão exibidos. Note que os três procedimentos indicaram o modelo com as explicativas X_1, X_2, X_3, X_4 e X_5 como o melhor segundo o critério AIC.

```
> summary(lm(Y ~ X1 + X2 + X3 + X4 + X5, data = prostata))

Call:
lm(formula = Y ~ X1 + X2 + X3 + X4 + X5, data = prostata)

Residuals:
      Min        1Q    Median        3Q       Max
-1.835049 -0.393961  0.004139  0.463365  1.578879

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.95100    0.83175   1.143 0.255882
X1           0.56561    0.07459   7.583 2.77e-11
X2           0.42369    0.16687   2.539 0.012814
X3          -0.01489    0.01075  -1.385 0.169528
X4           0.11184    0.05805   1.927 0.057160
X51          0.72095    0.20902   3.449 0.000854
---
Residual standard error: 0.7073 on 91 degrees of freedom
Multiple R-squared: 0.6441,     Adjusted R-squared: 0.6245
F-statistic: 32.94 on 5 and 91 DF,  p-value: < 2.2e-16
```

No entanto, observe que a variável X_3 não é significativa e deve ser retirada do modelo. A saída não será mostrada aqui, mas retirando a variável X_3, X_4 deixa de ser significativa. Assim, o modelo final ficaria apenas com as variáveis X_1, X_2 e X_5. Isso demonstra que os métodos automáticos auxiliam na escolha das variáveis, mas não devem substituir o bom senso e o conhecimento do problema.

9.2 Melhores Subconjuntos (*Best subsets*)

Ao invés de listar todas as possíveis regressões com as 8 variáveis candidatas a preditoras, a função `regsubsets` lista os k melhores modelos, segundo o critério de menor soma de quadrados residual, para subgrupos de preditoras de todos os tamanhos (desde uma até oito variáveis explicativas).

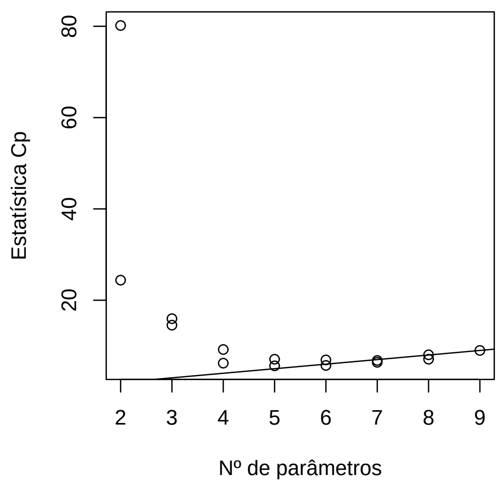
Além de mostrar as melhores regressões, a função também retorna os valores da soma de quadrados residual (s^2), da Estatística C_p de Mallows e de R^2 ajustado para cada regressão. Essas medidas podem ser utilizadas como critérios de seleção destes modelos, pois deseja-se modelos com o número p de parâmetros pequeno e s^2 pequeno, R^2 ajustado alto e Estatística C_p de Mallows com valor próximo de p.

Para exemplificar, encontraremos a seguir as duas melhores regressões ($k = 2$) para cada tamanho. Veja que k é o argumento `nbest` desta função.

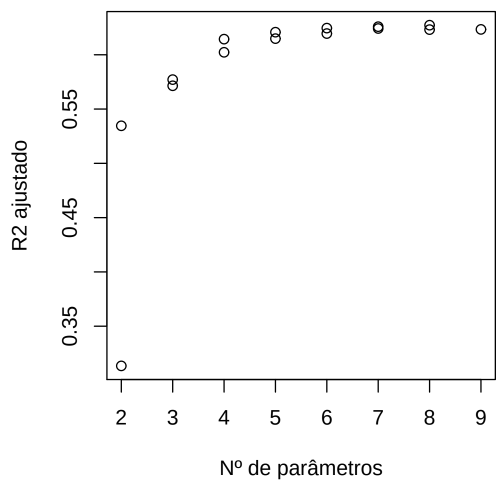
```
> require(leaps)
> rs = summary(regsubsets(Y~.,nbest=2,data=prostata))
> rs
Subset selection object
Call: regsubsets.formula(Y ~ ., nbest = 2, data = prostata)
8 Variables  (and intercept)
    Forced in Forced out
X1      FALSE      FALSE
X2      FALSE      FALSE
X3      FALSE      FALSE
X4      FALSE      FALSE
X51     FALSE      FALSE
X6      FALSE      FALSE
X7      FALSE      FALSE
X8      FALSE      FALSE
2 subsets of each size up to 8
Selection Algorithm: exhaustive
         X1  X2  X3  X4  X51 X6  X7  X8
1  ( 1 ) "*" " " " " " " " " " " " " " "
1  ( 2 ) " " " " " " " " "*" " " " " " "
2  ( 1 ) "*" "*" " " " " " " " " " " " "
2  ( 2 ) "*" " " " " " " "*" " " " " " "
3  ( 1 ) "*" "*" " " " " "*" " " " " " "
3  ( 2 ) "*" " " " " "*" "*" " " " " " "
4  ( 1 ) "*" "*" " " "*" "*" " " " " " "
4  ( 2 ) "*" "*" " " " " "*" " " " " "*"
5  ( 1 ) "*" "*" "*" "*" "*" " " " " " "
5  ( 2 ) "*" "*" " " "*" "*" " " " " "*"
6  ( 1 ) "*" "*" "*" "*" "*" " " " " "*"
6  ( 2 ) "*" "*" "*" "*" "*" " " "*" " "
7  ( 1 ) "*" "*" "*" "*" "*" "*" " " "*"
7  ( 2 ) "*" "*" "*" "*" "*" "*" "*" " "
8  ( 1 ) "*" "*" "*" "*" "*" "*" "*" "*"
```

As figuras a seguir, obtidas com os comandos abaixo, auxiliam na escolha do melhor modelo:

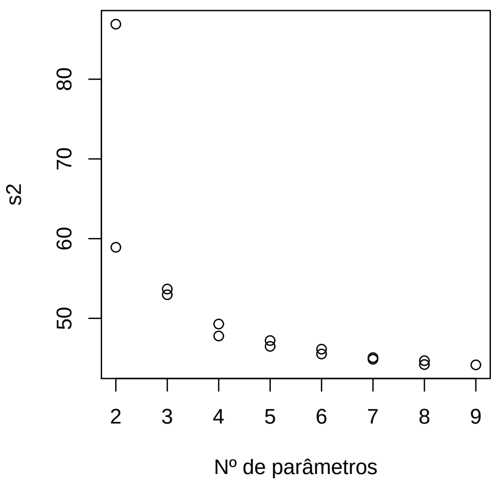
```
> n_parametros = as.numeric(rownames(rs$which))+1
> Cp = rs$cp
> R2_ajustado = rs$adjr2
> s2 = rs$rss
> plot(n_parametros,Cp,xlab="Nº de parâmetros",ylab="Estatística Cp")
> abline(0,1)
> plot(n_parametros,s2,xlab="Nº de parâmetros",ylab="s2")
> plot(n_parametros,R2_ajustado,xlab="Nº de parâmetros",ylab="R2 ajustado")
```



(a) Estatística C_p vs Nº de parâmetros



(b) R^2 ajustado vs Nº de parâmetros



(c) s^2 vs Nº de parâmetros

Figura 26: Gráficos para seleção de modelos

Nota-se da Figura 26 (a) que as Estatísticas C_p de Mallows já são próximas de p quando $p = 5$ ou $p = 6$. Das Figuras 26 (b) e (c), observa-se que o incremento em R^2 ajustado e o decremento em s^2 são bem pequeno ao passar de $p = 5$ para $p = 6$. Assim, tendo em vista um modelo mais parcimonioso, poderíamos escolher o valor de $p = 5$.

```
> n_variaveis = n_parametros-1
> cbind(n_variaveis,n_parametros,Cp,R2_ajustado,s2)
      n_variaveis n_parametros        Cp R2_ajustado       s2
 [1,]           1            2 24.394559   0.5345838 58.91476
 [2,]           1            2 80.172023   0.3134515 86.90682
 [3,]           2            3 14.541475   0.5771246 52.96626
 [4,]           2            3 15.958255   0.5714480 53.67727
 [5,]           3            4  6.216935   0.6143899 47.78486
 [6,]           3            4  9.208478   0.6022748 49.28617
 [7,]           4            5  5.626422   0.6208036 46.48480
 [8,]           4            5  7.074224   0.6148766 47.21139
 [9,]           5            6  5.715016   0.6245476 45.52556
[10,]           5            6  6.922392   0.6195505 46.13149
[11,]           6            7  6.401965   0.6258707 44.86660
[12,]           6            7  6.806372   0.6241784 45.06956
[13,]           7            8  7.082184   0.6272521 44.20427
[14,]           7            8  8.047624   0.6231666 44.68878
[15,]           8            9  9.000000   0.6233681 44.16302
```

Dentre os dois modelos com 5 parâmetros, o primeiro, constituído pelas variáveis X_1, X_2, X_4 e X_5, seria o mais adequado segundo todos os critérios.

10 Multicolinearidade

Em um estudo sobre a relação entre a *quantidade de gordura corporal* (Y) e as medidas de *espessura de dobra do tríceps* (X_1), *circunferência da coxa* (X_2) e *circunferência do antebraço* (X_3), foram coletados dados de uma amostra de 20 mulheres saudáveis entre 20 e 35 anos (Tabela 7).

Tabela 7: Medidas corporais relativas à gordura em 20 mulheres saudáveis

Tríceps(X_1)	Coxa(X_2)	Antebraço(X_3)	Gordura(Y)
22.5	26.8	8.7	22.9
20.1	24.0	7.6	19.8
18.4	22.1	7.5	8.0
22.8	27.2	8.8	28.6
21.4	25.2	8.5	19.2
21.4	25.7	8.4	20.6
19.9	23.9	8.0	18.3
22.2	26.5	8.8	30.1
20.1	23.9	8.3	16.9
19.3	23.0	7.7	12.7
20.6	24.5	8.0	21.8
18.7	22.1	7.5	13.9
21.6	25.7	8.3	22.6
19.8	23.6	8.0	18.1
21.3	25.6	8.6	22.7
21.3	25.4	8.1	20.8
20.8	24.8	7.7	25.4
19.5	23.0	7.7	11.5
20.1	23.9	8.2	11.7
20.7	24.7	8.2	15.3

Para a análise dos dados no `R`, suponha que a Tabela 7 esteja contida no seguinte arquivo de texto *gordura.txt*, da seguinte forma:

```
> medidas_corp = read.table("gordura.txt",header=T)
> attach(medidas_corp)
> medidas_corp
   Triceps Coxa Antebraco Gordura
1     22.5 26.8       8.7    22.9
2     20.1 24.0       7.6    19.8
3     18.4 22.1       7.5     8.0
.
.
.
```

Objetiva-se investigar se há multicolinearidade entre as variáveis explicativas, visto que a forte correlação entre elas acarreta vários efeitos negativos no ajuste do modelo de regressão. Alguns métodos exploratórios e formais de detecção e/ou mensuração da multicolinearidade serão aplicados neste exemplo.

10.1 Métodos Informais de Diagnóstico

Algumas indicações da presença de multicolinearidade são:

(1) Os coeficientes da correlação linear entre pares de variáveis apresentam valores muito próximos de 1 ou -1. Na Seção 2.3, o comando necessário foi aplicado somente a um par de variável. Se há muitas variáveis explicativas no modelo, uma outra maneira de utilizar aquele comando é útil:

```
> explicativas = medidas_corp[,1:3]
> cor(explicativas)
            Triceps      Coxa Antebraco
Triceps   1.0000000 0.9956813 0.8776755
Coxa      0.9956813 1.0000000 0.8773968
Antebraco 0.8776755 0.8773968 1.0000000
```

Observa-se que todas as variáveis explicativas são altamentes correlacionadas, de forma linear.

(2) Os gráficos de dispersão entre pares de variáveis apresentam configurações especiais, indicando algum tipo de relação entre elas. Na Seção 2.2, o comando necessário foi aplicado somente a um par de variável. Se há muitas variáveis explicativas no modelo, outro comando é útil:

```
> pairs(explicativas)
```

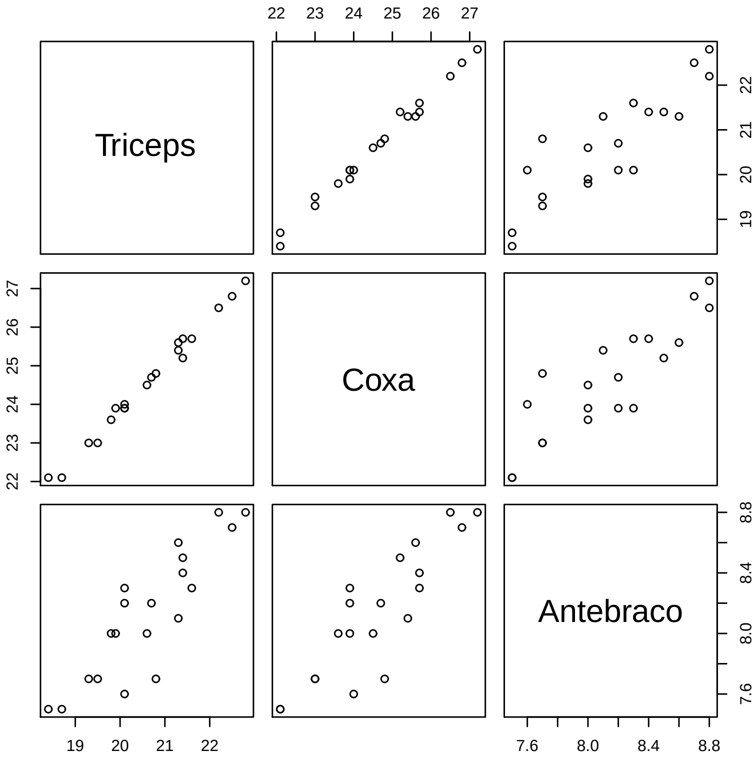


Figura 27: Diagramas de dispesão dos pares de variáveis explicativas

Conforme já verificado no item 1, observa-se da Figura 27 um relacionamento linear forte e positivo entre todos os pares de variáveis explicativas, em especial entre o par Tríceps e Coxa.

(3) Os coeficientes de regressão apresentam sinais algébricos opostos ao esperado a partir do conhecimento teórico. Neste caso, esperaria-se o sinal positivo para todos os coeficientes, pois quanto maior as medidas da espessura de dobra do tríceps, circunferência da coxa e circunferência do antebraço, espera-se uma maior quantidade de gordura corporal.

```
> ajuste = lm(Gordura ~ Triceps + Coxa + Antebraco,data = medidas_corp)
> summary(ajuste)

Call:
lm(formula = Gordura ~ Triceps + Coxa + Antebraco, data = medidas_corp)

Residuals:
     Min       1Q   Median       3Q      Max 
-3.95086 -2.34008 -0.02048  2.13954  5.51043 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -54.1322    13.3181  -4.065 0.000901
Triceps      -0.4802     6.1750  -0.078 0.938979
Coxa          5.1668     5.1096   1.011 0.326973
Antebraco    -5.4021     3.3996  -1.589 0.131610
---
Residual standard error: 2.996 on 16 degrees of freedom
Multiple R-squared: 0.7714,     Adjusted R-squared: 0.7285 
F-statistic:    18 on 3 and 16 DF,  p-value: 2.223e-05
```

(4) O teste F da significância da regressão rejeita $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$ (p-valor igual a 2.223e-05) , mas nenhuma das $H_0 : \beta_j = 0$, $j = 1, 2, 3$, é rejeitada pelos testes t individuais.

(5) Os erros padrão (na saída, a coluna `Std.Error`) dos coeficientes de regressão são muito altos.

10.2 Método Formal de Diagnóstico

Visto que os métodos anteriores são limitados, para detectar e medir a multicolineridade de maneira formal, a análise dos fatores de inflação da variância (em inglês, *Variance Inflation Factors - VIF*) é feita a seguir. Esta função está no pacote `faraway`, o qual deve ser, portanto, instalado e carregado.

```
> require(faraway)
> vif(explicativas)
   Triceps       Coxa  Antebraco 
117.001244 116.752602   4.381068 
```

Draper e Smith (1998) recomendam que valores de VIF maiores do que 10 podem causar sérios problemas na estimação dos coeficientes de regressão. Observa-se valores muito elevados para VIF_1 e VIF_2, confirmando a presença da multicolinearidade.

11 Regressão Sem Intercepto

Em determinadas situações, não faz sentido ajustar um modelo com intercepto. No exemplo anterior, se as medidas da espessura de dobra do tríceps, circunferência da coxa e circunferência do antebraço são nulas, então o valor esperado da gordura corporal também é zero. Assim, para ajustar um modelo sem intercepto, basta adicionar a constante -1 no comando comumente utilizado:

```
> ajuste = lm(Gordura~ -1 + Triceps + Coxa + Antebraco,data = medidas_corp)
> summary(ajuste)

Call:
lm(formula = Gordura ~ -1 + Triceps + Coxa + Antebraco, data = medidas_corp)

Residuals:
    Min      1Q  Median      3Q     Max
-7.5188 -3.1045  0.1102  2.0517  9.7035

Coefficients:
          Estimate Std. Error t value Pr(>|t|)
Triceps     -3.493      8.479  -0.412   0.6855
Coxa         7.589      7.019   1.081   0.2947
Antebraco  -11.724      4.181  -2.804   0.0122
---
Residual standard error: 4.143 on 17 degrees of freedom
Multiple R-squared: 0.963,      Adjusted R-squared: 0.9564
F-statistic: 147.4 on 3 and 17 DF,  p-value: 2.294e-12
```

Referências

[1] Draper, N. R. ; Smith, H. Applied Regression Analysis, 3a. edição. John Wiley and Sons, EUA, 706 p, 1998.

[2] FARAWAY, J. J. *Linear Models with R*.

[3] "Material da Disciplina Análise de Regressão". Disponível em: `http://www.est.ufmg.br/~edna/regressao.htm`