Regressão linear Simples. Aula 2 minicurso

10 de outubro de 2013

I congreso Internacional de Estadística. Trujillo-Perú. 2013

Resumo

Nesta unidade apresentaremos o modelo de regressão linear simples, e os modelos de regressão não linear (nos parâmetros), as equações dos modelos, suas propriedades e fundamentos, os pressupostos básicos para as variáveis independentes e para a variável dependente, assim como para o erro aleatório. As variáveis são fatores de qualquer fenômeno, podemos classificar-las em endógenas (dependentes) e exógenas (independentes), as endógenas são aquelas que recebem influência de outras variáveis, na área econométrica chamadas também de variáveis efeito. As variáveis independentes chamadas de exógenas, são aquelas que afetam o controlam a variável dependente.

1 Objetivo

De uma maneira geral, a partir dos dados (observações) realizar inferências sobre uma população. Em particular estudar o comportamento de uma variável (dependente) quando esta se relaciona com o comportamento de uma ou outras variáveis (independentes).

2 Definição

Quando falamos de modelo linear estamos tratando de modelos que são lineares nos parâmetros (Gujarati, 2000), assim um modelo de regressão linear simples ou análise de regressão de duas variáveis considere como variável independente X e como variável dependente Y de tal forma que, admitindo que $E(Y|X)$ seja linear em X, possamos escrever $Y_i = E(Y|X_i) + \epsilon_i$ ou

$$Y_i = \beta_1 + \beta_2 X_i + \epsilon_i, \tag{3.1}$$

onde ϵ_i é definido como termo de erro estocástico ou perturbação estocástica com distribuição normal com média 0 e variância σ^2, β_1 definido como parâmetro do intercepto, e β_2 como o parâmetro de inclinação (ou tangente). Com esta definição, são relevantes:

1. Linearidade de impacto: X deve ser a média de Y.
2. A variância é constante ou homoscedástica.

Para que o modelo seja tratável é necessário algumas suposições:

1. $Y_i = \beta_1 + \beta_2 X_i$ linearidade do modelo
2. $v(Y) = \sigma^2 \ (0 < \sigma^2 < \infty), \qquad \forall X$
3. Cada Y é não correlacionado com os demais
4. X deve assumir pelo menos dois valores distintos
5. Y possuí distribuição normal (nem sempre necessário)
6. O erro aleatório é dado por $\epsilon = Y_i - \beta_1 - \beta_2 X_i$

Até aqui temos definido o modelo populacional, porém quando nos limitamos a questões práticas surge a ideia de usar uma amostragem de valores de Y correspondentes a alguns valores fixos de X e nossa equação (3.1) será dada pela reta de regressão amostral e pode ser escrita como:

$$y_i = \widehat{\beta}_1 + \widehat{\beta}_2 x_i + \widehat{\epsilon}_i \tag{3.2}$$

A estimativa $\widehat{y}$ de Y é definida como:

$$\widehat{y}_i = \widehat{\beta}_1 + \widehat{\beta}_2 x_i$$

Onde $\widehat{\beta}_1$, $\widehat{\beta}_2$, são estimativas dos parâmetros acima descritos, $\widehat{\epsilon}_i$ é a estimativa do erro (ou perturbação) denominado de resíduo aleatório.

Inicialmente são exigidos os seguintes pressupostos básicos:

1. **O modelo é linear**, nos parâmetros.
2. **Normalidade**, Os ϵ_i's têm distribuição normal (Nem sempre necessario).
3. **Média Zero**, $E(\epsilon_i) = 0$.

4. **Homoscedasticidade**, $V(\epsilon_i) = \sigma^2$, a variabilidade dos erros é constante.
5. **Não colinearidade significativa** a pares nos regressores.
6. **Independência das perturbações** ou ausência de autocorrelação, $E(\epsilon_i, \epsilon_j) = 0$.
7. x **é determinística**.
8. **n ¿ p** O número de observações (n) tem que ser maior que o número de parâmetros (p).
9. **Covariância zero** entre ϵ_i e x_i, ou $E(\epsilon_i x_i) = 0$.
10. **Correta especificação** do modelo de regressão.
11. **Nenhum erro de medida nos x's**. As variáveis explicativas são medidas sem erro. Para maiores detalhes ver Gujarati, 2000.

Nossa tarefa agora é estimar a função de regressão. Aqui, apresentaremos dois métodos que são: *o método dos mínimos quadrados ordinários (MQO)* e o *método de máxima verossimilhança (MMV).*

3 Mínimos Quadrados Ordinários

Para estimar os parâmetros de regressão, inicialmente expressamos a perturbação como

$$\widehat{\epsilon_i} = y_i - \widehat{y_i} = y_i - \widehat{\beta}_1 - \widehat{\beta}_2 x_i. \tag{3.3}$$

Desta forma queremos determinar uma reta que esteja tão próxima quanto possível de Y real. Para tanto, teremos que minimizar a soma dos erros quadráticos

$$\sum \widehat{\epsilon_i}^2 = \sum (y_i - \widehat{\beta}_1 - \widehat{\beta}_2 x_i)^2 \tag{3.4}$$

de modo a torná-la menor possível. Diferenciando em relação aos parâmetros β_1 e β_2 do modelo, obtemos as seguintes equações:

$$\sum y_i = n\widehat{\beta}_1 + \widehat{\beta}_2 \sum x_i, \tag{3.5}$$

$$\sum y_i x_i = \widehat{\beta}_1 \sum x_i + \widehat{\beta}_2 \sum x_i^2 \tag{3.6}$$

Resolvendo essas equações simultaneamente, obtemos

$$\widehat{\beta}_2 = \frac{n\sum y_i x_i - \sum y_i \sum x_i}{n\sum x_i^2 - (\sum x_i)^2}, \tag{3.7}$$

$$\widehat{\beta}_1 = \bar{y} - \widehat{\beta}_2 \bar{x} \tag{3.8}$$

Como y é variável aleatória, se tiver distribuição normal, $\hat{\beta}_i$'s, são variáveis aleatórias e tem distribuição normal também. Se Y não tiver distribuição normal, $\hat{\beta}_i$'s serão aproximadamente normal, se o tamanho da amostra for grande.

4 Máxima Verosimilhança

Um outro método apresentado na literatura é o *método de máxima verossimilhança* (ver Carneiro, 1998). Considere o modelo em (3.1). De acordo com os pressupostos de validade de um modelo econométrico, y é um variável aleatória com distribuição normal, média $(\beta_1 + \beta_2 x)$ e variância σ^2. Considerando que os valores observados $y_1, y_2, \cdots, y_n$ são independentes, e n é o total de observações, então a função de probabilidade conjunta a partir da distribuição normal é dada pela função de verossimilhança:

$$\begin{aligned} L &= p(y_1)p(y_2)\cdots p(y_n) = L(y_1, y_2, \cdots, y_n, \beta_1, \beta_2, \sigma^2) \\ &= \prod_{i=1}^{n} 1/(2\pi\sigma^2)^{1/2} \exp[-1/2\sigma^2 (y_i - \beta_1 - \beta_2 x_i)^2] \end{aligned} \tag{3.9}$$

Para estimar os parâmetros β_1 e β_2 minimizamos a função de verossimilhança. Para facilitar os cálculos utilizamos a expressão

$$LogL = -\frac{n}{2}\log 2\pi - n\log\sigma - \frac{1}{2\sigma^2}\sum (y_i - \beta_1 - \beta_2 x_i)^2. \tag{3.10}$$

Desse modo, quando as derivadas parciais de LogL em relação aos parâmetros a serem estimados, $\widehat{\beta}_1$, $\widehat{\beta}_2$ e $\widehat{\sigma}^2$, são igualadas a zero, obtemos as equações (3.7) e (3.8).
Observação: para estimar os β_i's usando máxima verosimilhança precisamos uma distribuição para os erros aleatórios, no caso uma normal, já mínimos quadrados esta restrição não é necessária. Assim os EMV dos β_i's sob normalidade são exatamente os $\hat{\beta}_i$'s em MQO.

5 Teorema de Gauss - Markov

Considerando certas, as primeiras 5 suposições, $\widehat{\beta}$ é o melhor estimador linear não tendencioso (ou não-viesado) de β. Assim $\widehat{\beta}$ é MELNT de β.

6 Covariância de β

A matriz de variâncias e covariâncias dos parâmetros estimados é dado por:

$$cov(\hat{\beta}) = \begin{bmatrix} \frac{\sigma^2 \sum x_i^2}{\sum (x_i - \bar{x})^2} & \frac{-\sigma^2 \bar{x}}{\sum (x_i - \bar{x})^2} \\ \frac{-\sigma^2 \bar{x}}{\sum (x_i - \bar{x})^2} & \frac{\sigma^2}{\sum (x_i - \bar{x})^2} \end{bmatrix} \tag{3.11}$$

7 Previsões

Suponha que o interesse é prever o valor da variável dependente correspondente a um dado valor do regressor, digamos x_0, então:

$$y_0 = \beta_1 + \beta_2 x_0 + \epsilon_0 \tag{3.12}$$

cuja previsão é:

$$\hat{y}_0 = \hat{\beta}_1 + \hat{\beta}_2 x_0 \tag{3.13}$$

Onde o erro e previsão é dado por $y_0 - \hat{y}_0$, e o calculo da esperança é:

$$E(\beta - 1 - \beta_2 x_0 + \epsilon_0 - \hat{\beta}_1 - \hat{\beta}_2 x_0) = (\beta_1 - E(\hat{\beta}_1)) + x_0(\beta_2 - E(\hat{\beta}_2)) = 0 \tag{3.14}$$

Então a previsão não possui viés, ainda a variância do erro de previsão é dada por:

$$V(y_0 - \hat{y}_0) = \sigma[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2}] \tag{3.15}$$

Esta variância pode ser estimada por $\hat{V}(y_0 - \hat{y}_0)$, substituindo na equação: σ por $\hat{\sigma}$. $V(y_0 - \hat{y}_0)$ é mínima em $x_0 = \bar{x}$. Um intervalo de precisão de nível $1 - \alpha$ para y_0 é:

$$\hat{y}_0 \pm t_{\alpha/2, n-2} \sqrt{V(y_0 - \hat{y}_0)} \tag{3.16}$$

A qual é a previsão da variável aleatória y_0.
Suponha que desejamos inferir o valor médio da variável dependente correspondente a um dado valor do regressor, digamos x_0. Temos:

$$\eta_0 = E(y_0) = \beta_1 + \beta_2 x_0 \tag{3.17}$$

então, a esperança do erro é

$$E(erro) = E(\eta_0 - \hat{y}_0) = E(\beta_1 - \hat{\beta}_1 + x_0(\beta_2 - \hat{\beta}_2)) = 0 \tag{3.18}$$

e a variância é dada por:

$$V(erro) = V(\eta_0 - \hat{y}_0) = V(\hat{y}_0) = \sigma^2[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum(x_i - \bar{x})^2}] \tag{3.19}$$

que pode ser estimado por:

$$\hat{V}(\hat{y}_0) = \hat{\sigma}^2[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum(x_i - \bar{x})^2}] \tag{3.20}$$

Um intervalo de precisão de nível $1 - \alpha$ para y_0 é:

$$\hat{y}_0 \pm t_{\alpha/2,n-2}\sqrt{V(\hat{y}_0)} \tag{3.21}$$

cujo valor é a previsão para a média η_0.

8 Correlação

Regressão e correlação, estão relacionados da seguinte maneira, no caso da regressão estamos interessados nas estimativas dos parâmetros do modelo, já na correlação estamos interessados no grau (e em que direção) de intensidade na relação entre a variável dependente com a independente.

8.1 Coeficiente de correlação

$$r = \frac{\sum X_i Y_i}{\sqrt{(\sum X_i^2)(\sum Y_i^2)}} \qquad X_i = X_i - \bar{X}, \qquad y_i = Y_i - \bar{Y} \tag{3.22}$$

O coeficiente de correlação é um número entre -1 e 1, e de fácil interpretação indica que quando se aproxima de -1 a relação é forte e inversamente proporcional (Quando a variável independente aumenta a dependente diminui) e no caso de se aproximar a 1, a relação é forte e diretamente proporcional. Quando este coeficiente se aproxima de zero esta relação é fraca. Esta correlação é denominada de correlação de Pearson, e no R, é usada com o seguinte comando:

```
>cor(x,y)  #valor da relação entre x e y
```

8.2 Teste de correlação

Para determinar se existe ou não relação entre duas variáveis causais, usamos o teste de hipótese do coeficiente de correlação. Os n pares de valores (X,Y) de duas variáveis pode ser pensado como amostras de uma população de todos os pares possíveis. Uma vez que duas variáveis são envolvidas, isto é chamado de uma população bivariada, a qual supomos que segue uma distribuição normal bivariada. Podemos pensar em uma população teórica de coeficientes de correlação, denotados por ρ, a qual é estimada por um coeficiente de correlação $\hat{\rho}$.

8.2.1 Hipótese

Para a hipótese nula, $H0 : \rho = 0$, assumimos que este parâmetro tem distribuição simétrica.
A hipótese alternativa, $Ha : \rho \neq 0$

8.2.2 Nível de significância

Considere $\alpha = 0.05$, num teste bicaudal

8.2.3 Estatística de prova

A estatística envolvendo a distribuição de Student pode ser usada. Para $\rho \neq 0$, a distribuição é "assimetrica"(alongada). Nestes casos uma transformação desenvolvida por Fisher produz uma estatística que é aproximadamente normalmente distribuída. A estatística de teste é dada por:

$$t_0 = \hat{\rho} * \sqrt{n - 2/(1 - \rho^2)}$$

que segue aproximadamente uma distribuição t com $n-2$ graus de liberdade.

8.2.4 Regra de deisão

A regra é se $t_0 > \left|t_{(n-2)}(0.975)\right|$, não há evidencias para aceitar H0.
Uma forma alternativa é usando o $p - valor$; se o $p - valor < 0.05$, não há evidencias para aceitar H0. No R a função usada é `cor.test()`, dado por:

```
>cor.test(x,y) # teste para verificar a significância da correlação.
```

Tudo isto é valido sobre o pressuposto de conhecer a distribuição das variáveis (por exemplo normalidade). Contudo existe uma variante não paramétrica

quando não dependemos da distribuição das variáveis. Há duas variantes não paramétricas como a correlação de Spearman (ρ), baseada nos ranks de correlação; a outra de Kendall(τ) é baseada na contagem de número de pares concordantes e discordantes. No R é implementado da seguinte maneira:

```
>cor(x,y) , method="spearman") #correlação de spearman

>cor(x,y) , method="kendall") #correlação de kendall

>cor.test(x,y, method="spearman") #significância da correlação de spearman

>cor.test(x,y, method="kendall") #significância da correlação de kendall
```

Exemplo, no consumo de água em uma residência, para encontrar a relação entre valor e consumo:

```
> cor(valor,consumo)
[1] 0.9458655

> cor(valor,consumo,method="spearman")
[1] 0.9270707

> cor(valor,consumo,method="kendall")
[1] 0.8078577

> cor.test(valor,consumo) # Teste de Pearson

        Pearson's product-moment correlation

data:  valor and consumo
t = 23.1316, df = 63, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.9124691 0.9667414
sample estimates:
      cor
0.9458655
```

```
> cor.test(valor,consumo, method="spearman")

        Spearman's rank correlation rho

data:  valor and consumo
S = 3337.245, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.9270707

> cor.test(valor,consumo, method="kendall")

        Kendall's rank correlation tau

data:  valor and consumo
z = 8.9711, p-value < 2.2e-16
alternative hypothesis: true tau is not equal to 0
sample estimates:
      tau
0.8078577
```

8.3 Coeficiente de determinação

É o quadrado do coeficiente de correlação, e pode ser interpretado em percentual, assim mede em percentual quanto X explica Y

$$R^2 = \left\{ \frac{\sum x_i y_i}{\sqrt{(\sum x_i^2)(\sum y_i^2)}} \right\}^2 \tag{3.23}$$

Definição 1 *Um estimador $\hat{\theta}$ num espaço paramétrico ($\theta \in \Theta$), é dito ser não-viesado, se $E(\hat{\theta}) - \theta = 0, \ \forall \theta \in \Theta$*

9 Estimação de σ^2

O estimador de mínimos quadrados ordinários (EMQO) de σ^2 é dado por:

$$\hat{\sigma}^2 = \sum \hat{\epsilon}^2 / n - 2 \tag{3.24}$$

onde $\hat{\epsilon}$, são os resíduos dos modelo.
O estimador de máxima verosimilhança (EMV), assumindo normalidade dos erros, e dado por:

$$\tilde{\sigma}^2 = \sum \hat{\epsilon}^2/n \tag{3.25}$$

que é diferente de $\hat{\sigma}^2$, o EMQO. Como o EMV de σ^2 é viesado, quando n é grande, este estimador é assintoticamente não-viesado. Os estimadores de MQO é não viesado, pois

$$E(\hat{\sigma}^2) = \sigma^2, \qquad \forall\sigma^2 \tag{3.26}$$

Sobre normalidade temos que

$$\hat{\sigma}^2 \approx \frac{\sigma^2}{n-2} X^2_{n-2} \tag{3.27}$$

Asim:

$$E(\hat{\sigma}^2) = E(\frac{\sigma^2}{n-2} X^2_{n-2}) = \frac{\sigma^2}{n-2} E(X^2_{n-2}) = \sigma^2 \tag{3.28}$$

A variância é dada por:

$$V(\hat{\sigma}^2) = V(\frac{\sigma^2}{n-2} X^2_{n-2}) = \frac{\sigma^4}{(n-2)^2} V(X^2_{n-2}) = \frac{\sigma^4}{(n-2)^2} 2 * (n-2) = \frac{2\sigma^4}{n-2} \tag{3.29}$$

Note que se σ^2 é grande, maiores serão os estimadores, e a variância dos estimadores decresce quando n cresce.

10 Intervalo de confiança

Com a suposição de normalidade dos erros, os parâmetros do modelo tem distribuição normal:

$$\widehat{\beta}_1 \approx N(\beta_1, var(\widehat{\beta}_1)) \qquad e \qquad \widehat{\beta}_2 \approx N(\beta_2, var(\widehat{\beta}_2)) \tag{3.30}$$

Seja z_c, um número real, onde $P(Z > z_c) = P(Z < -z_c) = \alpha/2$, $0 < \alpha < 1$, e Z, tem distribuição normal padrão. Caso σ^2 ser conhecido podemos determinar o intervalo de confiança de β_i, para $i = 1, 2$ como:

$$P[-z_c \leq \frac{\widehat{\beta}_i - \beta_i}{\sqrt{var(\widehat{\beta}_i)}} \leq z_c] = 1 - \alpha \tag{3.31}$$

ou:

$$P[\widehat{\beta}_i - z_c\sqrt{var(\widehat{\beta}_i)} \leq \beta_i \leq \widehat{\beta}_i + z_c\sqrt{var(\widehat{\beta}_i)}] = 1 - \alpha \tag{3.32}$$

Em geral desconhecemos σ^2, por tanto substituímos o estimador de σ^2, por:

$$\widehat{\sigma}^2 = \frac{\sum(\hat{\epsilon}_i^2)}{n-2} \tag{3.33}$$

11 Análise de variância

Fontes de variação	Gl	SQ	QM	F	$p-valor$
Regressão	1	$\sum \hat{y}_i^2$	$\sum \hat{y}_i^2$	$\frac{SQReg}{SQRes}$	
Resíduo	$n-2$	$\sum \hat{\epsilon}_i^2$	$\sum \hat{\epsilon}_i^2/n-2 = \hat{\sigma}^2$		
Total	$n-1$	$\sum y_i^2$			

onde:

$SQReg$: significa soma de quadrados da regressão.
$SQRes$: significa soma de quadrados dos resíduos.
$SQTotal = SQReg + SQRes$
Uma forma alternativa de calcular R^2 usando o ANOVA é dado por:

$$R^2 = \frac{SQRes}{SQTotal} = 1 - \frac{\sum \hat{\epsilon}^2}{\sum y_i^2 - n\bar{y}^2}, \qquad 0 \leq R^2 \leq 1 \tag{3.34}$$

Observação: Se o modelo não tiver um intercepto, então $SQTotal \neq SQReg+ SQRes$, pois $\sum \hat{\epsilon}^2 \neq 0$, e

$$SQTotal = SQReg + vies + SQRes \tag{3.35}$$

Para modelos sem intercepto use a medida R^2 não centrado, dado por:

$$R^2_{nc} = 1 - \frac{\sum \hat{\epsilon}_i^2}{\sum y_i^2} \tag{3.36}$$

onde: y_i é uma variável centrada.

12 Exemplo 1.

Considere x uma variável independente e deterministica com valores iniciados em 1 e finalizados em 20 com espaçamento de 0.1, a variável dependente y, sera gerada dos valores de x acrescidas de um termo aleatório proveniente de uma distribuição normal padrão.

12.1 O modelo

Considere : $x = 1, 1.1, 1.2, \ldots, 19.9, 20; \qquad y = x + N(0,1) \quad (3.37)$

o modelo é dado por:

$$\widehat{Y}_i = \widehat{\beta}_1 + \widehat{\beta}_2 x_i$$

No R.

```
> set.seed(20)
> x=seq(1,20, by=0.1)
> y=x+rnorm(length(x))
> m1=lm(y~x)
```

12.2 A função lm()

É composta por vários comandos, para exemplificar considere u objeto m1:

1. O comando print

   Este comando apresenta o modelo proposto e as estimativas dos parâmetros do modelo

   ```
   > print(m1)      # ou simplesmente  >m1

   Call:
   lm(formula = y ~ x)

   Coefficients:

   (Intercept)            x
       -0.1749       1.0237
   ```

2. O comando summary
   Este comando apresenta um resumo de estatísticas do modelo de regressão.

   ```
   > summary(m1)
   ```

```
Call:
lm(formula = y ~ x)

Residuals:
     Min       1Q   Median       3Q      Max
-2.75273 -0.67023  0.05422  0.75149  2.14659

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.17487    0.15254  -1.146    0.253
x            1.02367    0.01286  79.589   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.9801 on 189 degrees of freedom
Multiple R-squared: 0.971,     Adjusted R-squared: 0.9709
F-statistic:  6334 on 1 and 189 DF,  p-value: < 2.2e-16
```

3. O comando coef
   Este comando apresenta simplesmente as estimativas dos parâmetros do modelo de regressão

```
> coef(m1)

(Intercept)           x
 -0.1748678   1.0236721
```

4. O comando residuals
   Apresenta os resíduos do modelo de regressão

```
> residuals(m1)

           1            2   ...          190          191
 1.313880999 -0.437095965   ... -1.501733127  0.381029987
```

5. O comando fitted
   Este comando calcula os valores estimados da função dependente ($\hat{y}$)

```
> fitted(m1)

         1          2  ...        190        191
 0.8488043  0.9511715  ... 20.1962067 20.2985739
```

6. O comando predict
   Para este comando precisamos identificar os valores que realizaremos as previsões, considere que os valores para previsão de x, sejam $1, 2$ e 3, o qual denominamos de **novo**:

```
> novo=data.frame(x=c(1,2,3))
> predict(lm(y ~ x), novo, se.fit = TRUE)

$fit
        1         2         3
0.8488043 1.8724764 2.8961485

$se.fit
        1         2         3
0.1412777 0.1303135 0.1197276

$df
[1] 189

$residual.scale
[1] 0.9800842
```

7. anova
   Este comando apresenta o análise de variância do modelo de regressão

```
> anova(m1)

Analysis of Variance Table
Response: y
           Df Sum Sq Mean Sq F value    Pr(>F)
x           1 6084.6  6084.6  6334.3 < 2.2e-16 ***
Residuals 189  181.5     1.0
```

8. O comando plot
   Este comando apresenta o gráfico das variáveis de interesse, considere o modelo m1, este objeto tem quatro gráficos internamente construídos, por tanto precisamos preparar a saída como o comando **par(mfrow)**, da seguinte maneira:

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

> plot(m1, main="m1=lm(y~x)")
```

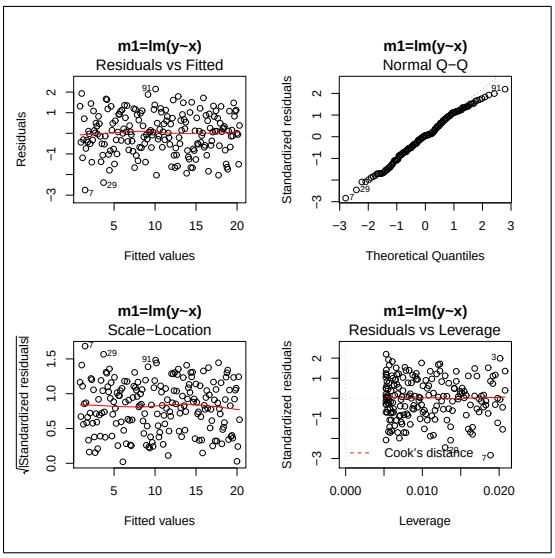


Figura 10. Saídas gráficas do modelo m1

Outro gráfico de interesse é o ajuste do modelo, dado a seguir:

```
> plot(x,y)
> abline(m1$coef, col=8)   #$ Ajusta a reta de regressão
```

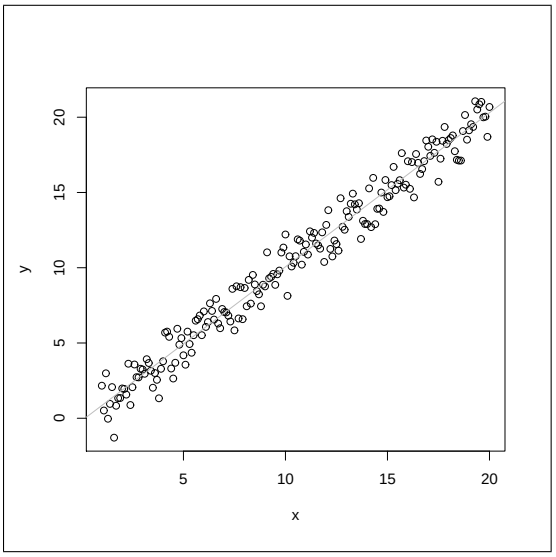


Figura 11. Diagrama de dispersão, exemplo 1

9. O comando confint
   Determina como padrão o intervalos de 0.95 de confiança para os verdadeiros valores dos parâmetros do modelo de regressão.

```
> confint(m1)
                 2.5 %    97.5 %
(Intercept) -0.4757647 0.1260291
x            0.9983005 1.0490437
```

   Para alterar o intervalo padrão, exemplo para 0.90 de confiança, acrescentamos o nível na função:

```
> confint(m1, level=0.90)
                   5 %      95 %
(Intercept) -0.4270074 0.0772718
x            1.0024117 1.0449325
```

10. O comando vcov
    Determina a matriz de variâncias e covariâncias das estimativas dos parâmetros do modelo de regressão:

```
> vcov(m1)
             (Intercept)             x
(Intercept)  0.023268029 -0.0017370374
x           -0.001737037  0.0001654321
```

11. Comandos adicionais
    A soma de quadrados dos resíduos, a log-verosimilhança do modelo assumindo normalidade dos resíduos e os critérios de informação AIC e BIC (estes critérios são discutidos no capítulo de séries temporais) pode ser calculada como:

```
> deviance(m1)
[1] 181.5468

> logLik(m1) #função log-verosimilhança
log Lik. -266.1697 (df=3)

> AIC(m1)
[1] 538.3394

> BIC(m1)
[1] 548.0962
```

Comandos adicionais com graus de liberdade, ajuste de y, resíduos, e coeficientes estimados, podem também se calculados com os comandos:

```
m1$df;m1$fitted.values;m1$residuals;m1$coeff
```

13 Exemplo 2

No exemplo apresentado na introdução referente ao consumo de água de uma residência durante o período de 12/05 a 04/11. Com as seguintes variáveis: valor em reais, consumo em m^3, dias de consumo, e construindo o imóvel (1: sim, 2: não). Ver dados na integra no apêndice.

```
ca=read.table("consumoagua2.txt",header=T)
ca
```

```
   dado valor consumo diasconsumo Construindo
1     1 14.76      11          30           0
2     2 13.47      10          31           0
.........................................
64   64 34.39      14          31           1
65   65 22.21      11          32           0
```

13.1 O modelo

Para esclarecimento usaremos os nomes das variáveis:

$$valorpago = \beta_1 + \beta_2 consumo + \epsilon \tag{3.38}$$

```
> attach(ca)
> md=lm(valor~consumo)
> summary(md)

Call:
lm(formula = valor ~ consumo)

Residuals:
    Min      1Q  Median      3Q     Max
-5.9517 -2.6347 -0.4547  1.6074  9.8732

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -12.365      1.704  -7.258 7.09e-10 ***
consumo        3.007      0.130  23.132  < 2e-16 ***

Residual standard error: 3.721 on 63 degrees of freedom
Multiple R-squared: 0.8947,     Adjusted R-squared: 0.893
F-statistic: 535.1 on 1 and 63 DF,  p-value: < 2.2e-16
```

13.2 Resumo dos resultados

Para fortalecer a interpretação do exemplo motivador dado na introdução, usaremos a saída do comando **summary()**. Inicialmente é apresentado o modelo de regressão proposto, posteriormente as estatísticas dos resíduos como valor mínimo, valor máximo, e os valores do primeiro segundo e terceiro quartil. Na parte dos coeficientes identificamos as estatísticas das estimativas dos parâmetros do modelo. O valor do intercepto $\hat{\beta}_1$ que corta o eixo vertical do valor de consumo é $-12,37$, com erro padrão de 1.704, o quociente de estes dois valores é $t_0 = -7,258$, num teste de hipótese bicaudal, se comparado com a estatística $t = t_{64}(0.025) = -1.99773$, temos que o valor observado $t_0 < t$, por tanto o parâmetro do intercepto é significativo (não há evidencias para aceitar a hipóteses nula H0, de que o intercepto é zero), isto pode ser confirmado por outra estatística chamada de $p-valor = 7.09e-10$, o qual é menor que $\alpha/2 = 0.025$, concluindo que a localização do $p - valor$ está fora da área de aceitação da hipótese nula H0.
O valor de variação pelo consumo ou variação do valor pago $\hat{\beta}_2$ é 3.007, indicando que a uma unidade de consumo em m^3, á um acréscimo de 3 reais em média no valor pago. O erro padrão de esta estimativa é 0.13 centavos de real, o quociente é $t_1 = 23.132$, num teste de hipótese bicaudal, $t = t_{64}(0.975) = 1.99773$, (valor observado maior que o tabular), por tanto o parâmetro de variação pelo consumo é significativo, isto pode ser confirmado por outra estatística chamada de $p - valor = 2e - 16$ o qual é menor que $\alpha/2 = 0.025$, o qual está fora da área de aceitação da hipótese nula H0.

14 linearidade nos parâmetros

14.1 Aceitação dos β_i

Região de aceitação de H0, num teste bicaudal.

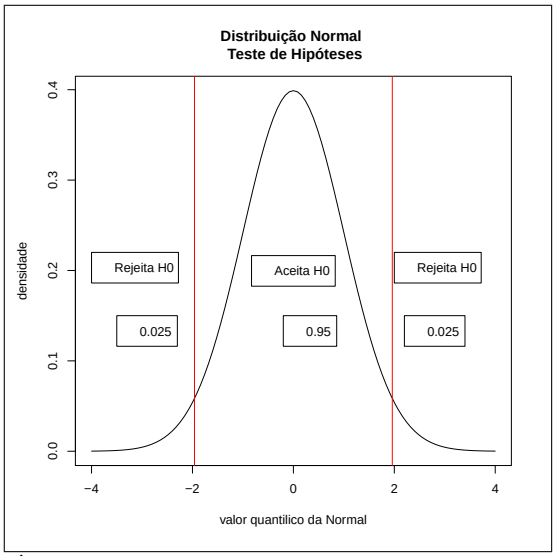


Figura 12. Áreas de Aceitação e rejeição na distribuição Normal.

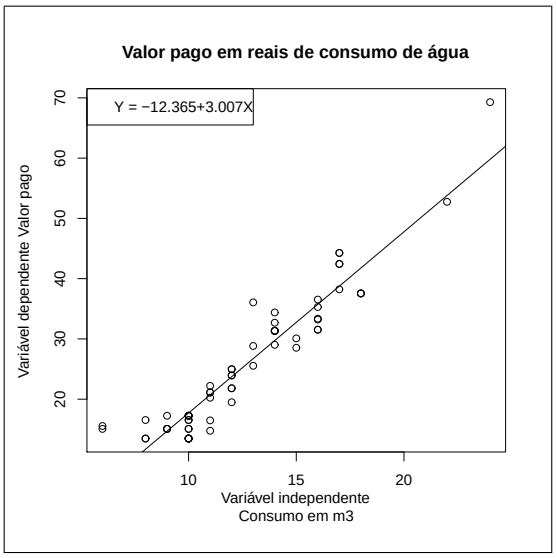


Figura 13. Valor pago em reais e reta de regressão do consumo

Acima observamos o gráfico de dispersão entre consumo e valor de consumo de água.

```
> plot(consumo,valor, main="Valor pago em reais de consumo de água",
ylab="Varíavel dependente Valor pago", xlab="Variável independente
```

```
Consumo em m3")

> legend("topleft", "Y = -12.365+3.007X")
> lines(consumo,fitted(md),lty=3)
```

14.2 ANOVA

Na parte da análise de variância é apresentado a resposta do consumo e dos resíduos, os graus de liberdade correspondentes são 1 e 63, a soma de quadrados do consumo é 7408.8, a soma de quadrados dos resíduos é 872.3, os quadrados médios correspondentes são calculados dividindo-os pelos graus de liberdade individuais, sendo o quadrado médio do consumo igual a 7408.8, é o quadrado médio dos resíduos, conhecido comumente com $\hat{\sigma}^2 = 13.8$. O valor da distribuição F, corresponde ao quociente dos quadrados médios do consumo sobre os quadrados médios dos resíduos, sendo o seu valor 537.07.

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

Response: valor
          Df Sum Sq Mean Sq F value    Pr(>F)
consumo    1 7408.8  7408.8  535.07 < 2.2e-16 ***
Residuals 63  872.3    13.8
```

14.3 Hipóteses do modelo

A tabela do ANOVA, é construída formulando os passos de um teste de hipótese:

1. H0: O modelo não é adequado (Conjuntamente os $\beta_i = 0$)
   Ha: Caso contrario (Conjuntamente os $\beta_i \neq 0$)
   Pela definição da nossa hipótese alternativa(Ha), nosso teste é bicaudal.

2. O nível de significância $\alpha = 0.05$

3. A estatística de prova F, é o quociente de duas qui-quadrado, no numerador o quadrado médio do consumo com 1 gl, e o denominador o quadrado médio dos resíduos ($\hat{\sigma}^2$) com 63 gl. Denominamos de valor calculado $F_0 = 535.07$.

4. O valor tabular corresponde a uma $F_{(1,63)}(0.975)$, no R é calculada por:

```
> qf(0.975,1,63)
[1] 5.272703
```

5. A regra de decisão é se $F_0 > F_{(1,63)}(0.975)$, não há evidencias para aceitar H0. Nosso exemplo $535.07 > 5.2727$, por tanto, concluímos em favor da hipótese alternativa, que conjuntamente os $\beta_i \neq 0$, ou o modelo é adequado.

14.4 As previsões

Para realizar ajustes de previsão, imaginemos que estamos interessados em saber quanto será o valor pago para um consumo de 20 m^3 e 25 m^3 podemos, realizar o seguinte comando:

```
> predict(md,newdata=data.frame(consumo=c(20,25)),
interval="confidence")

       fit      lwr      upr
1 47.77441 45.64589 49.90294
2 62.80927 59.46250 66.15605
```

Podemos interpretar que para um consumo de $20m^3$ de água, teremos em média um valor pago de 47.77 reais, com um intervalo de confiança que pode variar entre 45.65 e 49.90 reais no valor pago.
Para um consumo de $25m^3$ de água, teremos em média um valor pago de 62.81 reais, com um intervalo de confiança que pode variar entre 59.46 e 66.16 reais no valor pago.

14.5 Os resíduos

Podemos identificar que os resíduos do modelo proposto seguem uma distribuição normal ao longo do tempo, usamos o teste de normalidade de Jarque Bera. Atenção o teste com ks.test do pacote *stats*, é usado para normalidade de dados de coorte transversal.

```
>library(tseries)

> resíduos=residuals(md)

> jarque.bera.test(resíduos)

        Jarque Bera Test

data:  resíduos
X-squared = 9.5102, df = 2, p-value = 0.008608
```

Gráficamente a normalidade dos resíduos é obtido no R por:

```
> z=seq(-4,4,length=1000)

> respmodd=md$res/sd(md$res)

> plot(z, dnorm(z), main="resíduos padronizados e distribuição
teórica ", xlab="z", ylab="densidade", ylim=c(-0.1,0.5))

> lines(density(respmodd), col=8)

> legend(-3.8,0.38,legend=c("N(0,1)", "resíduos"),
col=c(1,8),lty=1:1)
```

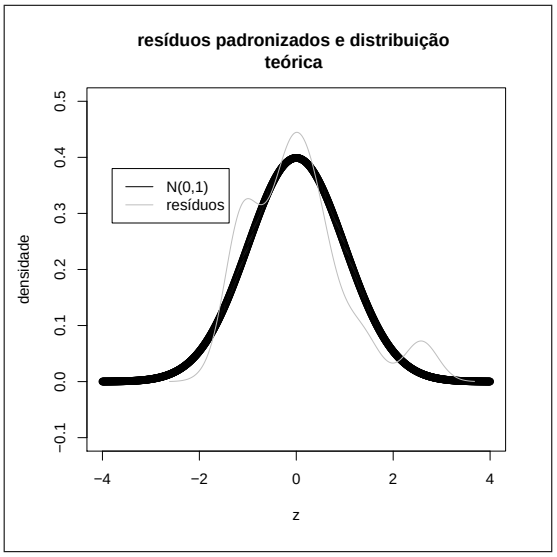


Figura 14. Distribuição empírica dos resíduos do modelo.

Pelo teste de normalidade de Jarque Bera, podemos afirmar que não há evidencias para aceitar a normalidade dos dados. Comandos gráficos adicionais podem ser obtidos com as seguintes funções:

```
qqnorm(residuals(md),plot.it=T)

qqline(residuals(md))
```

14.6 Objetos da função lm

Nosso modelo(md) tem componentes adicionais:

```
> names(md)

 [1] "coefficients"  "residuals"     "effects"       "rank"
 [5] "fitted.values" "assign"        "qr"            "df.residual"
 [9] "xlevels"       "call"          "terms"         "model"
```

A forma de ser chamados é :

```
> md$call
lm(formula = valor ~ consumo)

> md$rank
 [1] 2
```

15 O modelo quadrático

Em situações onde os dados não seguem uma tendência linear, uma proposta em favor do modelo quadrático, ou de potência, é adequada, no exemplo 2 do consumo de água, uma inspeção visual (ver figura 13) permite considerar um modelo quadrático, a fim de melhorar o ajuste da série de consumo de água.

15.1 Modelo matemático

$$y = \hat{\beta}_1 + \hat{\beta}_2 x + \hat{\beta}_3 x^2 + \epsilon \tag{3.39}$$

Para os dados do consumo de água:

$$valorpago = \hat{\beta}_1 + \hat{\beta}_2 consumo + \hat{\beta}_3 consumo^2 + \epsilon \tag{3.40}$$

15.2 Implementação no R

```
> m2=lm(valor~consumo+I(consumo^2))

> summary(m2)

Call:
lm(formula = valor ~ consumo + I(consumo^2))

Residuals:
    Min      1Q  Median      3Q     Max
-5.4546 -2.7179 -0.4616  2.1583 10.4806

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.89565    4.53139   0.639 0.525164
```

```
consumo       0.63696    0.67124   0.949 0.346337
I(consumo^2)  0.08523    0.02375   3.588 0.000658 ***

Residual standard error: 3.413 on 62 degrees of freedom
Multiple R-squared: 0.9128,     Adjusted R-squared:  0.91
F-statistic: 324.4 on 2 and 62 DF,  p-value: < 2.2e-16
```

Um modelo sem intercepto é proposto a seguir:

```
> m3=lm(valor~consumo+I(consumo^2)-1)

> summary(m3)

Call:
lm(formula = valor ~ consumo + I(consumo^2) - 1)

Residuals:
    Min      1Q  Median      3Q     Max
-5.4715 -2.6125 -0.2055  2.0475 10.3046

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
consumo      1.058471   0.123749   8.553 3.86e-12 ***
I(consumo^2) 0.070978   0.008153   8.705 2.10e-12 ***

Residual standard error: 3.397 on 63 degrees of freedom
Multiple R-squared: 0.9857,     Adjusted R-squared: 0.9852
F-statistic:  2168 on 2 and 63 DF,  p-value: < 2.2e-16
```

16 Modelos de potência

Os modelos de regressão potencial tem uma extensão natural a partir do modelo quadrático, asi podemos construir o modelo cúbico com a seguinte sentencia:

```
> m4=lm(valor~consumo+I(consumo^2)+I(consumo^3))
```

Outros modelos de potencia que o R permite construir é usando a função poly().

```
> m5=lm(valor~poly(consumo,3))
```

Para um modelo de potência k, basta fazer:

```
> mk=lm(valor~poly(consumo,k))
```

Nota: m4 e m5 não apresentam as mesmas estimativas.

17 Comparando modelos

Quando realizamos uma análise de regressão, estamos interessados se alguma variável independente não considerada no modelo é relevante para um modelo ser adequado, para comparar dois modelos já analisados usamos a função **anova()**, da seguinte forma

```
> anova(md,m2)

Analysis of Variance Table

Model 1: valor ~ consumo
Model 2: valor ~ consumo + I(consumo^2)
  Res.Df    RSS Df Sum of Sq      F   Pr(>F)
1     63 872.32
2     62 722.34  1    149.98 12.873 0.000658 ***
```

Podemos concluir que a inclusão do termo quadrático como regressor é significativo ($p - valor < 0,05$), e por tanto o modelo de regressão quadrático tem melhor performance de ajuste se comparado como o modelo de regressão linear simples. Para visualizar esta comparação podemos usar os seguintes comandos:

```
> plot(consumo,valor,axes=F)

> lines(consumo,fitted(md))

> lines(consumo,fitted(m2),col=2)
```

18 Derivação dos EMQO

A ideia intuitiva que temos para determinar os estimadores dos parâmetros do modelo de regressão linear simples é encontrar uma função que minimize os erros, ou que o ajuste do modelo seja o mais próximo possível da variável dependente Y, isto pode ser proposto da seguinte maneira:

$$\epsilon = Y - \hat{\beta}_1 - \hat{\beta}_2 X = Y - \hat{Y} \tag{3.41}$$

Como a esperança dos resíduos é zero: $E(\epsilon) = 0$, por hipóteses, não teria sentido minimizar uma função linear dos resíduos, consideraremos por tanto minimizar a soma dos quadrados dos resíduos, da seguinte forma:

$$Z = \sum_{i=1}^{n} \hat{\epsilon}^2 = \sum_{i=1}^{n} (Y - \hat{\beta}_1 - \hat{\beta}_2 X)^2 \tag{3.42}$$

Para determinar as estimativas dos parâmetros derivamos parcialmente a equação acima em termos de $\hat{\beta}_1$ e $\hat{\beta}_2$, da seguinte maneira:

$$\frac{\partial Z}{\partial \hat{\beta}_1} = 2 \sum (Y - \hat{\beta}_1 - \hat{\beta}_2 X)(-1) = 0 \tag{3.43}$$

$$\sum Y = \sum \hat{\beta}_1 + \sum \hat{\beta}_2 X \tag{3.44}$$

$$\sum Y = n\hat{\beta}_1 + \hat{\beta}_2 \sum X \tag{3.45}$$

$$\frac{\partial Z}{\partial \hat{\beta}_2} = 2 \sum (Y - \hat{\beta}_1 + \hat{\beta}_2 X)(-X) = 0 \tag{3.46}$$

$$\sum XY = \sum \hat{\beta}_1 X + \sum \hat{\beta}_2 X^2 \tag{3.47}$$

$$\sum XY = \hat{\beta}_1 \sum X + \hat{\beta}_2 \sum X^2 \tag{3.48}$$

Para um sistema de duas equações e duas incógnitas, podemos determinar as estimativas dos parâmetros da seguinte forma:

$$\hat{\beta}_2 = \frac{\sum (XY) - (\sum Y * \sum X)/n}{\sum X^2 - (\sum X^2)/n} = \frac{\sum (Y - \bar{Y})(X - \bar{X})}{\sum (X - \bar{X})^2} \tag{3.49}$$

$$\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2 \bar{X} = \frac{\sum Y - \hat{\beta}_2 \sum X}{n} \tag{3.50}$$

19 Regressão Multipla

Uma extensão natural do modelo de regressão linear simples é o modelo de regressão múltipla:

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + ... + \beta_p x_p + \epsilon \tag{1}$$

e na forma matricial escrita como:

$$y = X\beta + \epsilon, \tag{3.51}$$

onde:

- y é um vetor $n \times 1$ de observações da variável dependente;
- X é uma matriz fixa contendo observações sobre as variáveis explicativas, de dimensão $n \times p$ (sendo $p < n$);
- β é um vetor com $p \times 1$ parâmetros desconhecidos;
- ϵ é um vetor $n \times 1$ de resíduos aleatórios com média 0.

Utilizando o método de mínimos quadrados, obtemos o sistema de equações normais, dado na forma matricial por

$$X'X\widehat{\beta} = X'y. \tag{3.52}$$

onde a matrix de dados:

$$X' = \begin{bmatrix} 1 & 1 & 1 & \dots & 1 \\ x_{21} & x_{22} & x_{23} & \dots & x_{2n} \\ x_{31} & x_{32} & x_{33} & \dots & x_{3n} \\ \dots & \dots & \dots & \dots & \dots \\ x_{p1} & x_{p2} & x_{p3} & \dots & x_{pn} \end{bmatrix}$$

a variável dependente é:

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{bmatrix}$$

O vetor de parâmetros é $\beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \dots \\ \beta_p \end{bmatrix}$

$$X'X = \begin{bmatrix} n & \sum x_2 & \sum x_3 & \dots & \sum x_p \\ \sum x_2 & \sum x_2^2 & \sum x_2x_3 & \dots & \sum x_2x_p \\ \sum x_3 & \sum x_2x_3 & \sum x_3^2 & \dots & \sum x_3x_p \\ \dots & \dots & \dots & \dots & \dots \\ \sum x_p & \sum x_px_2 & \sum x_px_3 & \dots & \sum x_p^2 \end{bmatrix}$$

Se $X'X$ for uma matriz não singular, a solução do sistema será

$$\widehat{\beta} = (X'X)^{-1}X'y. \tag{3.53}$$

onde:

$$X'y = \begin{bmatrix} \sum y \\ \sum x_2y \\ \dots \\ \sum x_py \end{bmatrix}$$

Os elementos do vetor $\widehat{\beta}$ são as estimativas de mínimos quadrados dos parâmetros associados às variáveis explicativas. Tais estimativas medem o efeito linear de cada variável de X sobre y, após terem sido descontadas de ambas as influências lineares de todas as outras variáveis explicativas incluídas no modelo.

19.1 Matriz de dispersão

A matriz de dispersão D, é dada pela seguinte equação:

$$D = [\hat{\beta} - E(\hat{\beta})][\hat{\beta} - E(\hat{\beta})]' = (X'X)^{-1}\sigma^2 \tag{3.54}$$

19.2 ANOVA

Fontes de variação	Gl	SQ	QM	F	p-valor
Da Regressão	$p-1$	$\hat{\beta}'X'y - C$	$\frac{\hat{\beta}'X'y-C}{p-1}$	$\frac{QMReg}{QMRes}$	
Do Resíduo	$n-p$	$y'y - \hat{\beta}'X'y$	$\frac{y'y-\hat{\beta}'X'y}{n-p}$		
Total	$n-1$	$y'y - C$			

$QMReg$: significa quadrado médio da regressão.
$QMRes$: significa quadrado médio dos resíduos.
$C = n\bar{y}^2$.

20 Exemplo 3

Considere consumo de água de uma residência (extensão do exemplo 2), com as seguintes variáveis: valor pago em reais (valor), consumo de água em m^3 (consumo), dias de consumo (diasconsumo), e construção da residência (Construindo), onde o valor 1 é construindo e o valor 0, não construindo. Um modelo completo é descrito a seguir:

20.1 O modelo

Considere o modelo de regressão na forma matricial:

$$y = X\beta + \epsilon, \tag{3.57}$$

O qual pode ser escrito na forma de modelo de regressão múltipla da seguinte forma:

$$valorpago = \beta_1 + \beta_2 consumo + \beta_3 diasconsumo + \beta_4 Construindo \tag{3.58}$$

20.2 As hipóteses

Consideremos a hipótese nula $H0$, que o modelo não é adequado ($X\beta = 0$), contra a hipótese alternativa Ha, que o modelo é adequado:

1. $H0 : X\beta = 0$
   $Ha : X\beta \neq 0$

2. O nível de significância considerado $\alpha = 0.05$

3. A estatística de prova:

$$F = \frac{QMReg}{QMres} = \frac{(\hat{\beta}' X' y - C)/p - 1}{(y'y - \hat{\beta}' X' y)/(n - p)} \tag{3.59}$$

4. Regra de decisão: Se o $p - valor > 0.05$, há evidencias em favor de $H0$, caso contrario em favor de Ha, (ver ANOVA)

20.3 Estimação dos parâmetros

para estimar os 4 parâmetros do modelo proposto $(\beta_1, \beta_2, \beta_3, \beta_4)$, usamos os seguintes comandos no R:

```
> mc=lm(valor~consumo+diasconsumo+Construindo)

> summary(mc)

Call:

lm(formula = valor ~ consumo + diasconsumo + Construindo)
Residuals:
    Min      1Q  Median      3Q     Max
-7.2955 -1.9715 -0.3418  1.8726  9.7028

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.7294    12.2383  -0.141  0.88809
consumo       2.7564     0.1571  17.547  < 2e-16 ***
diasconsumo  -0.2747     0.4080  -0.673  0.50322
Construindo   3.6809     1.2387   2.972  0.00424 **

Residual standard error: 3.52 on 61 degrees of freedom
Multiple R-squared: 0.9087,     Adjusted R-squared: 0.9043
F-statistic: 202.5 on 3 and 61 DF,  p-value: < 2.2e-16
```

Interpretação: a partir das estimativas dos parâmetros do modelo, podemos afirmar, que um proposta em favor de um modelo sem intercepto (valor do $p - valor = 0.888$) é adequado (pois $p - valor > 0.05$), o qual é dado a seguir:

```
> m1=lm(valor~consumo+diasconsumo+Construindo-1)

> summary(m1)

Call:
lm(formula=valor ~ consumo + diasconsumo + Construindo - 1)
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-7.3419 -1.9108 -0.3446  1.8175  9.6925

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
consumo      2.75811    0.15536  17.753  < 2e-16 ***
diasconsumo -0.33177    0.05995  -5.534 6.72e-07 ***
Construindo  3.68759    1.22800   3.003  0.00385 **

Residual standard error: 3.492 on 62 degrees of freedom
Multiple R-squared: 0.9851,     Adjusted R-squared: 0.9844
F-statistic:  1367 on 3 and 62 DF,  p-value: < 2.2e-16
```

Interpretação: podemos observar que todas as variáveis independentes que compõem o modelo são significativas individualmente($p < 0.05$), por tanto estatisticamente estas variáveis fazem parte do modelo. O coeficiente de determinação é $R^2_{mult} = 0.985$. Para esclarecer se o modelo é adequado ($X\beta = 0$), realizamos o análise de variância ANOVA.

20.3.1 O ANOVA

```
> anova(m1)

Analysis of Variance Table

Response: valor
            Df Sum Sq Mean Sq   F value    Pr(>F)
consumo      1  49175   49175 4033.0722 < 2.2e-16 ***
diasconsumo  1    736     736   60.3480 9.986e-11 ***
Construindo  1    110     110    9.0175  0.003852 **
Residuals   62    756      12
```

21 Modelo linear parcial

Em R podemos adicionar termos de potência nas variáveis regressoras para construir modelos de regressão múltipla, porem a escolha pode tornar-se lenta, por tanto uma extensão semi-paramétrica pode contornar o desempenho na escolha da potencia, no modelo parcialmente linear em: $y = \beta_1 + g(X_1) + \beta_2 X_2 + \cdots + \epsilon$, onde g é uma função desconhecida que pode ser utilizada diretamente na função lm (), e assim ajustar o modelo desejado. Considere o modelo parcial:

$$y = \beta_1 + g(X_1) + \beta_2 X_2 + \cdots + \epsilon \tag{3.60}$$

Onde g é uma função desconhecida que será estimada a partir dos dados, usamos a função **bs()** do pacote **splines**, da seguinte forma:

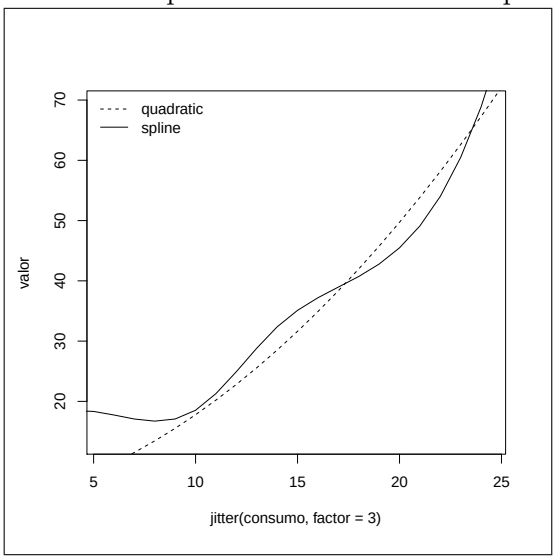
```
> library(splines)

> pm1=lm(valor~bs(consumo, df=5)+diasconsumo+Construindo)

> summary(pm1)

Call:
lm(formula=valor ~ bs(consumo, df=5) + diasconsumo + Construindo)

Residuals:
   Min     1Q Median     3Q    Max
-5.663 -1.436 -0.369  1.668  7.242

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)           18.4351     9.6864   1.903   0.0621 .
bs(consumo, df = 5)1  -0.9266     3.4455  -0.269   0.7890
bs(consumo, df = 5)2  -3.1656     2.5040  -1.264   0.2113
bs(consumo, df = 5)3  27.3786     3.5786   7.651 2.62e-10 ***
bs(consumo, df = 5)4  19.6916     4.5197   4.357 5.57e-05 ***
bs(consumo, df = 5)5  51.1325     3.4760  14.710  < 2e-16 ***
diasconsumo           -0.1033     0.3217  -0.321   0.7493
Construindo            2.5003     0.9987   2.503   0.0152 *

Residual standard error: 2.738 on 57 degrees of freedom
```

```
Multiple R-squared: 0.9484,     Adjusted R-squared: 0.942
F-statistic: 149.6 on 7 and 57 DF,  p-value: < 2.2e-16
```

A sugestão de usar graus de liberdade ($df = 5$), e dada no critério de seleção de modelo (AIC), o padrão é $df = 3$, para confirmar usamos a seguinte função:

```
ps_m1 <- lapply(3:9, function(i)
lm(valor ~ bs(consumo, df = i) + diasconsumo + Construindo))
structure(sapply(ps_m1, AIC, k = log(nrow(ca))),.Names = 3:9)
      3       4       5       6       7       8       9
356.0836 344.9000 344.4504 346.0979 350.3472 352.5446 353.3633
```

Figura 15. desempenho de dois modelos de potência



Houve uma mudança nas estimativas nas variáveis dias de consumo e construção. Para realizar o gráfico acima, usamos os seguintes comandos:

```
m2=lm(formula = valor ~ consumo + I(consumo^2))

pm <- data.frame(consumo= 0:40, diasconsumo=with(ca,
mean(diasconsumo[Construindo == 1])),Construindo = 1)
```

```
pm$yhat1 <- predict(m2, newdata = pm)
pm$yhat2 <- predict(pm1, newdata = pm)

plot(valor ~ jitter(consumo, factor = 3), pch = 30,
col = rgb(0.5, 0.5, 0.5, alpha = 0.02), data = ca)

lines(yhat1 ~ consumo, data = pm, lty = 2)
lines(yhat2 ~ consumo, data = pm)
legend("topleft", c("quadratic", "spline"), lty =
c(2,1),bty = "n")
```

22 Seleção de regressores

Em alguns casos é necessário se decidir quais variáveis incluir no modelo de regressão. Há varias formas de contornar esta situação:

1. Use o teste F, começando com modelos com vários regressores e teste exclusões de variáveis ou grupo de variáveis, pode usar a regressão **stepwise** que como padrão utiliza a técnica `backward`, da seguinte forma:

```
> mc=lm(valor~consumo+diasconsumo+Construindo)
> step(mc)
Start:  AIC=167.46
valor ~ consumo + diasconsumo + Construindo

              Df Sum of Sq    RSS    AIC
- diasconsumo  1       5.6  761.3 165.94
<none>                      755.7 167.46
- Construindo  1     109.4  865.1 174.25
- consumo      1    3814.5 4570.2 282.44

Step:  AIC=165.94
valor ~ consumo + Construindo

              Df Sum of Sq    RSS    AIC
<none>                      761.3 165.94
- Construindo  1     111.0  872.3 172.79
- consumo      1    3940.2 4701.5 282.28
```

```
Call:
lm(formula = valor ~ consumo + Construindo)

Coefficients:
(Intercept)      consumo  Construindo
     -9.880        2.733        3.706
```

Podemos observar que no modelo completo o AIC inicial é 167.46, e quando excluímos a variável `diasconsumo` o AIC inicial cai para 165.48. Por tanto por este critério a escolha do melhor modelo é: valor depende do consumo e da reforma da casa. Para usar a técnica `forward`, use o comando `>step(mc,direction = "forward")`.

2. Uso do coeficiente de determinação, R^2, este critério tem suas limitações, pois é uma função não decrescente, isto significa que a inclusão de uma variável não importante, aumenta o valor do R^2, da mesma forma que se a inclusão de uma variável regressora importante.

3. Uso do R^2 ajustado, pois pode aumentar ou diminuir quando incluímos novas variáveis, este valor é calculado da seguinte forma:

$$\bar{R}^2 = 1 - \frac{SQRes/n - p}{SQT/n - 1}, \tag{3.61}$$

Tanto o R^2 como $\bar{R}^2$, são disponibilizados pela função **summary()**. Podemos afirmar que os modelos $m1$ e $m3$, apresentam melhor performance por este critério.

4. Podemos usar o critério de seleção de modelos AIC, os quais surgem para penalizar a inclusão de novas variáveis. Assim cada variável adicionada ao modelo, apresenta 1 grau de liberdade a menos. Dentre estes critérios temos: Akaike, o qual usa a seguinte função de penalização

$$AIC = \log(\frac{SQRes}{n}) + \frac{2p}{n} \tag{3.55}$$

onde p, é o número de regressores do modelo, n, é o total de dados, e SQR, é a soma de quadrados do resíduo.

5. Um outro critério proposto por Scharz, é de informação Bayesiana, definido como

$$BIC = \log(\frac{SQRes}{n}) + \frac{p}{n}\log(n) \tag{3.56}$$

O BIC penaliza mais fortemente que o AIC, quando $n > 8$. Uma alteração do BIC, é o BICc, que é fornecida em alguns funções do projeto R. Vejamos os valores destes critérios nos modelos construídos:

```
> AIC(md,m1,m2,m3,m4)

   df      AIC
md  3 359.2519
m1  4 351.9462
m2  4 348.9891
m3  3 347.4158
m4  6 329.2190

> BIC(md,m1,m2,m3,m4)

   df      BIC
md  3 365.7751
m1  4 360.6437
m2  4 357.6867
m3  3 353.9390
m4  6 342.2653
```

Por este critério podemos observar que o modelo polinomial de ordem 4, apresenta melhor performance para o melhor ajuste do adequado.

6. Um outro critério de seleção de regressores é aplicar o teste que compara modelos não encaixados, por exemplo o modelo md esta encaixado em $m1, m2, m3$, e $m4$, porem $m1$ não é encaixado com $m3$ e $m4$, no exemplo temos:

```
> encomptest(m1,m3)
Encompassing test

Model 1: valor ~ consumo + diasconsumo + Construindo - 1
```

```
Model 2: valor ~ consumo + I(consumo^2) - 1
Model E: valor ~ consumo + diasconsumo + Construindo +
I(consumo^2) - 1
          Res.Df Df       F    Pr(>F)
M1 vs. ME     61 -1 13.6058 0.0004822 ***
M2 vs. ME     61 -2  5.3787 0.0070574 **
```

Um outro teste para modelos não encaixados é dado por Cox.

```
> coxtest(m1,m3)
Cox test

Model 1: valor ~ consumo + diasconsumo + Construindo - 1
Model 2: valor ~ consumo + I(consumo^2) - 1
                Estimate Std. Error z value  Pr(>|z|)
fitted(M1) ~ M2  -9.1559     2.1352 -4.2881 1.802e-05 ***
fitted(M2) ~ M1  -3.6272     2.5998 -1.3952     0.163
```

7. Wald propôs um teste que compara modelos encaixados (ou aninhados), por exemplo o modelo md esta encaixado em $m1, m2, m3$, e $m4$, no exemplo usaremos a função `waldtest()` da biblioteca `lmtest`.

```
> waldtest(md,m1)
Wald test

Model 1: valor ~ consumo
Model 2: valor ~ consumo + diasconsumo + Construindo - 1
  Res.Df Df      F    Pr(>F)
1     63
2     62  1 69.366 1.067e-11 ***
```

23 Especificação do modelo

Para superar o pressuposto de que o modelo esta bem especificado, como linearidade do modelo, inclusão de regressores importantes, e outras situações, usamos o teste de Ramsey.

23.1 Teste de Ramsey

As hipóteses do modelo são:
$H0$: O modelo esta corretamente especificado, $y = X\beta$
Ha: O modelo não esta corretamente especificado. $y = X\beta + Z\gamma$.
Onde Z, é uma matriz de regressores transformados de X que são fundamentais no modelo, e γ os parâmetros correspondentes de Z.

```
> m1=lm(valor~consumo+diasconsumo+Construindo-1)
> resettest(m1)

        RESET test

data:  m1
RESET = 7.923, df1 = 2, df2 = 60, p-value = 0.0008842
```

Não há evidencias para aceitar que o modelo completo sem intercepto (m1) esteja bem especificado. Um modelo do tipo quadrático é proposto:

```
> m3=lm(valor~consumo+I(consumo^2)-1)

> resettest(m3)

        RESET test
data:  m3
RESET = 0.0436, df1 = 2, df2 = 61, p-value = 0.9574
```

Este modelo quadrático (m3), esta bem especificado.

24 Normalidade dos resíduos

Para verificar a hipótese $H0$, de normalidade dos resíduos (nem sempre necessário), usamos o teste de Jarque Bera, da biblioteca `tseries`.

```
> library(tseries)
> jarque.bera.test(residuals(m1))

        Jarque Bera Test

data:  residuals(m1)
```

```
X-squared = 4.1774, df = 2, p-value = 0.1238


> jarque.bera.test(residuals(m3))

        Jarque Bera Test

data:  residuals(m3)
X-squared = 2.8264, df = 2, p-value = 0.2434
```

Não há evidencias para rejeitar $H0$, em consequência os resíduos de ambos modelos são normais.

25 Gráficos

Um gráfico do modelo m1 correspondente é:

```
> par(mfrow=c(2,2))
> plot(m1, main="modelo completo")
```

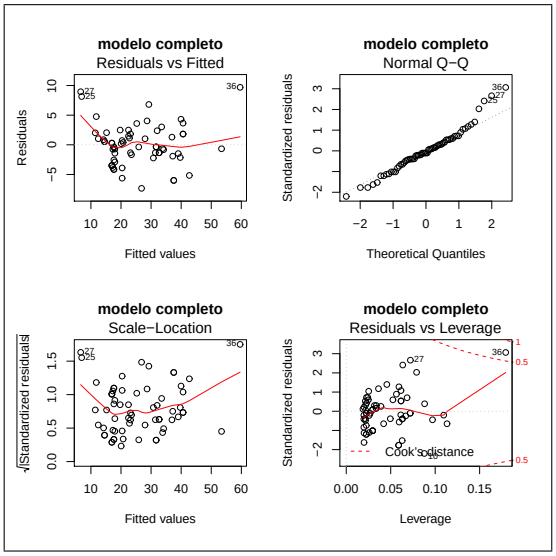


Figura 16. plot padrão do modelo m1

Da figura acima podemos observar o gráfico de dispersão entre os resíduos e os valores estimados, percebendo certa homogeneidade dos resíduos, no

gráfico seguinte observamos a normalidade dos resíduos, com leve afastamento das observações 25,27,36, continuando com o terceiro gráfico observamos que é uma alteração do primeiro gráfico, e que no lugar dos resíduos, é visualizado a raiz dos resíduos com os valores estimados, finalmente no último gráfico apresenta os pontos de alavanca dos resíduos utilizando a distancia de Cook, mostrando que a observação 36 (Novembro de 2008) é um outlier (valor fora do lugar).

26 Teste de mudança estrutural

É possível que alguns parâmetros não sejam constantes ao longo de todas as observações, em dados de séries temporais, separamos as primeiras n_{t1} observações, este é o ponto de separação onde houve mudança estrutural, a ideia é utilizar a soma de quadrados do modelo nestas primeiras observações e compará-las com a soma de quadrados do modelo nas $n - n_{t1}$ observações restantes. Podemos usar a função **sctest()** do pacote **strucchange**. O pacote **gap** disponibiliza a função **chow.test()**, como método alternativo de cálculo. Considere os dados de consumo de água, onde $n_{t1} = 30$.

26.1 O modelo

$valor_t = \beta_t + \beta_{2t} consumo_t + \beta_{3t} diasconsumo_t + \beta_{4t} Construindo_t + \epsilon_t$

Se dividimos nossos dados em dois grupos, então temos:

$valor_{t1} = \beta_{t1} + \beta_{2t1} consumo_{t1} + \beta_{3t1} diasconsumo_{t1} + \beta_{4t1} Construindo_{t1} + \epsilon_{t1}$

e

$valor_{t2} = \beta_{t2} + \beta_{2t2} consumo_{t2} + \beta_{3t2} diasconsumo_{t2} + \beta_{4t2} Construindo_{t2} + \epsilon_{t2}$

Considere a soma de quadrados dos resíduos no modelo completo (SQR), a SQR1, e SQR2, a soma do quadrado dos resíduos do modelo no tempo t1 e t2 respectivamente.

26.2 Hipóteses

A hipóteses nula $H0$: Acima dos primeiros 30 meses de consumo não há mudança estrutural (Ou $\beta_{t1} = \beta_{t2}, \beta_{2t1} = \beta_{2t2}, \beta_{3t1} = \beta_{3t2}, \beta_{4t1} = \beta_{4t2}$)
Ha: Caso contrario.

26.3 Nível de significância

Considere $\alpha = 0.05$

26.4 estatistica de prova

$$Chow = F = \frac{(SQR - (SQR1 + SQR2))/(p)}{(SQR1 + SQR2)/(n_{t1} + n_{t2} - 2p)}. \quad (2)$$

A estatística de teste segue a distribuição F com p e $n_{t1} + n_{t2} - 2p$ graus de liberdade.

```
> sctest(valor~consumo, type="Chow", point=30)

        Chow test

data:  valor ~ consumo
F = 16.995, p-value = 1.36e-06


> sctest(valor~consumo+diasconsumo+Construindo,
type="Chow", point=30)

        Chow test

data:  valor ~ consumo + diasconsumo + Construindo
F = 6.459, p-value = 0.0002335
```

Observamos que não há evidencias para aceitar $H0$, em favor da mudança estrutural, ou em outras palavras há câmbios significativos nas estimativas dos parâmetros.

27 Teste de outliers

O objetivo do teste é determinar se uma dada observação é um outlier (dado que não corresponde a média das observações). As hipóteses são:
$H0$: A n-ésima observação não corresponde a média das observações
Ha: caso contrario.
Use a função **outlierTest()** do pacote **car**, No exemplo precedente temos:

```
> outlierTest(m1)

No Studentized residuals with Bonferonni p < 0.05
Largest |rstudent|:
   rstudent unadjusted p-value Bonferonni p
36 3.299283          0.0016213      0.10538
```

Podemos afirmar que há evidencias para aceitar que o mês 36 não corresponde a média das observações, ou esta fora dos limites de controle da média, por tanto aquele mês é um outlier.

28 Detecção de observações atípicas

O objetivo é detetar observações que apresentem padrão atípico. Observações influentes são aquelas que apresentam uma contribuição relativamente grande para o ajuste, esta influencia pode ser por dois fatores, primeiro por padrão atípico de regressores ou segundo por erros demasiado grandes.

1. Considere alavancagem, as observações cujos regressores apresentam padrão atípico, também chamados de pontos de alavanca ou observações de alta alavancagem. Para medir os graus de alavancagem de diferentes observações a medida padrão é:

$$h_t = X_t(X'X)^{-1}X_t' \tag{3.62}$$

definida como medida de alavancagem da t-ésima observação, e X_t é um vetor de $p \times 1$ seja

$$H = X(X'X)^{-1}X' \tag{3.63}$$

Multiplicando H por y, temos:

$$Hy = X(X'X)^{-1}X'y = X\hat{\beta} = \hat{y} \tag{3.64}$$

onde $h_1, h_2, \cdots, h_t$ são elementos diagonais de H. No caso de uma regressão linear simples, temos que:

$$h_t = \frac{1}{n} + \frac{(x_t - \bar{x})^2}{\sum(x_t - \bar{x})^2} \tag{3.65}$$

Onde $\bar{x} = \sum x_t/n$, h_t aumenta a medida que x_t se afasta de $\bar{x}$. Inicialmente observações com alavancagem próximo de 2 ou 3 vezes maior que a média são pontos de alavanca ou se $h_t > 2p/n$. No R use a função **hatvalues()**, onde $\mathrm{V}(\hat{\epsilon}_i|X) = \sigma^2(1 - h_{ii})$.

```
> hatm1=hatvalues(m1)
> alavancam1=hatm1[hatm1>3*mean(hatvalues(m1))]
> alavancam1
       36
0.1793138

> alavancam1=hatm1[hatm1>2*mean(hatvalues(m1))]
> alavancam1
         4          15          16          36
0.10959219 0.09697215 0.11331092 0.17931385


> alavancam1=hatm1[hatm1>3*2/65]
> alavancam1
         4          15          16          36
0.10959219 0.09697215 0.11331092 0.17931385
```

2. Resíduos padronizados, observações cujos valores da variável dependente são atípicos são chamados de outliers, o problema desta medida é que no caso de erros grandes, os resíduos correspondentes nem sempre são uma boa indicação dos erros verdadeiros. O valor é dado por

$$r_i = \frac{\hat{\epsilon}}{\hat{\sigma}\sqrt{(1-h_{ii})}} \tag{3.66}$$

no R, usamos a função **rstandard()**

```
> respad=rstandard(m1)

> res=respad[respad>qt(0.975,64)]

> res
      13       25       27       36
2.029803 2.409691 2.662142 3.064047
```

3. Distancia de Cook, pode ser calculado a partir dos resíduos padronizados e os valores influentes, da seguinte forma

$$C_t = \frac{r_t^2 h_t}{p(1-h_t)} \tag{3.67}$$

regra de bolso, $C_t > \frac{4}{n-p}$

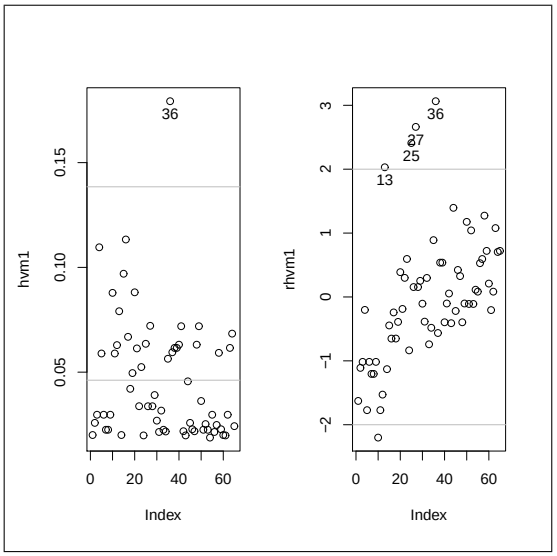


Figura 17. Valores influentes de duas medidas

Considere o modelo m1, a identificação dos valores influentes usando o R:

```
hvm1=hatvalues(m1)
hvm1
        1             2            65
0.01999192    0.02578687   -0.02422222
plot(hvm1); abline(h=4/62, col=2); identify(hvm1)
rhvm1=rstandard(m1)
par(mfrow=c(1,2))
plot(hvm1)
abline(h=c(1,3)*mean(hvm1),col=8)
iden=which(hvm1>3*mean(hvm1))
text(iden,hvm1[iden],rownames(ca)[iden],pos=1,xpd=TRUE)
plot(rhvm1)
abline(h=c(-2,2),col=8)
iden=which(rhvm1>2)
text(iden,rhvm1[iden],rownames(ca)[iden],pos=1,xpd=TRUE)
```

Um agoritmo para identificar gráficamente estes valores influentes é dado por:

```
 vinfluentes=function(modelo, scale=10, col=c(1,2),
 labels=names(rst),...){
 vi=hatvalues(modelo) rst=rstudent(modelo)
 dc=sqrt(as.vector(cookd(modelo)))
 escala=scale/max(dc)
 p=length(coef(modelo)) n=length(rst)
 co=sqrt(4/(n-p))
 plot(vi,rst,xlab="Valores infuentes",
 ylab="Residuos padronizados", type="n",
 ylim=c(-3,4), ...)
 abline(v=c(2,3)*p/n, lty=2)
 abline(h=c(-2,0,2), lty=2)
 for(i in 1:n)
 points(vi[i],rst[i], cex=escala*dc[i],
 col=if(dc[i]>co) col[2] else col[1])
 if(labels[1]!=FALSE) identify(vi,rst,labels)}
> library(car)
> vinfluentes(m1) # identifique com o mause
[1] 10 13 25 27 36
> dc=cooks.distance(m1)  #distancia de cook
```

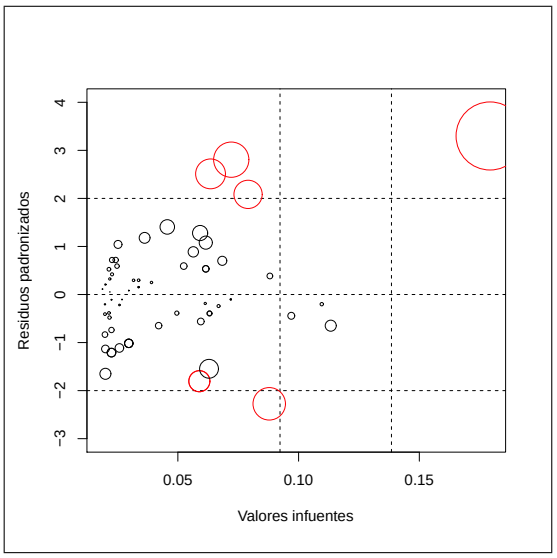


Figura 18. Identificação de valores influentes

29 Consistência

Definição 2 *Convergência em probabilidade. Seja $Z_n, n \in N$ uma sequencia de variáveis aleatórias e seja $c \in \Re$, dizemos que Z_n converge em probabilidade para c, denotado por $Z_n \xrightarrow{p} c$ ou $plim(Z_n) = c, \qquad \forall \epsilon > 0.$*

pode ser escrito como $\lim_{n\to\infty} P(|Z_n - c|] < \epsilon) = 1$

Definição 3 *Consistência. Um estimador $\hat{\theta}$ de um parâmetro $\theta \in \Theta$, sendo Θ um espaço paramétrico, é dito ser consistente se $\hat{\theta} \xrightarrow{p} \theta, \ \forall \theta \in \Theta$*

Para identificar a relação entre consistência e não viés, consideremos alguns estimadores.

1. $\hat{\theta}_1 = \frac{1}{n+1} \sum Z_n$
2. $\hat{\theta}_2 = \frac{0.5}{n} \sum Z_n$
3. $\hat{\theta}_3 = 0.01 Z_1 + \frac{0.99}{n-1} \sum Z_n$

Para o primeiro estimador temos que é viesado:

$$E(\hat{\theta}_1) = E(\frac{1}{n+1} \sum Z_n) = \frac{1}{n+1} \sum E(Z_n) = \frac{1}{n+1} n\theta \neq \theta$$

Para verificar a consistência fazemos:

1. $\lim_{n\to\infty} E(\hat{\theta}_1) = \lim_{n\to\infty} \frac{n}{n+1}\theta = \theta$
2. $\lim_{n\to\infty} V(\hat{\theta}_1) = \lim_{n\to\infty} \frac{1}{(n+1)^2} V(\sum Z_n) = \lim_{n\to\infty} \frac{n}{(n+1)^2}\sigma^2 = 0.$ Por tanto é consistente.

Para o segundo estimador temos que é viesado e inconsistente (prove).
Para o terceiro estimador temos que é não viesado e inconsistente (prove).

30 Iterações

Quando os regressores apresentam dependência existe interação, podemos identificar o grau de interação das variáveis, caso não há interação

dizemos que as variáveis são independentes, em nosso exemplo de consumo de água, é natural que quando estamos construindo há mais consumo de água, ou a mais dias de consumo, mais consumo. Para exemplificar:

```
> m_int=lm(valor~consumo+Construindo*consumo+diasconsumo*consumo)
> summary(m_int)

Call:
lm(formula = valor ~ consumo + Construindo * consumo + diasconsumo
*consumo)

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        -8.84548   43.60920  -0.203   0.8400
consumo             3.51575    3.43121   1.025   0.3097
Construindo        -8.55260    5.40296  -1.583   0.1188
diasconsumo         0.04108    1.42980   0.029   0.9772
consumo:Construindo 0.82826    0.35774   2.315   0.0241 *
consumo:diasconsumo -0.03210    0.11216  -0.286   0.7757

Residual standard error: 3.415 on 59 degrees of freedom
Multiple R-squared: 0.9169,     Adjusted R-squared: 0.9099
F-statistic: 130.3 on 5 and 59 DF,  p-value: < 2.2e-16
```

Podemos interpretar que há interação significativa ($p < 0.05$) entre consumo e construção, já no caso de consumo e dias de consumo não há interação significativa ($p > 0.05$).

31 Regressão separada

Em alguns casos estamos interessados em analisar regressões separadas por vários níveis, no exemplo do consumo de água estamos interessados em determinar qual o desempenho das estimativas quando estamos construindo e quando não estamos construindo, exemplificando

```
> m_cons=lm(valor~Construindo/(consumo+diasconsumo))
> msc=matrix(coef(m_cons), nrow=2)
> m1=lm(valor~consumo+diasconsumo+Construindo-1)
> rownames(msc)=levels(Construindo)
> colnames(msc)=names(coef(m1))[1:2]
> msc
```

```
       consumo diasconsumo
[1,] 21.152708   3.3981281
[2,] -9.665703  -0.9184493
```

Podemos confirmar que o valor de consumo e dias de consumo é diferente para os dois níveis de construção. Para visualizar um gráfico de regressão por separado podemos usar a função

```
> coplot(valor~Construindo|consumo+diasconsumo,panel = panel.smooth)
```

Uma outra alternativa, é separar o conjunto de dados em duas matrizes de dados, a ideia é comparar duas ou mais regressão num mesmo gráfico alem de verificar a variação da tangente de inclinação em cada estrato. Exemplo verificar as regressões de valor em função do consumo, quando estamos ou não construindo.

```
plot(consumo,valor,pch=as.numeric(Construindo))
legend("topleft",legend=c("não construindo","construindo"),pch=0:1)
noconstruindo=ca[Construindo==0,]
construindo=ca[Construindo==1,]
mnc=lm(valor~consumo,data=noconstruindo)
mc=lm(valor~consumo,data=construindo)
summary(mnc)
summary(mnc)
abline(mnc,lty=2)
abline(mc,lty=3)
```

32 Regressão quantílica

Em algumas situações, as estimativas dos parâmetros apresentam mudanças bruscas ou repentinas, a cada quantidade das observações processada. A função **rq()** do pacote **quantreg**, proporciona resultados, para monitorar as mudanças das estimativas a cada percentual de observações de interesse, um exemplo é dado a seguir:

```
> library(quantreg)
> mrq1=rq(m1, tau=seq(0.2,0.8,by=0.15),data=ca)
> summary(mrq1)

Call: rq(formula = m1, tau = seq(0.2, 0.8, by = 0.15), data = ca)
```

```
tau: [1] 0.2
Coefficients:
            coefficients lower bd upper bd
consumo      3.01000      2.45247  3.13831
diasconsumo -0.51969     -0.59977 -0.35503
Construindo  0.00000     -1.13537  4.59762

Call: rq(formula = m1, tau = seq(0.2, 0.8, by = 0.15), data = ca)

tau: [1] 0.35
Coefficients:
            coefficients lower bd upper bd
consumo      2.66710      2.40001  2.92237
diasconsumo -0.32681     -0.52897 -0.22469
Construindo  3.47798     -0.67739  4.78711

Call: rq(formula = m1, tau = seq(0.2, 0.8, by = 0.15), data = ca)

tau: [1] 0.5
Coefficients:
            coefficients lower bd upper bd
consumo      2.61371      2.52489  2.97180
diasconsumo -0.27772     -0.40779 -0.23535
Construindo  3.63515      1.71949  6.55572

Call: rq(formula = m1, tau = seq(0.2, 0.8, by = 0.15), data = ca)

tau: [1] 0.65
Coefficients:
            coefficients lower bd upper bd
consumo      2.72463      2.43338  3.13545
diasconsumo -0.29537     -0.40229 -0.14411
Construindo  4.98220      1.08493  7.49433

Call: rq(formula = m1, tau = seq(0.2, 0.8, by = 0.15), data = ca)

tau: [1] 0.8
Coefficients:
            coefficients lower bd upper bd
consumo      2.65200      2.10188  3.22867
```

```
diasconsumo -0.22847     -0.38592 -0.00817
Construindo  6.04000      2.83924  9.45696
```

Podemos observar que as estimativas dos parâmetros, teve algumas mudanças como é o caso da Construção, a medida que atingimos o total das observações temos um crescimento maior no valor do consumo de água, as outras duas variáveis não teve mudanças repentinas expressivas.

33 Derivação dos EMQO na regressão Múltipla

A ideia é estender as estimativa para $p+1$ parâmetros, isto pode ser proposto da seguinte maneira:

$$\epsilon = Y - \hat{\beta}_1 - \hat{\beta}_2 X_1 - \hat{\beta}_3 X_2 - \cdots - \hat{\beta}_{p+1} X_p = Y - \hat{Y} \tag{3.68}$$

consideraremos por tanto minimizar a soma dos quadrados dos resíduos, da seguinte forma:

$$Z = \sum_{i=1}^{n} \hat{\epsilon}^2 = \sum_{i=1}^{n} (Y - \hat{\beta}_1 - \hat{\beta}_2 X_1 - \hat{\beta}_3 X_2 - \cdots - \hat{\beta}_{p+1} X_p)^2 \tag{3.69}$$

Para determinar as estimativas dos parâmetros da regressão múltipla, temos:

$$\frac{\partial Z}{\partial \hat{\beta}_1} = 2 \sum (Y - \hat{\beta}_1 - \hat{\beta}_2 X_1 - \hat{\beta}_3 X_2 - \cdots - \hat{\beta}_{p+1} X_p)(-1) = 0 \tag{3.70}$$

$$\sum Y = \sum \hat{\beta}_1 + \sum \hat{\beta}_2 X_1 + \sum \hat{\beta}_3 X_2 + \cdots + \sum \hat{\beta}_{p+1} X_p \tag{3.71}$$

$$\sum Y = n\hat{\beta}_1 + \hat{\beta}_2 \sum X_1 + \hat{\beta}_3 \sum X_2 + \cdots + \hat{\beta}_{p+1} \sum X_p \tag{3.72}$$

$$\frac{\partial Z}{\partial \hat{\beta}_2} = 2 \sum (Y - \hat{\beta}_1 - \hat{\beta}_2 X_1 - \hat{\beta}_3 X_2 - \cdots - \hat{\beta}_{p+1} X_p)(-X_1) = 0 \tag{3.73}$$

$$\sum X_1 Y = \sum \hat{\beta}_1 X_1 + \sum \hat{\beta}_2 X_1^2 + \sum \hat{\beta}_3 X_1 X_2 + \cdots + \sum \hat{\beta}_{p+1} X_1 X_p \tag{3.74}$$

$$\sum X_1 Y = \hat{\beta}_1 \sum X_1 + \hat{\beta}_2 \sum X_1^2 + \hat{\beta}_2 \sum X_1 X_2 + \cdots + \sum \hat{\beta}_{p+1} X_p \tag{3.75}$$

$$\frac{\partial Z}{\partial \hat{\beta}_{j+1}} = 2\sum(Y - \hat{\beta}_1 - \hat{\beta}_2 X_1 - \cdots - \hat{\beta}_{j+1} X_j - \cdots)(-X_j) = 0 \tag{3.76}$$

$$\sum X_j Y = \sum \hat{\beta}_1 X_j + \sum \hat{\beta}_2 X_1 X_j + \cdots + \sum \hat{\beta}_{j+1} X_j^2 + \cdots \tag{3.77}$$

$$\sum X_j Y = \hat{\beta}_1 \sum X_j + \hat{\beta}_2 \sum X_1 X_j + \hat{\beta}_2 \sum X_2 X_j + \cdots + \hat{\beta}_{p+1} \sum X_j Xp \tag{3.78}$$

34 Violação de um pressuposto básico

Nos capítulos prévios temos tratado os modelos de regressão sob a perspectiva de não violar algum dos pressupostos básicos descritos anteriormente. No entanto, em muitas aplicações práticas e em particular séries econômicas que envolvem modelagem de regressão, onde o comportamento de uma variável de interesse é explicado a partir de sua relação com variáveis auxiliares assumindo-se, em geral, que esta relação seja linear. No caso de violar a suposição do modelo ser verdadeiro (adequado) isto pode ser explicado por não incluir em nosso modelo variáveis que são importantes ou por que incluímos sem poder explicativo, sendo assim a omissão de regressores relevantes causa viés, e a inclusão de regressores irrelevantes causa ineficiência.
Uma outra suposição constantemente feita é a de homoscedasticidade, ou seja, assume-se que todos os erros do modelo possuem variâncias idênticas. Esta suposição, contudo, é violada em muitas situações, em especial, quando o interesse reside na modelagem de dados de corte transversal. Neste caso, é muito comum que os dados apresentem heterocedasticidade, ou seja, variâncias condicionais não-constantes, ver Cribari-Neto, onde

$$y_i = \beta_1 + \beta_2 x_i + \sigma_i \epsilon_i \qquad i = 1, 2, \ldots n \tag{3.79}$$

a variância $V(\epsilon) = \sigma_i^2$, não é constante. Dessa forma, a variância da perturbação depende, por exemplo, da variável independente X.

Vale destacar que apenas sobre normalidade da estrutura de erros, o estimador de mínimos quadrados ordinários dos parâmetros coincide com o estimador de máxima verossimilhança. Além disso, sob normalidade, o estimador de MQO é o melhor estimador na classe dos estimadores não viesados, ou seja, o estimador de MQO é eficiente, Ferreira. Com estas hipóteses do modelo clássico de regressão linear, os estimadores MQO possuem certas características estatísticas desejáveis, sintetizadas nas proprie-

dades de MELNT (Melhor Estimativa Linear não-Tendenciosa). Contudo, na prática, como sabemos se a propriedade MELNT é valida? Por exemplo, como podemos verificar se os EMQO são não-viesados? Isto é feito através dos experimentos MONTE CARLO, que são basicamente experimentos de simulação realizados com auxilio de um computador. O modelo de regressão linear geral, descrito é da forma $y = X\beta + \epsilon$; onde y é um vetor $(n \times 1)$ de observações da variável dependente, X é uma matriz fixa de posto completo de dimensão $(n \times p)$, onde $p < n$, contendo observações sobre as variáveis explicativas. $\beta = (\beta_1, \beta_2, \cdots, \beta_p)$ é um vetor $(p \times 1)$ de parâmetros desconhecidos, considere ϵ como um vetor $(n \times 1)$ de distúrbios aleatórios (erros) com média 0 e matriz de covariância $\Omega = \text{diag}(\sigma_1^2, \sigma_2^2, \cdots, \sigma_n^2)$. Quando os erros são homoscedásticos, então $\sigma_i^2 = \sigma^2 > 0$, ou seja, $\Omega = \sigma^2 I_n$, onde I_n é a matriz identidade de ordem n. O estimador de mínimos quadrados ordinários de β é dado por $\widehat{\beta} = (X'X)^{-1}X'y$, cuja média é β (isto é, ele é não-viesado) e cuja variância é dada por

$$\Psi = (X'X)^{-1}(X'\Omega X)(X'X)^{-1}. \tag{3.80}$$

Quando todos os erros possuem a mesma variância, ou seja, $\Omega = \sigma^2 I_n$, esta expressão é simplificada para $\sigma^2(X'X)^{-1}$, podendo ser facilmente estimada como $\widehat{\sigma}^2(X'X)^{-1}$, onde $\widehat{\sigma}^{\,2} = \widehat{\epsilon}'\widehat{\epsilon}/(n-p)$. Aqui,

$$\widehat{\epsilon} = (I_n - X(X'X)^{-1}X')y = My$$

onde $\widehat{\epsilon}$ representa o vetor $(n \times 1)$ de resíduos de mínimos quadrados.

35 Aleatoriedade

Segundo Zafon, o termo aleatoriedade, baseado em conceitos estatísticos, significa algum evento que acontece com qualquer distribuição de probabilidade, onde normalmente se associa uma falta de influência ou correlação, a menos que especificado de outra forma. Em termos de computação, aleatoriedade refere-se à geração ou uso de um conjunto de sequências de números aleatórios dentro de algum conjunto finito.
Mas, o que é um número aleatório? Na verdade não existe um número aleatório, mas sim um conjunto sequencial de números independentes que possuem uma distribuição de probabilidade específica de ocorrer. Isso significa que cada número na sequência possui uma probabilidade de ocorrer independente do anterior e este número é obtido meramente ao acaso. Computacionalmente falando, existem técnicas para gerar uma sequência de números pseudo-aleatórios, pois é impossível gerar, apenas através de algoritmos computacionais, uma sequência de números realmente aleatória.

36 Gerador de Números Aleatórios

A disponibilidade de números aleatórios é extremamente útil em várias situações como, por exemplo, em amostragem de uma população ou em tomada de decisões. Algumas das propriedades que um bom gerador de números pseudo-aleatórios deve possuir:

a) Os números gerados devem seguir uma distribuição uniforme;

b) Os números devem ser estatisticamente independentes entre si;

c) A sequência não deve se repetir nunca (teoricamente isso é impossível mas, na prática, um período suficientemente grande é o bastante);

d) Os algoritmos de geração desses números devem ser rápidos de modo que os recursos computacionais possam ser concentrados nas simulações.

O processo de geração de números aleatórios, para ser eficiente, deve atender a alguns requisitos, como velocidade e conformidade. Infelizmente, esses requisitos, são contraditórios, pois geradores muito rápidos em geral não fazem operações mais complexas, o que permitiria melhor conformidade segundo Goldreich. Um outro trabalho de construção de geradores independentes de números aleatórios é dado por Zafon e Manacero. Nesse trabalho foi adotado um gerador congruencial linear como base do gerador uniforme. Uma definição formal de um Gerador de Números Aleatórios (GNA) pode ser encontrada em L'Ecuyer

Definição. Um GNA é uma estrutura $G = (S, s_0, f, U, g)$, onde

a) S é um conjunto finito de estados.
b) $s_0 \in S$ é o estado inicial do gerador, também chamado de semente.
c) f é a função de transição entre os estados.
d) U é um conjunto finito de estados de saída do gerador.
e) g é a função de saída do gerador. O GNA inicia na semente s_0 e $u_0 = g(s_0)$. A partir deste ponto, para todo $i = 1, 2, \ldots n$, tem-se $s_i = f(s_{i-1})$ e $u_i = g(s_i)$.
f) Dada a mesma semente, será produzida sempre a mesma sequência de números pseudo-aleatórios.
h) Valores sucessivos da seqüência devem ser independentes e uniformemente distribuídos.

Uma outra característica interessante de um GNA é seu período. Dado que S é finito, a sequência de estados é necessariamente periódica. O período é definido como o menor inteiro positivo p tal que $S_{p+n} = S_n$ para todo n.

O R tem como padrão de GNA, o Mersenne-Twister, este gerador de números pseudo-aleatório desenvolvido em 1997 por Makoto Matsumoto, e Takuji Nishimura, o qual se baseia em uma matriz de recorrência linear ao longo de um campo finito binário. Para a geração rápida ele fornece inteiros pseudo-aleatórios de alta qualidade, tendo sido projetado especificamente para corrigir falhas encontradas em algoritmos mais velhos. Seu nome deriva do fato de ser um primo Mersenne e que a duração do período é escolhido. Existem, pelo menos, duas variantes comuns do algoritmo, que apenas diferem no tamanho dos números primos de Mersenne utilizados. O mais novo e mais comumente usado é o Mersenne Twister $MT19937$, com comprimento de 32 bits. Há também uma variante com comprimento de 64 bits, $MT19937-64$, o qual gera uma sequência diferente. Para um comprimento de palavra de k bits, o Mersenne Twister, gera números inteiros com uma distribuição quase uniforme no intervalo $[0, 2^k - 1]$. Para gerar por exemplo 5 números aleatórios normais com semente inicial (20) de geração, usamos:

```
> set.seed(20) #Padrão é Mersenne-Twister
> rnorm(5)
[1]  1.1626853 -0.5859245  1.7854650 -1.3325937 -0.4465668

> set.seed(20, kind = "Mersenne-Twister")
> rnorm(5)
[1]  1.1626853 -0.5859245  1.7854650 -1.3325937 -0.4465668
```

Para alterar por outro gerador a critério do pesquisador, como alternativas temos:

```
> set.seed(20, kind = "Wichmann-Hill")
> set.seed(20, kind ="Marsaglia-Multicarry")
> set.seed(20, kind ="Super-Duper")
> set.seed(20, kind ="Knuth-TAOCP-2002")
> set.seed(20, kind ="Knuth-TAOCP")
> set.seed(20, kind ="L'Ecuyer-CMRG")
```

37 Testes de Normalidade

Dentro da literatura existem diversos testes propostos para avaliar se uma determinada amostra provém de uma população com distribuição normal ou não normal. Geralmente, isto é medido com o chamado poder do teste (conhecido como $1-\beta$, onde β é a probabilidade de aceitar a hipóteses nula H_0, dado que esta é falsa).

Ferreira, apresenta um estudo de análise de sensibilidade dos testes de normalidade de Jarque-Bera e Lilliefors em modelos de regressão linear, violando alguns dos pressupostos básicos como o de não autocorrelação das perturbações, e homoscedasticidade. Para o estudo utiliza 10000 réplicas de Monte Carlo, com tamanhos de amostras de $10, 50, 100, 500$ e 1000 e níveis de significância de $10\%, 5\%$, e 1%. O autor concluiu que em modelos de regressão com estrutura de erros não correlacionados e homoscedásticos, os testes de Jarque-Bera e Lilliefors apresentaram comportamento bastante satisfatórios. Além disso, quando há autocorrelação fraca dos erros os desempenhos dos testes não sofrem alterações significativas; já na autocorrelação forte os desempenhos dos testes ficam comprometidos, pois as taxas de rejeição superam os valores nominais dos níveis de significância sugeridos acima. Para o caso da violação do pressuposto de homoscedasticidade (heteroscedasticidade), as taxas de rejeição atingem 100% em muitos casos. Ainda apresenta o poder do teste, concluindo que os testes apresentaram bom poder, destacando-se o teste de Jarque-Bera na maioria dos casos.

37.1 Teste Jarque-Bera (JB)

Trata-se de um teste assintótico. As hipóteses a serem testadas são:

H_0 : o erro ϵ_i do modelo de regressão linear possui distribuição normal

contra

H_1 : o erro ϵ_i do modelo de regressão linear possui distribuição não normal.

O erro do tipo I dado por $\alpha = \mathrm{P}(\text{Rejeitar } H_0 | H_0 \text{ é verdadeira})$, escolhe-se o nível de significância, que tipicamente é 1%, 5% e 10%. Para a construção do teste de Jarque-Bera utilizaremos as medidas de tendência central:

$$\bar{\epsilon} = \frac{1}{n}\sum_{i=1}^{n} \epsilon_i \tag{3.81}$$

onde n é o número de observações da amostra, os ϵ_i são os erros não observáveis da amostra, e $\bar{\epsilon}$ é a média amostral, conhecido também como momento centrado em torno da média de ordem 1. Temos,

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(\epsilon_i - \bar{\epsilon})^2, \tag{3.82}$$

$$DP = \sqrt{\widehat{\sigma}^2}, \tag{3.83}$$

onde $\widehat{\sigma}^2$ é definido como estimação da variância dos erros, determinando a variabilidade dos dados, conhecido também como momento centrado em torno da média de ordem 2. DP é comumente conhecido como desvio padrão. Extensões destes momentos são

$$\widehat{\sigma}^3 = \frac{1}{n}\sum_{i=1}^{n}(\hat{\epsilon}_i - \bar{\epsilon})^3, \tag{3.84}$$

$$A = \widehat{\sigma}^3/(DP)^3, \tag{3.85}$$

onde $\widehat{\sigma}^3$ é conhecido como momento centrado em torno da média de ordem 3. O valor de A mede o grau de assimetria da curva, que pode ser assimétrica negativa ($A > 0$), assimétrica positiva ($A < 0$), ou simétrica ($A = 0$).

Outro momento que é necessário na construção do teste Jarque-Bera é

$$\widehat{\sigma}^4 = \frac{1}{n}\sum_{i=1}^{n}(\hat{\epsilon}_i - \bar{\epsilon})^4 \tag{3.86}$$

$$k = \widehat{\sigma}^4/(DP)^4, \tag{3.87}$$

onde k é denominado de Curtose, podendo ser platicurtose ou achatada se $k < 3$, mesocurtose ou normal se $k = 3$ e lepto curtose se $k > 3$.
A estatística de Jarque-Bera é dada por

$$JB = n\left[\frac{A^2}{6} + \frac{(k-3)^2}{24}\right]. \tag{3.88}$$

No R use a função: `jarque.bera.test()` do pacote `tseries`

37.1.1 exemplo1

considere um objeto do tipo lm, como os resíduos do modelo m1:

```
> library(tseries)
> m1= lm(valor ~ consumo + diasconsumo + Construindo - 1)
> jarque.bera.test(residuals(m1))

        Jarque Bera Test

data:  residuals(m1)
X-squared = 4.1774, df = 2, p-value = 0.1238
```

37.1.2 exemplo2

considere a construção de números aleatórios de várias distribuições:

```
> set.seed(20)

> f=runif(1000)

> g=rexp(1000,1)

> h=rnorm(1000)

> k=rnorm(1000)


> jarque.bera.test(diff(f-g))

        Jarque Bera Test

data:  diff(f - g)
X-squared = 116.4544, df = 2, p-value < 2.2e-16

> jarque.bera.test(diff(g-h))

        Jarque Bera Test

data:  diff(g - h)
```

```
X-squared = 24.2139, df = 2, p-value = 5.521e-06


> jarque.bera.test(diff(k-h))

        Jarque Bera Test

data:  diff(k - h)
X-squared = 0.613, df = 2, p-value = 0.736
```

37.2 Teste Lilliefors

O teste de Lilliefors é uma adaptação do Teste de Kolmogorov-Smirnov para testar normalidade. Como antes, queremos estudar erros do tipo I. Dada uma amostra $\hat{\epsilon}_i$, para $i = 1, 2, \ldots, n$, procede-se da seguinte forma:

1. Calcula-se a média amostral
$$\overline{\epsilon} = \frac{1}{n} \sum \hat{\epsilon}_i \tag{3.89}$$

2. Calcula-se a variância amostral
$$s^2 = \frac{1}{n-1} \sum (\hat{\epsilon}_i - \overline{\epsilon})^2 \tag{3.90}$$

3. Transformamos a amostra, calculando
$$Z_i = \frac{\hat{\epsilon}_i - \overline{\epsilon}}{s} \tag{3.91}$$

4. Para cada valor Z_i, calcula-se a proporção $\mathcal{L}(Z_i)$ dos valores da amostra que não excedem Z_i.
5. Finalmente, determina-se a probabilidade $\mathcal{N}(Z_i)$ de que Z_i tenha sido obtido de uma normal com média 0 e desvio-padrão 1.

O critério de Lilliefors é baseado no número

$$L = \max_i \left\{ |\mathcal{L}(Z_i) - \mathcal{N}(Z_i)|, |\mathcal{L}(Z_i) - \mathcal{N}(Z_{i-1})| \right\}. \tag{3.92}$$

Os quantis da distribuição dessa estatística podem ser encontrados, por exemplo, em Connover (1999). No R, use a função `lillie.test()`, do pacote `nortest`.)

37.2.1 exemplo1

considere um objeto do tipo lm, como os resíduos do modelo m1:

```
> library(nortest)

> m1= lm(valor ~ consumo + diasconsumo + Construindo - 1)

> lillie.test(residuals(m1))


        Lilliefors (Kolmogorov-Smirnov) normality test

data:  residuals(m1)
D = 0.0804, p-value = 0.3729
```

37.2.2 exemplo2

considere a construção de números aleatórios de várias distribuições:

```
> set.seed(20)

> f=runif(1000)

> g=rexp(1000,1)

> h=rnorm(1000)

> k=rnorm(1000)


> lillie.test(diff(f-g))

         Lilliefors (Kolmogorov-Smirnov) normality test

data:  diff(f - g)
D = 0.0614, p-value = 1.752e-09


> lillie.test(diff(g-h))
```

```
        Lilliefors (Kolmogorov-Smirnov) normality test

data:  diff(g - h)
D = 0.0296, p-value = 0.03907



> lillie.test(diff(k-h))

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  diff(k - h)
D = 0.0189, p-value = 0.5224
```

38 Simulação de Monte Carlo

Simulação de Monte Carlo em modelos de regressão traz a confirmação da verdadeira distribuição dos parâmetros quando n é grande. A teoria assintótica reafirma o teorema central do limite, onde qualquer conjunto de observações independentes e identicamente distribuídas, a medida que n cresce a distribuição da média amostral desta variável segue uma normal. Um comentário sobre os experimentos Monte Carlo, utilizando a função de regressão populacional de duas variáveis, é estimado por $y = \beta_1 + \beta_2 * x_i + \epsilon_i$ Os experimentos de Monte Carlo é composto dos seguintes passos:

- Suponha que no modelo $y = \beta_1 + \beta_2 * x_i + \epsilon_i$ os valores reais dos parâmetros sejam $\beta_1 = 20$ e $\beta_2 = 6$.
- Escolha o tamanho da amostra, digamos $n = 100$.
- Fixe os valores do regressor para cada observação. Ao todo você terá 100 valores de x .
- Suponha que, de uma tabela de números aleatórios normais com média zero e variância=20, você escolha 100 valores e os chame de ϵ_i.
- Como você conhece β_1, β_2, x_i e ϵ_i, obterá 100 valores de y_i .

- Utilizando agora os 100 valores de y_i assim gerados, você os regressa nos 100 valores de x_i escolhidos no passo 3, obtendo os estimadores de mínimos quadrados: $\hat{\beta}_1$ e $\hat{\beta}_2$.
- Suponha que você repita este experimento mais 999 vezes, utilizando a cada vez os mesmos valores de β_1, β_2 e x_i. Naturalmente, os valores ϵ_i irão variar de um experimento para outro. Por tanto você tem ao todo 1000 experimentos, gerando assim 1000 valores de : $\hat{\beta}_1$ e $\hat{\beta}_2$.
- Tome as médias dessas 1000 estimativas e as chame de $\bar{\hat{\beta}}_1$ e $\bar{\hat{\beta}}_2$.
- Se esses valores médios forem quase iguais aos verdadeiros valores de β_1 e β_2, admitidos por hipóteses no passo 1, o experimento Monte Carlo "demonstra"que os estimadores por mínimos quadrados são realmente não-viesado ou não-tendenciosos.

Estes passos caracterizam a natureza geral dos experimentos de Monte Carlo.

38.1 Implementação no R

Para exemplificar um número de 1000 simulações (repetições) de experimentos de Monte Carlo de um modelo de regressão linear simples sem violação de um pressuposto básico e proposto a seguir, usamos uma semente inicial. A variável independente x, é gerada a partir de uma amostra aleatória de uma distribuição normal com média 20 e desvio padrão 5. Na saída são apresentados os histogramas e a função acumulada correspondente a cada estimativa.

```
mc.sim=function(r=1000)
{
set.seed(100)
beta1=20
beta2=0.6
x=rnorm(100,20,5)
sigma2=20
n=length(x)
x=as.matrix(x)
X=cbind(1,x)
lin.pred=beta1+beta2*x
y.simulated=lin.pred[,1]+matrix(rnorm(n*r, mean=0, sd=sqrt(sigma2)),n,r)
estimate=solve(t(X)%*%X)%*%t(X)%*%y.simulated
par(mfrow=c(2,2), pty="s")
```

```
hist(estimate[1,], col="red", breaks=12, xlab="estimativa de b1",
main="histograma de b1", ylab="frequência", xlim=c(15,25))
hist(estimate[2,], col="blue", breaks=12, xlab="estimativa de b2",
main="histograma de b2", ylab="frequência", xlim=c(0.4,0.8))
plot(ecdf(estimate[1,]),main="distribuição de b1")
plot(ecdf(estimate[2,]),main="distribuição de b2")
return(c(mean(estimate[1,]),mean(estimate[2,])))
}
```

Para executar esta função utilizamos o seguinte comando:

```
mc.sim(1000)
[1] 19.9941537  0.6006474
```

Observação: podemos verificar o teorema de Gauss-Markov, que a medida que aumentamos o tamanho da amostra as estimativas dos parâmetros não apresentam viés e a sua variância são eficientes.

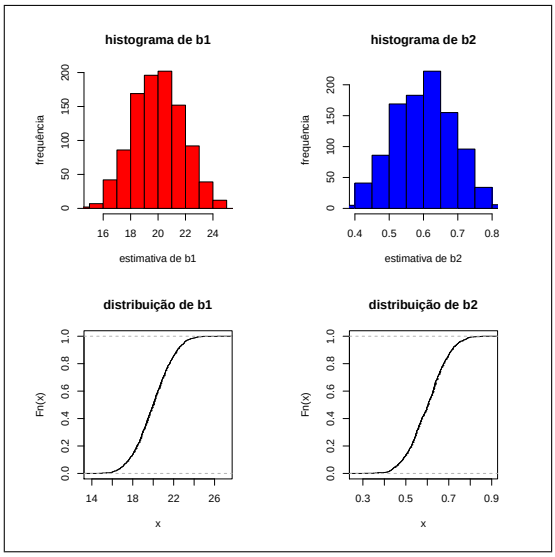


Figura 19. Simulação MC de um modelo de regressão linear simples.

Observação: Para amostras pequenas, a variabilidade de β_1e β_2 é grande. (verifique).
Se assumirmos o fato do modelo ser não homocedástico, podemos modificar a linha 7 da seguinte forma:

```
sigma2=rchisq(length(x),2*x^2) # com a saída seguinte:
mc.sim(1000)
[1] 21.9977275  0.6001847
```

Se além do modelo ser não homocedástico acrescentarmos outra violação do pressuposto básico de não normalidade dos erro (erros uniformes por exemplo), podemos modificar a linha da seguinte forma:

```
 y.simulated=lin.pred[,1]+matrix(runif(n*r,-10, 10),n,r),
```

A saída é:

```
> mc.sim(1000)
[1] 20.0170569  0.5995204
```

39 Simulação de modelos

39.1 Caso 1. Modelo de Regressão Linear Simples

O seguinte código no R mostra a simulação de uma amostra de n observações para o modelo: $y = \beta_0 + \beta_1 * x + \epsilon$. Onde y é a variável dependente e x a variável independente ou determinística, ϵ é o erro aleatório distribuído normalmente com média zero e desvio padrão σ; β_0 o intercepto, β_1 a tangente o variação de y, quando x acontece. Considere $n = 10$, $\beta_0 = 2$, $\beta_1 = 3$, e $\sigma = 2.5$.

```
options(digits=3); set.seed(10)
n=10; x=seq(1,n); sigma=2.5; b0=2; b1=3
erro=rnorm(n,sd=sigma)
y=b0+b1*x+erro
t(data.frame(x,y))
```

39.2 Caso 2. Simulações que serão ajustados a partir de modelos em um banco de dados

Considere o consumo de água em uma residência em 65 meses, com as seguintes variáveis: valor pago em reais (valor), consumo de água em m3 (consumo), dias de consumo contabilizada pela empresa fornecedora da água (diasconsumo) e a variável dicotômica, sendo 1: reformando a casa e 0: caso contrario (Construindo).

```
ca=read.table(file.choose(), header=T) #consumoagua2
attach(ca)
m1=lm(formula = valor ~ consumo + diasconsumo + Construindo)
m1.sim=simulate(m1,nsim=20)
```

Exercício: Implemente este algoritmo e interprete para:
1. O numero de simulações é 1000.
2. apply(m1.sim,1,mean); apply(m1.sim,2,mean)
3. tapply(ca[,2],apply(m1.sim,1,mean),mean)

39.3 Caso 3. Usando Bootstrapping

O procedimento Bootstrap é uma técnica de re-amostragem, bastante utilizada em diferentes situações estatísticas. A base da técnica é a obtenção de um novo conjunto de dados, por reamostragem do conjunto de dados original (Efron e Tibishirani, 1993).

39.4 Desempenho da variação de consumo

```
library(boot)
ca.fn=function(ca,index){
ca.reamostra=ca[index,]
ca.lm=lm(valor~consumo, data=ca.reamostra)
coef(ca.lm)[2]  #coeficiente de variação doconsumo
}
set.seed(10)
ca.boot=boot(ca,R=999, statistic=ca.fn)
ca.boot
plot(ca.boot)
```

Exercício: Implemente este algoritmo para: 1. O parâmetro do intercepto 2. O tamanho do bootstraps seja 1000.

39.4.1 IC das previsões do valor pago

```
attach(ca)
library(boot)
ca2.fn=function(ca,index){
ca.reamostra=ca[index,]
ca2.lm=lm(valor~consumo, data=ca.reamostra)
```

```
predict(ca2.lm, newdata=data.frame(consumo=c(20,40)))
#intervalo de confiança das previsões do valor pago pelo consumo de água.
}
set.seed(10)
ca2.boot=boot(ca,R=999, statistic=ca2.fn)
 ca2.boot
plot(ca2.boot)
par(mfrow=c(2,2))  #saida gráfica 2X2
plot(ca2.boot$t); hist(ca2.boot$t), qqnorm(ca2.boot$t); boxplot(ca2.boot$t)
# adicionais da reamostragem
```

Exercício: Implemente este algoritmo para: 1. Modelos de regressão múltipla.
2. O tamanho do bootstraps seja 1000.
3. Identifique e interprete os objetos de ca2.boot com o comando summary.

39.4.2 Algoritmo para o AIC

```
caAIC.fn=function(ca,index){
ca.reamostra=ca[index,]
ca2.lm=lm(valor~consumo, data=ca.reamostra)
AIC(ca2.lm) #IC para o critério AKAIKE no modelo valor~consumo.
}
set.seed(10)
ca2.boot=boot(ca,R=999, statistic=caAIC.fn)
ca2.boot
plot(ca2.boot)
```

39.5 caso 4. Estimação Bayesiana

Para modelos de regressão linear simples com Estimação Bayesiana usamos Cadeias de Markov Monte Carlo (MCMC): Esta técnica gera sucessivas estimativas dos parâmetros considerando estes parâmetros como variáveis aleatórias. Este processo de simulação descarta as primeiras estimativas pois inicia-se o aquecimento da cadeia (geralmente os primeiros 1000 chamados de burnings), e obtém a distribuição a posterior dos parâmetros (maiores detalhes Jim Albert, 2009).

39.5.1 Função MCMCregress()

O seguinte código no R mostra a simulação de uma amostra de n observações para o modelo: $y = \beta_0 + \beta_1 * x + \epsilon$. Onde y é a variável dependente e x a variável independente ou determinística, ϵ é o erro aleatório distribuído normalmente com média zero e desvio padrão σ; β_0 o intercepto, β_1 a tangente o variação de y, quando x acontece.

```
library(MCMCpack)
ca.mcmc=MCMCregress(valor~consumo)
summary(ca.mcmc)
plot(ca.mcmc)
Iterations = 1001:11000 #descarta as primeiras 1000 estimativas
Thinning interval = 1 ; Number of chains = 1; Sample size per chain = 10000

1. Empirical mean and standard deviation for each variable,
   plus standard error of the mean:

               Mean     SD     Naive SE  Time-series SE
(Intercept) -12.351   1.7301   0.017301   0.017301
consumo       3.006   0.1325   0.001325   0.001325
sigma2       14.324   2.6546   0.026546   0.027411

2. Quantiles for each variable:
               2.5%      25%       50%       75%      97.5%
(Intercept)  -15.728   -13.494   -12.334   -11.205   -8.996
consumo        2.745     2.919     3.005     3.092    3.265
sigma2        10.058    12.450    13.990    15.884   20.423
```

No caso da convergência da cadeia podemos usar a fração da cadeia para verificar se as diferencias das médias das estimativas dos parâmetros são iguais. Este teste é chamado de Geweke. Pra implementar no R, realizamos o seguinte comando:

```
geweke.diag(ca.mcmc, frac1=0.1, frac2=0.5)
```

Exercicios: 1. Implemente este algoritmo para o caso de regressão múltipla por exemplo o valor consumo+diasconsumo+Construindo
2. Implemente esta técnica para as variáveis econômicas disponíveis em "decon.txt". Importante Interprete.

39.5.2 Função blinreg()

Regressão Bayesiana usando a função blinreg do pacote LearnBayes. Útil para encontrar a distribuição a posterior conjunta dos parâmetros de um modelo a partir de simulações. Considere os dados de consumo de água, onde o valor pago esta em função do consumo, o modelo linear com o intercepto β_1 e a variação pelo consumo β_2 é simulado por:

```
library(LearnBayes)
md=lm(valor~consumo,x=T,y=T)
tsample=blinreg(md$y,md$x,5000)
par(mfrow=c(3,2))
hist(tsample$beta[,1])
hist(tsample$beta[,2])
boxplot(tsample$beta[,1], col=4)
boxplot(tsample$beta[,2], col=2)
plot(density(tsample$beta[,1], col=4))
plot(density(tsample$beta[,2], col=2))
```

39.6 Exercícios:

- Implemente simulações de experimentos de Monte Carlo de um modelo de Regressão quadrática sem violação de um pressuposto básico.
- Implemente simulações de experimentos de Monte Carlo de um modelo de Regressão quadrática com violação de não normalidade dos erros.
- Implemente simulações de experimentos de Monte Carlo de um modelo de Regressão quadrática com violação da homogeneidade da variância dos erros.

40 Heterocedasticidade

Homoscedasticidade é a variância constante dos resíduos. Esta é uma propriedade fundamental, que deve ser garantida, sob pena de invalidar toda a análise estatística nos modelos de regressão. Deseja-se que os erros sejam aleatórios, se isto não ocorre, há heterocedasticidade. Significa dizer que ha chances de ocorrerem erros grandes (ou pequenos). Há tendências nos erros. Por exemplo, se na avaliação de terrenos a equação obtida indica erros maiores para os imóveis mais caros, progressivamente (quanto maior o imóvel, maior o erro), não há variância constante.

As consequências da heterocedasticidade são que as estimativas dos parâmetros da regressão os $\hat{\beta}_i$ não são tendenciosas (não viesados), continua sendo consistente, mas o teorema de Gaus Markov deixa de valer ou seja as estimativas dos parâmetros são ineficientes e suas respectivas variâncias são tendenciosas. Os testes t e F tendem a dar resultados incorretos. Neste caso, os resultados não são confiáveis, ou seja, o modelo pode parecer bom, mas ele não é adequado aos dados, na verdade. A heterocedasticidade inicialmente pode ser verificada através de gráficos de resíduos. Os gráficos dos resíduos contra os valores reais e contra os valores calculados pela equação são importantes. Se os pontos estão distribuídos aleatoriamente, sem mostrar um comportamento definido, há homoscedasticidade. Mas se existe alguma tendência (crescimento/decrescimento/oscilação), então há heterocedasticidade. Havendo heterocedasticidade, podem ser tentadas transformações nas variáveis (geralmente logarítmicas) ou outras soluções mais complexas. O modelo deve ser modificado. No modelo heterocedástico assumimos que a matriz de variâncias e covariâncias dos erros é dada por:

$$\Phi = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 & 0 \\ 0 & \sigma_2^2 & \cdots & 0 & 0 \\ \vdots & & & & \vdots \\ 0 & 0 & \cdots & 0 & \sigma_n^2 \end{bmatrix}$$

onde $\Phi = dig(\sigma_1^2, \sigma_2^2, \cdots, \sigma_n^2)$ A covariancia dos parâmetros estimados é dado por:

$$Cov(\hat{\beta}) = (X'X)^{-1}X'\Phi X(X'X)^{-1} \tag{3}$$

Para exemplificar situações onde a presença de heterocedasticidade é comum, considere os índices de cotação da Bovespa ($\times 1000$) em formato "ibovm.txt", durante os períodos de $2001-2010$, disponibilizados pela média mensal a seguir:

```
 2001  2002  2003  2004  2005  2006  2007  2008  2009  2010
17.12 13.40 11.49 23.41 24.43 36.19 43.44 59.65 39.57 68.58
16.60 13.23 10.35 21.97 26.55 37.55 45.16 62.54 40.17 65.94
15.33 13.90 10.92 21.92 27.67 37.77 44.00 61.54 39.48 69.07
14.43 13.34 12.09 21.81 25.51 39.19 48.05 64.24 45.22 69.74
14.69 12.52 13.09 18.88 24.81 39.04 51.23 71.21 50.89 62.58
14.87 11.69 13.49 20.22 25.43 35.07 53.65 67.23 52.06 63.33
13.95 10.31 13.55 21.74 25.25 36.30 56.20 59.77 52.07 64.14
13.41  9.79 14.01 22.27 27.01 36.92 52.16 55.46 56.66 66.58
11.01  9.64 16.09 22.70 29.86 36.17 56.36 50.59 59.20 67.79
```

```
10.96  9.18 17.78 23.37 29.84 38.63 62.68 38.14 63.99 70.62
12.80 10.03 19.01 24.05 31.15 41.20 62.45 35.91 66.00 70.38
13.32 10.84 21.16 25.54 33.13 43.32 63.47 37.56 68.10 68.55
```

No R os dados são disponibilizados seguindo os seguintes comandos:

```
> ibov=read.table("ibovm.txt",header=T)
> colnames(ibov)=c("2001","2002","2003","2004","2005","2006","2007",
  "2008","2009","2010")
> ibovm=apply(ibov,2,mean)
> xi=c(1:10)
> mi=lm(ibovm~xi-1)
> summary(mi)

Call:
lm(formula = ibovm ~ xi - 1)

Residuals:
   Min     1Q Median     3Q    Max
-5.635 -4.587 -1.163  3.135  7.800

Coefficients:
   Estimate Std. Error t value Pr(>|t|)
xi    6.491      0.259   25.06 1.23e-09 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 5.082 on 9 degrees of freedom
Multiple R-squared: 0.9859,     Adjusted R-squared: 0.9843
F-statistic: 628.1 on 1 and 9 DF,  p-value: 1.232e-09

> anova(mi)
Analysis of Variance Table

Response: ibovm
          Df  Sum Sq Mean Sq F value    Pr(>F)
xi         1 16221.3 16221.3  628.08 1.232e-09 ***
Residuals  9   232.4    25.8
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Na saída acima, considerando as médias anuais determinamos um modelo de regressão adequado, porem considerando os meses da cotação, pode negar o pressuposto da variância constante, um boxplot, monstra visualmente a presença de heterocedasticidade.

```
> bplot(ibov2,ylab="cotação mensal", main="Bovespa - Brasil")
```

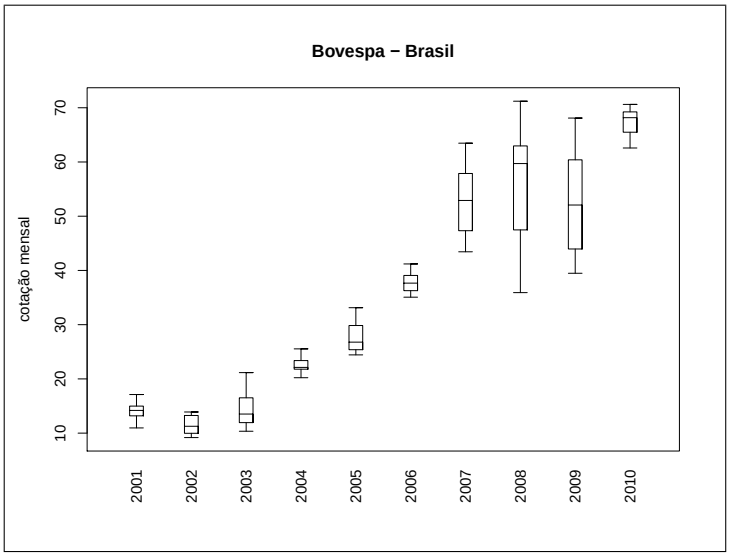


Figura 21. Cotação da bovespa por ano.

40.1 Testes de Heterocedasticidade

Na literatură á vários testes que identificam a presença de heterocedasticidade, nesta unidade, mostraremos apenas aqueles que são disponibilizados pelo pacote **lmtest** do R.

40.1.1 Teste de Quandt-Goldfeld

Os passos deste teste são os seguintes:

1. $H0 : \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_n^2$
   $Ha : \sigma_1^2 \neq \sigma_2^2, ou, \sigma_1^2 \neq \sigma_3^2 \cdots$, pelo menos uma desigualdade estrita.
2. Ordenar os dados segundo a magnitude de X.

3. Eliminar c observações amostrais, em geral $c = 1/4$ dos dados. O teste fica mais poderoso
4. Estimam-se as duas regressões separadas pelas c observações.
5. A estatística de prova é a F_0, com a $SQRes_1/SQRes_2$.
6. Os graus de liberdade é $[(n-c)/2-p-1]/[(n-c)/2-p-1]$, p é o número de parâmetros
7. Se o F_0 é próximo de 1, não há presença de Heterocedasticidade.
8. Se o F_0 se afastar de 1, a presença de heterocedasticidade é seria.

Onde $SQRes$, representa a soma de quadrados do resíduo.
Para implementar o teste, considere a função **gqtest()** do pacote **lmtest**, aplicando no modelo **mi**, onde:

```
> library(lmtest)
> gqtest(mi)

        Goldfeld-Quandt test

data:  mi
GQ = 1.3331, df1 = 4, df2 = 4, p-value = 0.3936
```

Pelo teste de Quandt-Goldfeld, pode-se afirmar que para dados anuais do Bovespa, não há evidencias para a presença de heterocedasticidade.
Este mesmo teste é aplicado para a cotação diária da Bovespa, (a cotação foi extraída do IPEA-DATA), os resultados são dados a seguir:

```
> bov=read.table("bovespa.txt",header=T)
> dim(bov)
[1] 2703    1
> dias=c(1:2703)
> bov=cbind(dias,bov)
> attach(bov)
The following object(s) are masked _by_ '.GlobalEnv':

    dias
> md=lm(cotacao~dias)
> summary(md)
```

```
Call:
lm(formula = cotacao ~ dias)

Residuals:
   Min     1Q Median     3Q    Max
-23607  -6419  -1038   6218  23101

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 2236.0514   313.5662   7.131 1.27e-12 ***
dias          23.4455     0.2009 116.717  < 2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 8149 on 2701 degrees of freedom
Multiple R-squared: 0.8345,     Adjusted R-squared: 0.8345
F-statistic: 1.362e+04 on 1 and 2701 DF,  p-value: < 2.2e-16

> gqtest(md)

        Goldfeld-Quandt test

data:  md
GQ = 4.5783, df1 = 1350, df2 = 1349, p-value < 2.2e-16
```

Por tanto há evidencias da presença de heterocedasticidade num nível mais desagregado (dias úteis).

40.1.2 Teste de Breusch-Pagan

O teste anterior é exato, não envolvendo aproximação, no teste a seguir assumimos que:

$$\sigma_i^2 = h(\alpha_1 + \alpha_2 Z_{i2} + \cdots, \alpha_s Z_{is}) \tag{4}$$

onde $Z's$ são as variáveis que afetam as variâncias, h é uma função duas vezes diferenciável. As hipóteses são:
$H0 : \alpha_1 = \alpha_2 = \cdots = \alpha_s$ (Homocedástico)
$Ha : min|\alpha_1|, |\alpha_2|, \cdots, |\alpha_s| > 0$ (Heterocedástico)

Estatística do teste
Seja $\hat{\epsilon}_i$, i-ésimo resíduo de MQO, obtenha $\tilde{\sigma}^2 = \frac{\sum \hat{\epsilon}_i^2}{n}$, considere

$$m_i = \hat{\epsilon}_i^2 - \tilde{\sigma}^2 \tag{5}$$

e regredindo m em Z, obtenha a soma de quadrados do resíduo da regressão auxiliar (SQR_A), e faça $BP = (SQR_A)/2$, forma alternativa

$$BP = \frac{m'Z(Z'Z)^{-1}Z'm}{2\tilde{\sigma}^4} \tag{6}$$

Um problema é que a distribuição de BP é desconhecida, uma aproximação sobre $H0$, é dada pela distribuição qui-quadrado ($\chi^2_{s-1}gl$)
Regra de decisão: Rejeita-se $H0$ se $BP > \chi^2_{\alpha,s-1}$. Este teste é aproximado e geral, onde não precisamos especificar h.

40.1.3 Teste de Koenker's

As hipóteses do teste anterior permanecem válidas, porem no teste anterior não funciona bem não violação de normalidade dos erros, por tanto uma modificação de padronização é feita no teste BP, da seguinte forma:

$$BP_m = n * R_A^2 \tag{7}$$

ou alternativamente

$$n\frac{m'Z(Z'Z)^{-1}Z'm}{m'm} \tag{8}$$

que tem aproximação da distribuição qui-quadrado($\chi^2_{s-1}gl$)

Implementando os dois testes nos dados diários da Bovespa, temos

```
> bptest(md,studentize=FALSE) # Teste BP

        Breusch-Pagan test

data:  md
BP = 14.9066, df = 1, p-value = 0.000113

> bptest(md,studentize=TRUE)  # Teste Koenker

        studentized Breusch-Pagan test
```

```
data:  md
BP = 19.8011, df = 1, p-value = 8.593e-06
```

Nos dois testes concluímos que não há evidencias para aceitar $H0$, em favor de aceitar que pelo menos há duas variâncias dos erros distintas. Alternativa: este teste pode ser usado o **ncv.test()** do pacote **car**.

40.1.4 Harrison McCabe

O teste Harrison-McCabe fracciona a soma de quadrados do resíduos a partir de um certo ponto, A hipótese nula é determinada pelo tamanho da fração que como padrão assume o valor 0.5, define-se, assim $H0 =$ a fração é 0.5, a hipótese nula é rejeitada se o valor for menor que 0.5. Implementando o teste no R com o pacote **lmtest** temos:

```
> hmctest(md)

        Harrison-McCabe test

data:  md
HMC = 0.42, p-value < 2.2e-16
```

40.2 Corrigindo a Heterocedasticidade

Intuitivamente quando X é pequeno, a variabilidade é pequena, indicando que as observações são muito informativas sobre a reta, e recebem peso alto na estimação de mínimos quadrados ponderados, quando X é grande a variabilidade é grande, então as observações são pouco informativas e elas recebem peso baixo no processo de estimação. Para corrigir a Heterocedasticidade conhecendo σ^2, podemos usar o método de mínimos quadrados ponderados, que no modelo de regressão simples é dado por:

$$\frac{y_i}{\sigma} = \frac{\beta_0}{\sigma} + \frac{\beta_1 X_1}{\sigma} + v_i \tag{9}$$

Nosso exemplo de regressão com os dados agregados anuais da Bovespa, podemos eliminar a heterocedasticidade, da seguinte forma

$$\frac{cotacao_i}{\sigma} = \frac{\beta_1 ano_i}{\sigma} + v_i \tag{10}$$

onde: $v_i = \epsilon_i/\sigma$

Implementado no R temos:

```
> lmh=lm(ibovm/ibovsd ~ xi/ibovsd -1)
> summary(lmh)

Call:
lm(formula = ibovm/ibovsd ~ xi/ibovsd - 1)

Residuals:
    Min      1Q  Median      3Q     Max
-2.5566 -1.9388  0.7331  2.0788  4.7476

Coefficients:
          Estimate Std. Error t value Pr(>|t|)
xi         3.04248    0.25257  12.046 2.08e-06 ***
xi:ibovsd -0.22827    0.03353  -6.807 0.000137 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 2.624 on 8 degrees of freedom
Multiple R-squared: 0.9589,     Adjusted R-squared: 0.9486
F-statistic: 93.32 on 2 and 8 DF,  p-value: 2.854e-06
```

Para corrigir a presença de heterocedasticidade quando σ^2 é desconhecido, podemos admitir que σ^2 é diretamente proporcional com a variável regressora que ordena os dados (no caso os dias de cotação), assim podemos dividir na equação do modelo (md) os dias, para tal efeito propomos dois modelos, o primeiro modelo dividido pela raiz dos dias:

```
> cr=cotacao/sqrt(dias)
> dr=sqrt(dias)
> invb=1/sqrt(dias)
> lm1=lm(cr~dr+invb)
> summary(lm1)

Call:
lm(formula = cr ~ dr + invb)

Residuals:
     Min       1Q   Median       3Q      Max
-2024.03  -142.58   -27.03   130.38   646.89

Coefficients:
```

```
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1070.5233    18.9513  -56.49   <2e-16 ***
dr             37.7680     0.4338   87.07   <2e-16 ***
invb        19986.7852   128.9229  155.03   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 215.8 on 2700 degrees of freedom
Multiple R-squared: 0.8996,     Adjusted R-squared: 0.8995
F-statistic: 1.209e+04 on 2 and 2700 DF,  p-value: < 2.2e-16
```

O segundo modelo dividido pelos dias.

```
> cd=cotacao/dias
> invd=1/dias
> lm2=lm(cd~invd)
> summary(lm2)

Call:
lm(formula = cd ~ invd)

Residuals:
    Min      1Q  Median      3Q     Max
-411.99  -11.43    1.20   10.24  262.61

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 7.598e+00  3.154e-01   24.09   <2e-16 ***
invd        1.666e+04  1.279e+01 1302.94   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 16.26 on 2701 degrees of freedom
Multiple R-squared: 0.9984,     Adjusted R-squared: 0.9984
F-statistic: 1.698e+06 on 1 and 2701 DF,  p-value: < 2.2e-16
```

40.3 Estimador HC0 de White

Uma questão que precisamos resolver é encontrar um bom estimador para $Cov(\hat{\beta})$ na presença de heterocedasticidade. Sobre homoscedasticidade

temos

$$Cov(\hat{\beta}) = \sigma^2(X'X)^{-1}$$

e por tanto

$$\widehat{Cov}(\hat{\beta}) = \hat{\sigma}^2(X'X)^{-1} \tag{11}$$

no caso de heterocedasticidade, é estimada por

$$\widehat{Cov}(\hat{\beta}) = \hat{\Psi}_0 = (X'X)^{-1}X'\hat{\Phi}_0X(X'X)^{-1} \tag{12}$$

Para estimar a consistência de Φ, que tem como estimador $\hat{\Phi}_0$ dado por:

$$\hat{\Phi}_0 = \begin{bmatrix} \hat{\epsilon}_1^2 & 0 & \cdots & 0 & 0 \\ 0 & \hat{\epsilon}_2^2 & \cdots & 0 & 0 \\ \vdots & & & & \vdots \\ 0 & 0 & \cdots & 0 & \hat{\epsilon}_n^2 \end{bmatrix}$$

Este estimador é consistente sobre homoscedasticidade e heterocedasticidade de forma desconhecida, pois:

$$plim(\widehat{cov}(\hat{\beta})(\widehat{cov}(\hat{\beta})^{-1} = I_p, \qquad n \to \infty \tag{13}$$

Mesmo sem normalidade dos erros e sem homoscedasticidade, $\hat{\beta}_j \xrightarrow{d} N(\beta_j, V(\hat{\beta}_j))$, para $j = 1, 2, \cdots, p$, assim:

$$\frac{\hat{\beta}_j - \beta_j}{\sqrt{V(\hat{\beta}_j)}} \xrightarrow{d} N(0,1) \tag{14}$$

O estimador $\hat{V}(\hat{\beta}_j)$ é consistente para $V(\hat{\beta}_j)$. Por tanto para contornar este situação usamos a estatística quasi-t, com a seguinte hipótese:
$H0 : \beta_j = \beta_j^0$, contra a hipótese alterativa de $Ha : \beta_j \neq \beta_j^0$. A estatística de prova é dada por:

$$\frac{\hat{\beta}_j - \beta_j^0}{\sqrt{\hat{V}(\hat{\beta}_j)}} \xrightarrow{d} N(0,1) \tag{15}$$

Onde: $\hat{V}(\hat{\beta}_j) = (X'X)^{-1}X'\hat{\Phi}_0X(X'X)^{-1}$

Um problema com este estimador é que tende a ser bastante viesado quando n não é grande, especialmente quando há presença de pontos de alavanca. Tende a subestimar as variâncias verdadeiras, e por tanto o teste torna-se liberal (não conservador), pois o tamanho real do teste tende a ser maior do que a probabilidade do erro tipo I, para corrigir esta deficiência surge o estimador HC1.

40.4 Estimador HC1 de Hinkler

Podemos construir consistentemente a matriz de covariâncias dos parâmetros estimados da seguinte forma:

$$\widehat{Cov}(\hat{\beta}) = \hat{\Psi}_1 = (X'X)^{-1}X'\hat{\Phi}_1X(X'X)^{-1} \tag{16}$$

Onde

$$\hat{\Phi}_1 = \frac{n}{n-p}diag(\hat{\epsilon}_1^2, \hat{\epsilon}_1^2, \cdots, \hat{\epsilon}_n^2) \tag{17}$$

este estimador no limite ($n \to \infty$) é igual ao HCO.

40.5 Estimador HC2 de Horn-Duncan

Este estimador é dado por:

$$\widehat{Cov}(\hat{\beta}) = \hat{\Psi}_2 = (X'X)^{-1}X'\hat{\Phi}_2X(X'X)^{-1} \tag{18}$$

Onde

$$\hat{\Phi}_2 = diag(\frac{\hat{\epsilon}_1^2}{1-h_1}, \frac{\hat{\epsilon}_1^2}{1-h_2}, \cdots, \frac{\hat{\epsilon}_n^2}{1-h_n}) \tag{19}$$

O estimador HC2 é não viesado sobre homoscedasticidade.

40.6 Estimador HC3 de Davidson-MacKinnon

Este estimador é uma aproximação de "Jackknife", dado por:

$$\widehat{Cov}(\hat{\beta}) = \hat{\Psi}_3 = (X'X)^{-1}X'\hat{\Phi}_3X(X'X)^{-1} \tag{20}$$

Onde

$$\hat{\Phi}_3 = diag(\frac{\hat{\epsilon}_1^2}{(1-h_1)^2}, \frac{\hat{\epsilon}_1^2}{(1-h_2)^2}, \cdots, \frac{\hat{\epsilon}_n^2}{(1-h_n)^2}) \tag{21}$$

Os testes que usam HC3, tendem a funcionar melhor que os que usam HC0.

40.7 Estimador HC4 de Cribari Neto

Este estimador não considera uma constante quadrática para a potencia dos valores influentes e sim um potencia nos valores influentes, as quais variam de observação para observação, dado por:

$$\widehat{Cov}(\hat{\beta}) = \hat{\Psi}_4 = (X'X)^{-1}X'\hat{\Phi}_4X(X'X)^{-1} \tag{22}$$

Onde

$$\hat{\Phi}_4 = diag(\frac{\hat{\epsilon}_1^2}{(1-h_1)^{\delta_1}}, \frac{\hat{\epsilon}_1^2}{(1-h_2)^{\delta_2}}, \cdots, \frac{\hat{\epsilon}_n^2}{(1-h_n)^{\delta_n}}) \tag{23}$$

$$\delta_n = min(4, \frac{h_n}{\bar{h}}) = min(4, \frac{nh_n}{p}) \tag{24}$$

e

$$\bar{h} = \frac{\sum h_n}{n} = p/n \tag{25}$$

40.8 Estimador HC5 de Cribari-Souza

Uma alteração do HC4 é dado por:

$$\widehat{Cov}(\hat{\beta}) = \hat{\Psi}_5 = (X'X)^{-1}X'\hat{\Phi}_5 X(X'X)^{-1} \tag{26}$$

Onde

$$\hat{\Phi}_5 = diag(\frac{\hat{\epsilon}_1^2}{(1-h_1)^{\delta_1}}, \frac{\hat{\epsilon}_1^2}{(1-h_2)^{\delta_2}}, \cdots, \frac{\hat{\epsilon}_n^2}{(1-h_n)^{\delta_n}}) \tag{27}$$

$$\delta_n = \min(\frac{nh_n}{p}, \max(4, \frac{n\kappa h_{max}}{p})) \tag{28}$$

e h_{max} : alavancagem máxima, κ : constante entre 0 e 1 (Usa-se $\kappa = 0.7$ nas simulações). Para exemplificar estes estimadores consideremos o consumo de água em uma residência.

```
library(lmtest)
library(sandwich)
coeftest(m1,vcov=vcovHC(m1))
vcov=vcovHC(m1)
vcov
               consumo    diasconsumo Construindo
consumo      0.06220938  -0.024252971 -0.17343741
diasconsumo -0.02425297   0.009637151  0.06351919
Construindo -0.17343741   0.063519195  1.94060840
# Para incorporar a matriz de covariância usamos
t(sapply(c("const","HC0","HC1","HC2","HC3","HC4","HC5"),
function(x) sqrt(diag(vcovHC(m1,type=x)))))

consumo diasconsumo Construindo
const 0.1553605  0.05995221    1.228001
HC0   0.2178130  0.08653240    1.289346
HC1   0.2230204  0.08860119    1.320171
```

```
HC2   0.2328432  0.09206926    1.340030
HC3   0.2494181  0.09816899    1.393057
HC4   0.2764422  0.10747053    1.380511
HC5   0.2437678  0.09577545    1.333420
```

41 Multicolinearidade

Duas ou mais variáveis são colineares se possuem relação exata, ou seja, se um dos vetores é uma combinação linear dos outros (como se fossem retas paralelas). A correlação exata raramente ocorre e maiores detalhes deste fato pode ser encontrado em Gujarati (2000), porem correlações fortes (correlação r acima de $|0.8|$) já são perigosas, assim como regressores auxiliares de cada regressor sobre os demais for alto.
Apenas a correlação entre variáveis independentes é problemática. A relação forte de cada uma das variáveis independentes x_i com a variável dependente y_i é desejável. Quando existem mais de duas variáveis independentes relacionadas fortemente fala-se em *multicolinearidade*.

A multicolinearidade afeta os coeficientes da equação de regressão de forma significativa, os denominados $\hat{\beta}_i$, alterando o valor e até o sinal em relação ao que ocorreria se não houvesse este problema. Na presença de correlação alta, os coeficientes de regressão estimados tendem a ser imprecisos e as estimativas dos coeficientes variam bastante de uma amostra para outra. Quando há colinearidade, as estimativas dos mínimos quadrados ainda são não-tendenciosas e eficientes, porém o erro padrão dos coeficientes tende a ser grande, e o teste baseado na estatística t de Student calculará significância menor que a real. Os coeficientes não são confiáveis, impossibilitando o uso dos modelos para análise do mercado ou previsão de valores. Outro efeito da colinearidade é que torna-se difícil obter interpretações sobre o efeito isolado de cada uma das variáveis. Nos casos de correlação alta, uma das alternativas é a remoção da variável mais afetada. Isso pode introduzir tendências, sendo mais adequado substituir esta variável por outra menos colinear mas que tenha aproximadamente a mesma construção teórica, ou por uma variável que seja a combinação das colineares. Nem sempre a remoção ou substituição da variável afetada é uma boa solução. Quando se trabalha com predição de valores e existem indicações de que a colinearidade encontrada continuará no futuro e o modelo poderá apresentar bons resultados.

Podemos observar a existência de colinearidade através da matriz de correlação das variáveis independentes.

A sequencia a seguir é identificar a natureza da multicolinearidade, se de fato é um problema e quando, as suas consequências práticas.

41.1 Natureza

A natureza da multicolinearidade ou existência de uma perfeita ou exata relação linear entre algumas ou todas as variáveis independentes de um modelo de regressão, em geral esta suposição não é realista. É comum a relação quase perfeita entre as variáveis independentes (as variáveis são inter correlacionadas). A multicolinearidade é uma questão de grau e não de natureza. A distinção significativa não esta entre a presença ou ausência de multicolinearidade mas entre seus vários graus. Como a multicolinearidade se refere à condição de variáveis independentes que se presume não estocásticas, é uma característica da amostra, e não da população.
Uma das consequências da multicolinearidade perfeita é que não pode ser estimado os parâmetro, invalidando o MELNT.
Se é quase-perfeita os estimadores de M.Q.O são não tendenciosos, eficientes e consistentes MELNT. Alem disso os teste de hipóteses não são afetados.
Se o interesse é a previsão a multicolinearidade não representa um problema, os contrario diminuem o erro dos resíduos, e as previsões serão não tendenciosas, também os intervalos de confiança para previsão permanecem válidos.

41.2 Identificação

1. O R^2 alto, porem $t - ratio$ baixos, implica teste F altamente significativo.

2. A correlação (a pares) das variáveis independentes muito alto.

3. Estimativa dos parâmetros sensível a especificação ou correlação a pares não é muito alta mas existe uma relação quase perfeita entre mais do que duas variáveis.

41.3 Corrigindo

1. Se objetivo é previsão ignorar multicolinearidade.

2. Eliminação de variáveis, geral para o particular.

3. Reformulando o modelo, usando razão entre variáveis ou transformando as variáveis para taxa.

4. Usando informação externa, assumir que um dos parâmetros foi estimado fora do modelo e impor como verdadeiro.

5. Usar estimativa de cross-section da elasticidade renda para obter a elasticidade preço numa especificação de séries temporais.

6. Aumentar o tamanho da amostra.

41.4 Testes de Multicolinearidade

1. Padronização dos regressores: este teste para detectar multicolinearidade é usando o determinante de $X'_m X_m$, onde $X_m = [x_2^m, x_3^m, \cdots, x_p^m]$, padronize os regressores da seguinte forma:

$$X_{nj}^m = \frac{x_{nj} - \bar{x}_j}{\sqrt{\sum (x_{nj} - \bar{x}_j)^2}}, \qquad \bar{x}_j = \frac{1}{n} \sum x_{nj} \tag{29}$$

   onde, $j = 2, 3, \cdots, p$. A dimensão de $X_m = n \times (p-1)$. Calcule o determinante de $X'_m X_m$, onde se $|X'_m X_m| = 0$, há multicolinearidade exata, e se $|X'_m X_m| = 1$, as colunas de X são ortogonais. Quanto mais próximo de 0 mais severo o problema.

2. Fatores de inflação de variância (vif): observando os elementos diagonais de $(X'_m X_m)^{-1}$, estes elementos são chamados de fatores de inflação de variância. Se um **vif** for maoir que 5, isto é indicativo de multicolinearidade, no R usamos a função **vif()** do pacote **car**

3. teste de Farrar e Glauber:
   Considere as seguintes hipóteses, $H0$: Há ausência de multicolinearidade, versus a hipótese Ha: caso contrário, com a estatística de teste qui-quadrado definido pela seguinte fórmula:

$$X^2 = -[n - 1 - 1/6 * (2p + 5)] \ln det(mcor(X)) \tag{30}$$

   onde $mcor(X)$, e a matriz de correlação das variáveis independentes.

41.5 Exemplo

Considere os dados de consumo de água de uma residência durante o período dos meses de 12/2005 a 04/2011, fornecidos em formato txt (consumoagua2.txt) salvo na pasta padrão do R. Usando a função **vif()**

```
> ca=read.table("consumoagua2.txt", header=TRUE)
> attach(ca)
> m1=lm(valor~consumo+diasconsumo+Construindo-1)
> vif(m1)
    consumo diasconsumo Construindo
  22.100045   18.091433    2.102514
```

Podemos observar que há presença de multicolinearidade ($vif > 5$), para contornar eliminaremos a variável consumo

```
>m6=lm(valor~diasconsumo+Construindo)
> vif(m6)
diasconsumo Construindo
    1.01605     1.01605
```

Podemos observar que com o modelo m6 não há presença de multicolinearidade ($vif < 5$)
Teste da presença de alta multicolinearidade de Farrar e Glauber. Inicialmente determinamos a matriz de correlações dos regressores.

```
# mm matriz de variáveis independentes
> mm=cbind(consumo,diasconsumo,Construindo)
> cor(mm)

              consumo diasconsumo Construindo
consumo     1.0000000   0.2496357   0.5968966
diasconsumo 0.2496357   1.0000000   0.1256844
Construindo 0.5968966   0.1256844   1.0000000
```

Posteriormente aplicamos o teste:

```
> x2=-(65-1-(1/(6*11)))*log(det(cor(mm)))
> x2
[1] 32.36007
> xt=qchisq(0.95,3)
> xt
[1] 7.814728
```

O valor observado x2= 32.36, é maior que o valor tabelado xt= 7.815, concluem-se que há presença forte de multicolinearidade. Em consequência a variável consumo sai da analise, o modelo que elimina a multicolinearidade é:

```
> m5=lm(valor~diasconsumo+Construindo-1)
> mm2=cbind(diasconsumo,Construindo)
> cor(mm2)
            diasconsumo Construindo
diasconsumo   1.0000000   0.1256844
Construindo   0.1256844   1.0000000
> x2=-(65-1-(1/(6*9)))*log(det(cor(mm2)))
> x2
[1] 1.018756
> xt=qchisq(0.95,1)
> xt
[1] 3.841459
```

Como o valor observado x2= 1.019, é menor que o valor tabelado xt= $3,84$, concluem-se que a presença de multicolinearidade não é significativa. Em consequência as duas variáveis dias consumo e Construindo, compõem as variáveis independentes. Cabe lembrar que nosso interesse é melhorar a significância das estimativas do parâmetros. Quando a presença de multicolinearidade é quase exata, podemos contornar-la com a obtenção de mas dados, uma outra alternativa é usando a regressão ridge.

41.6 Regressão ridge

Os estimadores de mínimos quadrados são não viciados na presença de multicolinearidade e as previsões podem ser melhoradas usando um método chamado de regressão ridge. A ideia é que introduzindo um viés nas estimativas dos parâmetros do modelo podemos encontrar uma matriz de variâncias e covariâncias menor que os estimados por MQO. A regressão ridge é uma família de estimadores dada por:

$$\hat{\beta}_\lambda = (X'X + \lambda I)^{-1}X'y, \qquad \lambda > 0 \tag{31}$$

Note que se $\lambda = 0$, obtemos o EMQO, se $\lambda > 0$, e não estocástico, então temos:

$$E(\hat{\beta}_\lambda) = E((X'X + \lambda I)^{-1}X'y) = (X'X)^{-1}X'X\beta \neq \beta \tag{32}$$

Por tanto o estimaor ridge é viesado. Para determinar o tamanho do vies calculamos:

$$E(\hat{\beta}_\lambda) = \beta - [\lambda^{-1}(X'X') + I]^{-1}\beta \tag{33}$$

Por tanto, $VIES(\hat{\beta}_\lambda = -\lambda[\lambda I + (X'X)]^{-1}\beta$. Para o cálculo de este estimador usamos as propriedade de descomposição de matrizes. Como estamos introduzindo um viés nas estimativas dos parâmetros, podemos Encontrar, que $V(\hat{\beta}_j) > V(\hat{\beta}_j(\lambda))$, (uma medida para confirmar o ganho com este estimador é usar o mape, ver detalhes no capítulo de séries temporais). O estimador ridge é uma transformação linear do EMQO, para corrigir a multicolinearidade, onde λ é um parâmetro de encolhimento. Para exemplificar

```
>library(MASS)
> rr=lm.ridge(m1,lambda = seq(0,0.1,0.01))
> rr
      consumo diasconsumo Construindo
0.00 2.758109  -0.3317662    3.687592
0.01 2.746590  -0.3274165    3.738761
0.02 2.735211  -0.3231202    3.789309
0.03 2.723969  -0.3188762    3.839249
0.04 2.712861  -0.3146837    3.888592
0.05 2.701886  -0.3105417    3.937347
0.06 2.691040  -0.3064493    3.985525
0.07 2.680321  -0.3024057    4.033137
0.08 2.669727  -0.2984100    4.080192
0.09 2.659257  -0.2944613    4.126701
0.10 2.648907  -0.2905588    4.172672

> m1
> plot(lm.ridge(m1,lambda = seq(0,0.1,0.01)))
> rr1=lm.ridge(m1,lambda =0.1)
>rr1
    consumo diasconsumo Construindo
  2.6489068  -0.2905588   4.1726722
> rrc=2.6489068*consumo-0.2905588*diasconsumo+4.1726722*Construindo
> rr2=lm.ridge(m1,lambda =0.2)
    consumo diasconsumo Construindo
  2.5516146  -0.2539091   4.6047813
>rr2
> rr2c=2.5516146*consumo-0.2539091*diasconsumo+4.6047813*Construindo
> f1=fitted(m1)
> f=cbind(f1,rrc,rr2c)
> mape=abs((valor-f)/f)*100
```

```
> apply(mape,2,mean)
      f1      rrc     rr2c
13.47194 12.71836 12.24492
```

Para a escolha do valor ideal de λ, podemos usar o estimador proposto por Hoel, Kennard e Baldwin:

$$\hat{\lambda} = \frac{(\lambda - 1)\hat{\sigma}_m^2}{\hat{\beta}'\hat{\beta}} \tag{34}$$

onde:

$$\hat{\sigma}_m^2 = \frac{(y_m - X_m\hat{\beta}_m)'(y_m - X_m\hat{\beta}_m)}{n - p} \tag{35}$$

$$\hat{\beta}_m = (X_m'X_m)^{-1}X_m'y_m \tag{36}$$

42 Autocorrelação

O nosso objetivo é tenta dar respostas a natureza da autocorrelação, as consequências teóricas e praticas, se a autocorrelação esta relacionada com as perturbações não observáveis, como saber se há autocorrelação numa dada situação, e como corrigir o problema da autocorrelação. Tanto na presença de heterocedasticidade como de autocorrelação, os estimadores usuais de MQO, embora não viesados, consistentes e assintoticamente normal, já não possuem variância mínima (não é eficiente) entre todos os estimadores lineares não tendenciosos. O seja não são MELNT.

42.1 Natureza da autocorrelação

Um dos pressupostos do modelo de regressão linear clássico é:

$$E(\epsilon_i, \epsilon_j) = 0 \qquad \forall i \neq j \tag{37}$$

Como se pressupõe que a média de $E(\epsilon_i) = 0 = E(\epsilon_j)$, podemos escrever que a covariância, $cov(\epsilon_i, \epsilon_j) = 0$. Esta característica das perturbações de regressão é conhecida como não auto-regressão. Porem caso haja dependência nas perturbações, temos autocorrelação. Simbolicamente:

$$E(\epsilon_i, \epsilon_j) \neq 0 \tag{38}$$

Dizemos que os erros são esféricos quando $cov(\epsilon) = E(\epsilon\epsilon') = \sigma^2 I_n$, esta esfericidade pode ser violada por ter elementos na diagonal não constantes (heterocedasticidade) ou por elementos fora da diagonal diferente de zero, neste último caso dizemos que á presença de autocorrelação. Um exemplo dado em Gujarati (2000), considera que si estivermos lidando com dados de corte transversal com regressão de consumo familiar sobre a renda familiar, não vamos esperar que o efeito no consumo de uma família, decorrente de um aumento em sua renda, vai afetar o consumo de uma outra família. No entanto se o acompanhamento de consumo familiar sobre a renda familiar de uma mesma família ao longo de uma série temporal (corte longitudinal, espera-se que no aumento em sua renda aumente também seu consumo).

42.1.1 Terminologia

Autocorrelação: Correlação defasada de uma dada série consigo mesma, defasada em um número de unidades de tempo.

Correlação serial: Correlação defasada entre duas séries diferentes. Assim uma correlação entre duas séries defasada em um período de tempo do tipo: $u_1, u_2, ..., u_{10}$, e $u_2, u_3, \ldots, u_{11}$, é um exemplo de autocorrelação, enquanto uma correlação entre séries temporais do tipo: $u_1, u_2, ..., u_{10}$, e $v_2, v_3, ..., v_{11}$, em que u e v são duas séries temporais diferentes.
O presuposto de que no tempo as relações estimadas a partir de observações, envolvem perturbações auto-regressivas é tão comum e por tanto:

$$E(\epsilon_t, \epsilon_{t-s}) \neq 0 \qquad \forall t > s \tag{39}$$

Esta expressão implica que a perturbação que ocorre no tempo t relaciona-se à perturbação que ocorre no tempo $t - s$.
As consequências da auto-regressão na estimação pressupõe que:

$$E(\epsilon_t, \epsilon_{t-s}) = cov(\epsilon_t, \epsilon_{t-s}) = \rho^s \sigma_\epsilon^2 \tag{40}$$

42.1.2 Geração das Perturbações

A questão que tem que ser respondida agora refere-se à maneira pela qual as perturbações são geradas de modo que elas se relacionam uma à outra. O modelo mais usado é geradas no seguinte esquema:

$$\epsilon_t = \rho\epsilon_{t-1} + \nu_t \qquad t = 1, 2, \ldots \tag{41}$$

Onde $\nu_t \approx N(0, \sigma_\nu^2)$ e pressupõe-se independente de ϵ_t. Uma relação como na equação acima é conhecida como esquema auto-regressivo de primeira

ordem. Implica que cada perturbação corrente seja igual a uma *porção* da perturbação precedente mais um efeito aleatório. Através de uma substituição, temos:
$\epsilon_t = \rho\epsilon_{t-1} + \nu_t$
$\epsilon_t = \rho(\rho\epsilon_{t-2} + \nu_{t-1}) + \nu_t$
$\dots$
$\epsilon_t = \rho^t\epsilon_0 + \rho^{t-1}\nu_1 + \ldots + \rho\nu_{t-1} + \nu_t$
Assumindo que

$$\epsilon_0 \approx N(0, \frac{\sigma_\nu^2}{1-\rho^2}) \tag{42}$$

temos

$$var(\epsilon_t) = \frac{\sigma_\nu^2}{1-\rho^2} \tag{43}$$

A correlação é por tanto:

$$corr(\epsilon_t, \epsilon_{t-1}) = \frac{cov(\epsilon_t, \epsilon_{t-1})}{\sqrt{var(\epsilon_t)var(\epsilon_{t-1})}} = \rho \tag{44}$$

ρ é a correlação entre erros sucessivos, por tanto

$$cov(\epsilon_t, \epsilon_{t-k}) = \rho^k\sigma_\nu^2 \tag{45}$$

A matriz de covariâncias dos resíduos é

$$\Phi = E(\epsilon\epsilon') = \frac{\sigma_\nu^2}{1-\rho^2}\begin{bmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{n-1} \\ \rho & 1 & \rho & \cdots & \rho^{n-2} \\ \vdots & & & & \vdots \\ \rho^{n-1} & \rho^{n-2} & \rho^{n-3} & \cdots & 1 \end{bmatrix}$$

e $\Phi = \sigma_\epsilon^2\Omega$ as correlações naturalmente decaem geometricamente a medida que os erros se afastam.

42.2 Diagnostico da Autocorrelação

42.2.1 Teste de Durbin Watson

Este teste é o mais comum na literatura, e serve para diagnosticar em que grau a autocorrelação de primeira ordem esta presente. As hipóteses são as seguintes:

1. H0: Não há presença de autocorrelação de primeira ordem ($d = 2$).
   Ha: Há presença de autocorrelação.

2. nível de significância: $\alpha = 0,05$

3. Estatística de Prova:

$$d = \frac{\sum_{t=2}^{n}(\hat{\epsilon}_t - \hat{\epsilon}_{t-1})^2}{\sum_{t=1}^{n}\hat{\epsilon}_t^2} \tag{46}$$

4. Regra de Decisão:
   Se $d \approx 2$, há ausência de autocorrelação;
   Se $d \approx 0$, há autocorrelação positiva;
   Se $d \approx 4$, há autocorrelação negativa;

Onde $\approx$ significa próximo do valor. Ver maiores detalhes em Gujarati(2000).

42.2.2 Exemplo 1

Considere a cotação diária da Bovespa, para determinar a presença de autocorrelação, temos:

```
> md=lm(cotacao~dias)
> dwtest(md)

        Durbin-Watson test

data:  md
DW = 0.008, p-value < 2.2e-16
alternative hypothesis: true autocorrelation is greater than 0
```

Conclusão, como o valor Durbin-Watson é próximo de zero e o p-valor menor que 0.05, concluímos que não há evidencias de aceitar H0, em favor de presença de autocorrelação positiva de primeira ordem. Podemos visualizar a autocorrelação dos resíduos com o seguinte comando:

```
acf(residuals(md),lag=200, main="Resíduos do Modelo md \n cotação~dias")
```

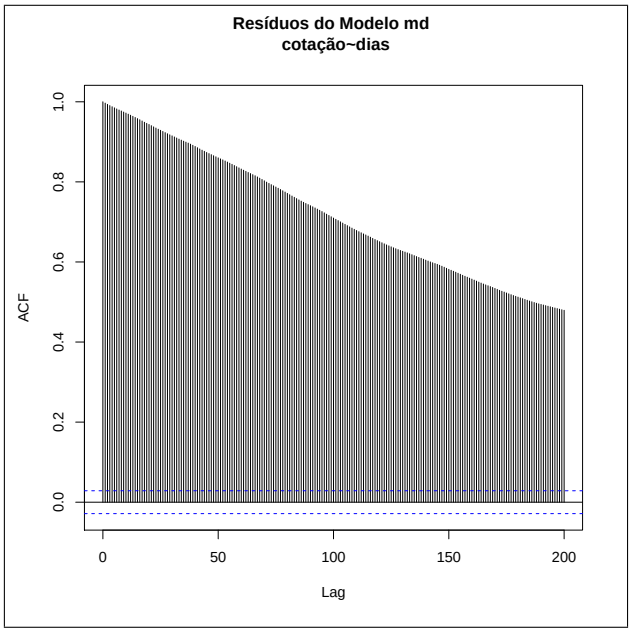


Figura 21.b Função de autocorrelação.

42.2.3 Exemplo 2

Considere o consumo de água em uma residência, aplicando o teste de Durbin, temos:

```
> m1 = lm(valor ~ consumo + diasconsumo + Construindo - 1)
> dwtest(m1)
        Durbin-Watson test

data:  m1
DW = 1.4882, p-value = 0.01273
alternative hypothesis: true autocorrelation is greater than 0
```

Conclusão, como o valor Durbin-Watson é próximo de zero e o p-valor menor que 0.05, concluímos que não há evidencias de aceitar H0, em favor de presença de autocorrelação positiva de primeira ordem.

42.2.4 Teste assintótico

Quando n é grande, podemos construir as hipóteses
$H0 : \rho = 0$, versus $Ha : \rho \neq 0$, onde:

$$\hat{\rho} \approx N(\rho, \frac{1-\rho^2}{n}) \tag{47}$$

Sobre $H0, \hat{\rho} \approx N(0, 1/n)$, com estatística de teste $\sqrt{n}\hat{\rho} \approx N(0, 1)$.
A regra de decisão: Rejeita-se $H0$ se $\sqrt{n}|\hat{\rho}| > z_{\alpha/2}$, onde, $z_{\alpha/2}$, é o quantil $1 - \alpha/2$ da distribuição normal padrão.
Para facilitar o entendimento de uma série temporal consideraremos daqui em diante $n = t$

42.2.5 Teste de Durbin

Quando y_{t-1} é regressor (é estocástico), viola a pressuposição básica do modelo. Para regressar y_{t-1} realizamos:

$$y_t = \gamma y_{t-1} + X_t'\beta + \epsilon_t, \qquad t = 1, 2, \cdots, T \tag{48}$$

Aqui $\epsilon_t = \rho\epsilon_{t-1} + \nu_t$, cujas hipóteses são:
$H0 : \rho = 0$, versus $Ha : \rho \neq 0$.
A estatística de teste é:

$$h = \hat{\rho}\sqrt{\frac{T}{1 - T(\hat{V}(\hat{\gamma}))}} \tag{49}$$

Assumindo que $T\hat{V}(\hat{\gamma}) < 1$, sobre $H0$, $h \xrightarrow{d} N(0, 1)$. Quando T não é muito grande o teste não funciona muito bem.

42.3 Eliminação da Autocorrelação

Para eliminar a autocorrelação, onde ϵ_t é gerado por um processo autorregressivo de primeira ordem, podemos construir o seguinte modelo:

$$y_t - \rho y_{t-1} = \beta_0 * (1 - \rho) + \beta_1(X_t - \rho X_{t-1}) + \nu_t \tag{50}$$

No exemplo da cotação da Bovespa, para a eliminação de autocorrelação:

```
> diast=dias[-1]
> diast1=dias[2703]
> cotacaot=cotacao[-1]
```

```
> cotacaot1=cotacao[-2703]
> diast1=dias[-2703]
> e=residuals(md)
> et=e[-1]
> et1=e[-2703]
> core=cor(et,et1)
> yt=cotacaot-core*cotacaot1
> xt=diast-core*diast1
> mea=lm(yt~xt-1)
> mea
> dwtest(mea)

        Durbin-Watson test

data:  mea
DW = 2.0505, p-value = 0.9054
alternative hypothesis: true autocorrelation is greater than 0
```

No exemplo acima com a transformação do modelo podemos concluir que não há evidencias de presença de autocorrelação.

Para o exemplo de consumo de água em uma residência, eliminamos:

```
> e=residuals(m1)
> e1=cor(e[-1],e[-65])
> wt=valor[-1]-e1*valor[-65]
> z1=consumo[-1]-e1*consumo[-65]
> z2=diasconsumo[-1]-e1*diasconsumo[-65]
> z3=Construindo[-1]-e1*Construindo[-65]
> msa=lm(wt~z1+z2+z3-1)
> dwtest(msa)

        Durbin-Watson test
data:  msa
DW = 2.1924, p-value = 0.7585
alternative hypothesis: true autocorrelation is greater than 0
```

O resultado do teste de Durbin-Watson acusa a ausência de autocorrelação, ou em outras palavras não há evidencias para rejeitar $H0$.

Uma alternativa simples, que em geral elimina a autocorrelação, é usar as primeiras diferenças, tanto nas variáveis endógenas quanto as exógenas.

43 Exercícios

1. Na sua opinião você prefere modelos com muitas variáveis regressoras (independentes), ou não, justifique sua resposta.
2. O que entende por parsimonia na construção de um modelo de regressão?
3. Pesquise a visão de Ballentine sobre a multicolinearidade.
4. Ao final, a multicolinearidade é um problema ou uma solução? justifique suas afirmações.
5. Determine as estimativas dos parâmetros na presença da multicolinearidade Perfeita. Ver Gujarati, 2000.
6. Determine as estimativas dos parâmetros na presença de multicolinearidade quasi-perfeita.
7. Construa um modelo polinomial (ou de potência) para os dados de consumo de água de uma residência, onde o valor pago esta em função do consumo num modelo cúbico. Diagnostique a presença ou não de multicolinearidade.
8. No exemplo anterior faça uma regressão passo passo para incluir qual o modelo de potência com o coeficiente de correlação quadrático alto sem presença de multicolinearidade.
9. Que acontece com a matriz de variâncias e covariâncias $cov(\hat{\beta})$ na presença de multicolinearidade perfeita.
10. Com os dados de consumo de água, identifique um modelo que supere todos os pressupostos básicos e comente.
11. Use os experimentos de Monte Carlo e construa modelos de regressão com 5 variáveis independentes com presença de multicolinearidade. O que acontece com a distribuição das estimativas dos parâmetros quando n é grande (maior que 1000000)
12. use a fução **raintest()** do pacote **lmtest**, para identificar a forma funcional do modelo m1.
13. use a fução **Box.test()**, para identificar a autocorrelação dos resíduos do modelo m1.

44 Regressão não Linear

Nos modelos lineares, a estimação dos parâmetros, cai no problema de resolver um sistema de equações lineares com relação aos coeficientes de regressão desconhecidos. Existe uma solução única e, portanto, obtemos uma forma analítica de estimação dos parâmetros. Esta forma é a mesma para qualquer modelo e qualquer conjunto de dados. Além disso, como os coeficientes são combinações lineares das observações, pela teoria estatística, demonstra-se que a distribuição amostral dos coeficientes estimados de regressão segue uma distribuição t ou normal, assim, podemos realizar testes de hipóteses, calcular intervalos de confiança para esses coeficientes populacionais de regressão. Quando falamos de modelos de regressão não linear estamos tratando de modelos que são não lineares nos parâmetros. Existem muitas situações nas quais não é desejável, ou mesmo possível, descrever um fenômeno através de um modelo de regressão linear. Ao invés de se fazer uma descrição puramente empírica do fenômeno em estudo, pode-se, a partir de suposições importantes sobre o problema (frequentemente dadas através de uma ou mais equações diferenciais), trabalhar no sentido de obter uma relação teórica entre as variáveis observáveis de interesse. O problema, diferentemente do caso linear, é que os parâmetros entram na equação de forma não linear, assim, nós não podemos simplesmente aplicar fórmulas para estimar os parâmetros do modelo. Em muitas situações, necessitam-se menos parâmetros nos modelos não lineares do que nos lineares, isto simplifica e facilita a interpretação.

Os modelos não lineares podem ser escritos como:

$$Y_i = f(X_i, \beta) + \epsilon_i \tag{51}$$

onde $f(X_i, \beta)$ é uma função não linear; os erros, ϵ_i, tem média zero, variância constante, e não são correlacionados. Assume-se que os erros seguem distribuição normal, são independentes e têm variância constante. β é o vetor de parâmetros do modelo. Para iniciar apresentamos três modelos não lineares:

45 Modelo exponencial

$$Y_i = \beta_0 \exp(\beta_1 X_i) + \epsilon_i \tag{52}$$

onde β_0 e β_1 são os parâmetros do modelo; X_i são constantes conhecidas como variável preditora e ϵ_i são os termos do erro, independentes, com distribuição normal de média 0 (zero) e variância σ^2. Diferenciando f com

respeito aos parâmetros β_0 e β_1 obtemos:

$$\frac{\partial f}{\partial \beta_0} = \exp(\beta_1 X_i) \tag{53}$$

$$\frac{\partial f}{\partial \beta_1} = \beta_0 X_i \exp(\beta_1 X_i) \tag{54}$$

45.1 Medida de ajuste

Em modelos de regressão há varias medidas de qualidade de ajuste dos dados, tais como o coeficiente de correlação, coeficiente de determinação, critérios de informação (AIC, BIC), porem neste livro usaremos uma medida universal para ajustar os dados próximos dos seus valores reais, esta medida chama-se de erro absoluto médio percentual, e na literatura é definido por **mape**, o calculo é dada por:

$$mape = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i/y_i| * 100 \tag{55}$$

45.2 Exemplo

Considere as projeções a cada década do tamanho da população Brasileira (ver no apêndice a chamada dos dados), desde 1870 até 2010. Um modelo de regressão não linear da família exponencial é

```
> mpop.ex= glm(popbrasil ~ano, family=Gamma(link=log))
>summary(mpop.ex)
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.952e+01  6.904e-01  -57.24   <2e-16 ***
ano          2.233e-02  3.558e-04   62.76   <2e-16 ***
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1  1

(Dispersion parameter for Gamma family taken to be 0.003544175)
    Null deviance: 12.735063  on 14  degrees of freedom
Residual deviance:  0.046475  on 13  degrees of freedom
AIC: 75.832
Number of Fisher Scoring iterations: 3
> mape1=(sum(abs((popbrasil-fitted(mpop.ex))/popbrasil))*100)/15
> mape1
[1] 4.366599
```

```
> plot(ano,popbrasil)
> lines(ano,fitted(mpop.ex),col=8)
```

Podemos afirmar que o modelo é adequado pelo teste F, os parâmetros são significativos pelo teste t para cada parâmetro, e as previsoes do modelo tem um bom ajuste, pois o erro de previsão em percentual (mape)é de 4.33

46 Modelo logístico

$$Y_i = \frac{\beta_0}{1 + \beta_1 \exp(\beta_2 X_i)} + \epsilon_i \tag{56}$$

da mesma forma que no modelo anterior, diferenciando obtemos:

$$\frac{\partial f}{\partial \beta_0} = \frac{1}{1 + \beta_1 \exp(\beta_2 X_i)} \tag{57}$$

$$\frac{\partial f}{\partial \beta_1} = \frac{\exp(\beta_2 X_i)}{(1 + \beta_1 \exp(\beta_2 X_i))^2} \tag{58}$$

$$\frac{\partial f}{\partial \beta_2} = \frac{\beta_1 \exp(\beta_2 X_i X_i)}{(1 + \beta_1 \exp(\beta_2 X_i))^2} \tag{59}$$

Para ajustar um modelo logístico para as projeções do tamanho da população do Brasil (décadas), precisamos de valores iniciais para os parâmetros, isto é rotineiro quando usamos a estimação de modelos não lineares, porem a eleição destes valores iniciais devem seguir alguns critérios e ser razoavelmente próximos dos seus verdadeiros valores, assim calcularemos os valores iniciais para a estimativa dos parâmetros (Os $\beta_i's$). Para modelo crescentes, utilizamos de modo geral que $\beta_1 ini$ é maior que a última projeção da população Brasileira, assim sera maior que 190, por tanto um valor inicial será $\beta_1 = 220$. Para estimar o valor inicial de $\beta_2 ini$, usamos a primeira projeção estimada em milhões de habitantes do Brasil (9.533659) então substituímos em nosso modelo inicial da seguinte forma:

$$9.533659 = \frac{220}{1 + \exp^{\beta_2 + \beta_3 0}} \tag{60}$$

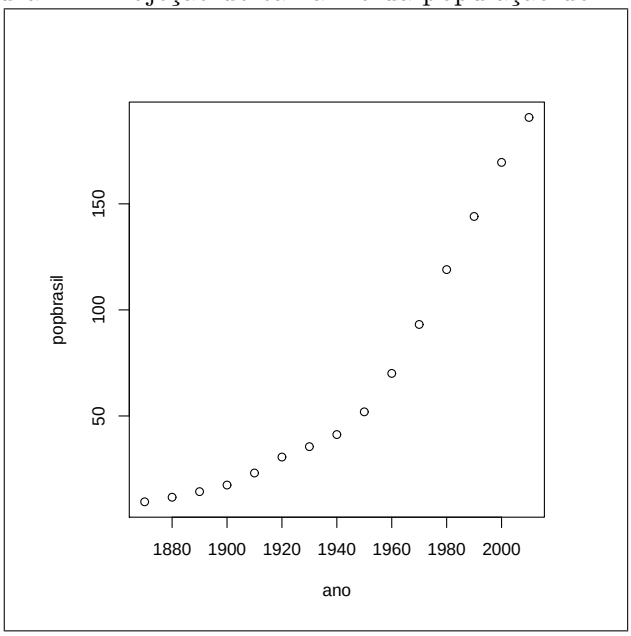
Resolvendo esta equação temos que $\beta_2 ini \cong 3.1$. Finalmente o valor inicial de $\beta_3 ini$, pode ser calculado usando a segunda projeção da população, dada a seguir

$$11.689938 = \frac{220}{1 + \exp^{3.1 + \beta_3}} \tag{61}$$

A solução desta equação estima o valor inicial de $\beta_3 ini \cong -0.22$. Com estes valores iniciais, usamos a função **nls()** do pacote **nls2**, da seguinte forma:

```
library(nls2)
tempo= 0:14
pop.b <- nls(popbrasil ~ beta1/(1 + exp(beta2 + beta3*tempo)),
start=list(beta1 = 220, beta2 = 3.1, beta3 = -0.22), trace=T)
```

Figura 22. Projeção do tamanho da população do Brasil



A saída dos resultados é dado na plataforma R.

```
pop.b <- nls(popbrasil ~ beta1/(1 + exp(beta2 + beta3*tempo)),
start=list(beta1 = 220, beta2 = 3.1, beta3 = -0.22), trace=T)
 18689.5 :  220.0000000   3.1000000  -0.2200000
12062.99 :  372.1359626   3.9467894  -0.2394392
229.5459 :  333.6289610   3.9016451  -0.3001283
217.1894 :  380.9831173   4.0275197  -0.2880894
179.4752 :  384.8339780   4.0395886  -0.2901861
179.4711 :  384.0791519   4.0396609  -0.2904526
179.4711 :  384.0546846   4.0397399  -0.2904686
179.4711 :  384.0494772   4.0397417  -0.2904707
```

As estimativas dos parâmetros são dados na saída do comando **summary()**.

```
> summary(pop.b)
```

```
Formula: popbrasil ~ beta1/(1 + exp(beta2 + beta3 * tempo))

Parameters:
       Estimate Std. Error t value Pr(>|t|)
beta1 384.04948   60.14841   6.385 3.48e-05 ***
beta2   4.03974    0.09384  43.049 1.60e-14 ***
beta3  -0.29047    0.01913 -15.180 3.39e-09 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 3.867 on 12 degrees of freedom

Number of iterations to convergence: 7
Achieved convergence tolerance: 2.861e-06
```

Podemos concluir que os três parâmetros são estatisticamente significativos.

47 Modelo alternativo

Um modelo linear generalizado usando **glm()** é proposto para estimar o tamanho da população e comparar-lo com o modelo *logístico* da seção anterior.

```
mpop.b <- glm(popbrasil ~ano, family=gaussian(link=log))
mpop.b
mape1=(sum(abs((popbrasil-fitted(mpop.ex))/popbrasil))*100)/15
mape1
[1] 4.366599
mape2=(sum(abs((popbrasil-fitted(pop.b))/popbrasil))*100)/15
mape2
[1] 9.872469
mape3=(sum(abs((popbrasil-fitted(mpop.b))/popbrasil))*100)/15
mape3
9.89976
```

O modelo alternativo, tem um erro médio absoluto percentual maior que o modelo logístico e exponencial. Para comparar o tamanho da população projetada e as estimativas dos dois modelos (exponencial e alternativo), podemos fazer

```
plot(ano, popbrasil)
lines(ano,fitted(pop.b))
lines(ano,fitted(mpop.ex), lty=4)
legend("topleft", legend=c("Logistico","Exponencial"),lty=1:4)
```

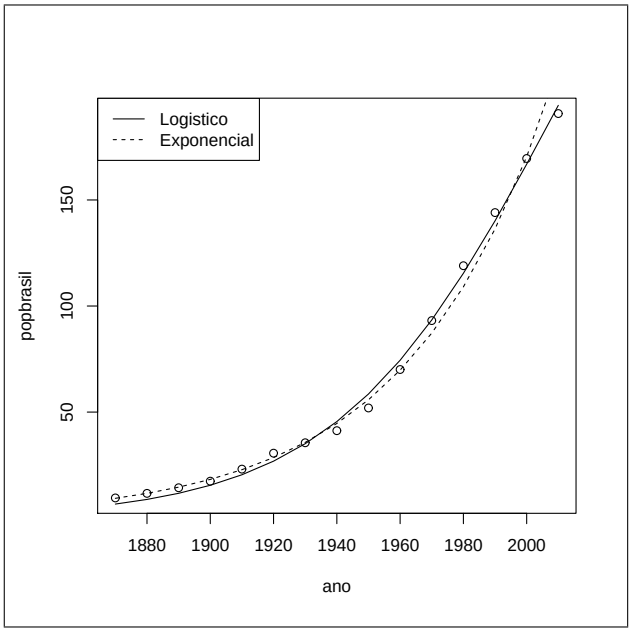


Figura 23. Dois modelos para ajustar o tamanho da população

Embora o valor do *mape* seja menor do modelo exponencial, isto não significa que seja melhor que o modelo proposto, para explicar este fato podemos observar que este modelo tem uma aceleração maior que a taxa de crescimento populacional real e que a taxa de crescimento do modelo proposto. Veja a comparação do ajuste dos modelos no gráfico acima. Observamos que o modelo exponencial ajusta-se bem ao tamanho de população, porem no final da série (última década) acelera o crescimento populacional, o que implica que suas previsões futuras não refletem o verdadeiro crescimento populacional, como podemos observar na previsão com seus respectivos erros de previssãopara 2015 e 2020 para os dois modelos.

```
novo=data.frame(tempo=c(14.5,15))
predict(pop.b, novo, se.fit = TRUE)
[1] 208.5062 222.2381
novo2=data.frame(ano=c(2015,2020))
```

```
p = exp(predict(mpop.ex, novo2, interval="'confidence"'))
p
       1        2
238.1832 266.3136
```

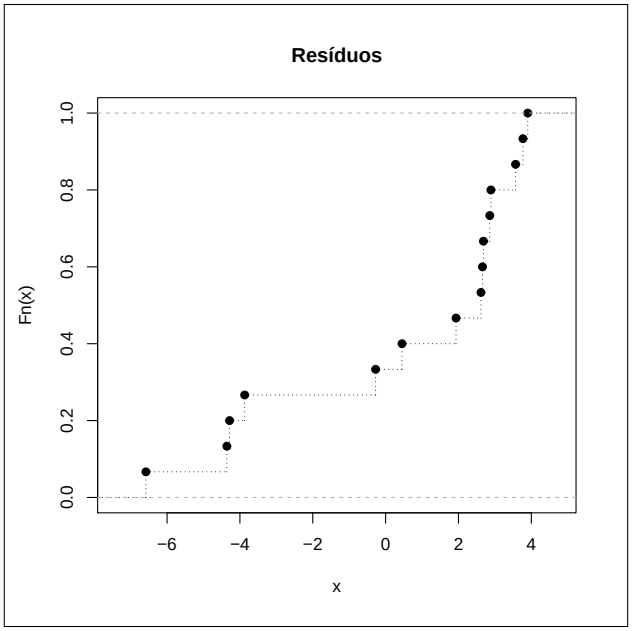


Figura 24. Função acumulada dos resíduos do modelo logistico

Podemos observar que as previsões do tamanho da população para os anos de 2015 e 2020, são conservadoras para o modelo proposto, e são não conservadores para o modelo exponencial. A distribuição acumulada pode ser observada no gráfico acima e pode ser escrito como:

```
plot(ecdf(residuals(pop.b)),verticals=T,lty=3, main="Resíduos")
```

48 Regressão sobre variáveis dummies

A introdução de variáveis qualitativas, frequentemente de variáveis dummies, torna o modelo de regressão linear uma ferramenta extremamente flexível, capaz de lidar com muitos problemas encontrados em estudos empíricos.

49 Natureza

Na analise de regressão, a variável dependente é muitas vezes influenciada, não somente pelas variáveis que podem ser quantificadas em alguma escala bem definida (como: renda, produtos, preços, custos, altura e temperatura), mas também por variáveis de natureza essencialmente qualitativas (como: raça, sexo, cor, religião, nacionalidade, guerras terremotos, greves, mudanças na política de governo). Por exemplo, mantendo constante todos os demais fatores verifica-se que professoras universitárias ganham menos que seus colegas homens, e que os não brancos ganham menos que os brancos. Este padrão pode resultar da discriminação sexual ou racial, mas que qualquer que seja a razão, variáveis qualitativas como sexo ou raça de fato influencia a variável dependente, e claramente devem ser incluídas nas variáveis explicativas. Como tais variáveis qualitativas geralmente indicam a presença ou ausência de uma qualidade ou atributo, tais como homem e mulher, negro ou branco, católico e não católico, um método para quantificar tais atributos e construir variáveis artificiais que assumam valores com valores: 1 ou 0, onde 0 indica ausência de um atributo e 1 indica a presença de um atributo, por exemplo, 1 - homem, 0 - mulher; ou 1 - formação superior e 0 - indica o contrario, estas variáveis são chamadas de variáveis dummies. São nomes alternativos: variáveis indicadoras, variáveis binárias, variáveis categóricas, variáveis dicotômicas. As variáveis dummies são usadas nos modelos de regressão tão facilmente quanto ás variáveis quantitativas, alias um modelo de regressão pode conter variáveis explicativas que são exclusivamente dummies, por natureza tais modelos são chamados de modelos de analise de variância (ANOVA). Como exemplo considere os dados em Gujarati (2000), com o seguinte modelo:

$$y_i = \alpha + \beta D_i + \epsilon_i \tag{62}$$

Onde $Yi =$ Salário anual de um professor universitário.
$Di = 0$: sexo feminino; 1: sexo masculino.

O modelo acima permite verificar se o sexo provoca alguma diferença no salário de um professor universitário, supondo naturalmente que idade, título acadêmico, e anos de experiência são mantidos constantes assim,
Salário médio da professoras universitárias: $E(Y|Di = 0) = \alpha$
Salário médio dos professores universitários: $E(Y|Di = 0) = \alpha + \beta$
Onde α indica o termo intercepto como salário médio de professoras, e β informa enquanto o salário de um professor difere da sua colega.

49.0.1 Os dados no R

```
> D=c(0,0,0,0,0,1,1,1,1,1)
> y = c( 19, 18, 18.5, 17, 17.5, 22, 21.7, 21, 20.5, 21.2)
> regdummi=lm(y~D)
> summary(regdummi)
> plot(y, main="salário do professor por sexo")
> abline(a=21,28, b=0, col=8)
> abline(a=18, b=0 )
> legend(2,22,legend="homem")
> legend(8,17.5,legend="mulher")
```

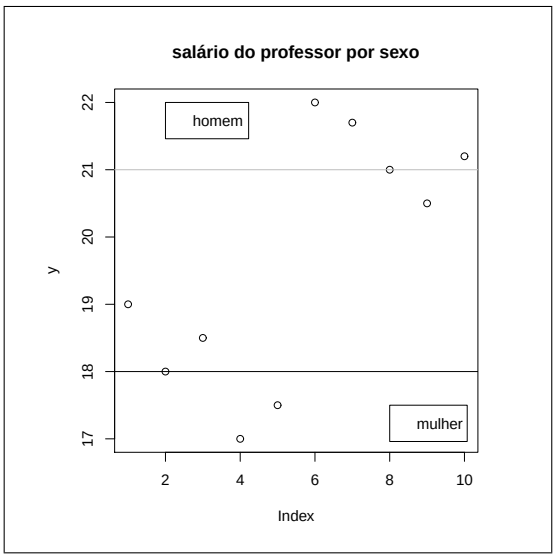


Figura 25. Comparação de salários para os professores

Interpretação: A salário médio de uma professora é de 18 unidades monetárias (α) e o salário de um professor é de 21 unidades monetárias ($\alpha + \beta$)

49.1 Regressão sobre variáveis qualitativas e quantitativas

Considere o seguinte modelo:

$$y_i = \alpha_1 + \alpha_2 D_i + \beta x_i + \epsilon_i \tag{63}$$

Onde $Yi =$ Salário anual de um professor universitário.
$Di = 0$, sexo feminino; 1: sexo masculino.

x_i = anos de experiência de ensino.
Admitindo $E(\epsilon_i) = 0$, podemos identificar os salários da seguinte forma.
Salário médio de uma professora universitária

$$E(y_i|x_i, D_i = 0) = \alpha_1 + \beta x_i \tag{64}$$

E o salário médio de um professor universitário

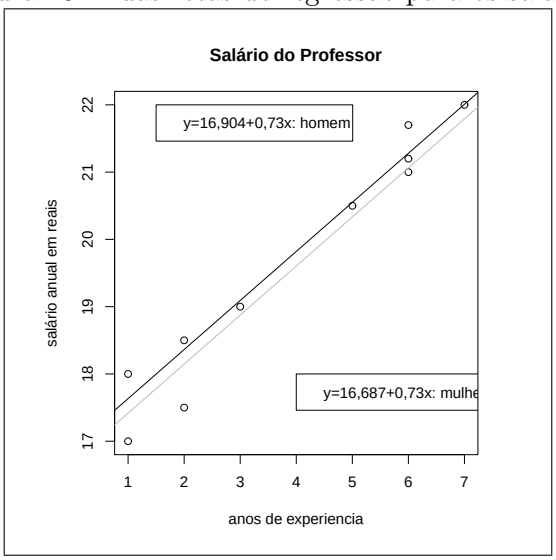
$$E(y_i|x_i, D_i = 1) = \alpha_1 + \alpha_2 + \beta x_i \tag{65}$$

No R, acrescentando a os dados dos salários dos professores os anos de experiencia, temos:

```
> x=c(3,1,2,1,2,7,6,6,5,6)
> mrd=lm(y~D+x)
> summary(mrd)

Call:

lm(formula = y ~ D + x)

Residuals:

    Min      1Q  Median      3Q     Max
-0.6458 -0.2300 -0.0300  0.2969  0.5833

Coefficients:

            Estimate Std. Error t value Pr(>|t|)
(Intercept)  16.6875     0.4086  40.836 1.38e-09 ***
D             0.2175     0.8819   0.247   0.8123
x             0.7292     0.1994   3.657   0.0081 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.4369 on 7 degrees of freedom
Multiple R-squared: 0.9566,     Adjusted R-squared: 0.9442
F-statistic: 77.15 on 2 and 7 DF,  p-value: 1.703e-05
```

A interpretação geométrica deste modelo, é que as funções salário de professores universitários homens e mulheres, em relação aos anos de experiência de ensino, têm a mesma inclinação (β), porem diferentes interceptos. No R podemos construir o gráfico a seguir:

Figura 26. Duas retas de regressão para os salários




```
> plot(x,y, main="Salário do Professor", xlab="anos de experiencia",
> ylab="salário anual em reais")
> abline(a=16.6875,b=0.7292,col=8)
> abline(a=16.904,b=0.7292)
> legend(4,18,legend="y=16,687+0,73x: mulher")
> legend(1.5,22,legend="y=16,904+0,73x: homem")
```

49.1.1 Teste de Chow

Para testar a estabilidade estrutural dos modelos de regressão usamos a função **sctest()**, da biblioteca **strucchange**. Na suposição das variáveis não apenas afetam o intercepto, mais sim o coeficiente de inclinação, deve se testar se este coeficiente de inclinação apresenta diferença significativa nos subgrupos. No exemplo do salário dos professores, determinamos:

```
> library(strucchange)
Carregando pacotes exigidos: sandwich
> sctest(y~D+x,type="Chow", point=5)

        Chow test

data:  y ~ D + x
F = 0.0015, p-value = 0.9999
```

Pelo valor do p-valor, podemos afirmar que não há evidencias para rejeitar a hipótese de igualdade na inclinação dos salários (Não há mudanças no coeficiente de inclinação dos salários entre professoras e professores).

50 Modelos de regressão para respostas binarias

Modelar o comportamento de variáveis dependentes que só assumem valores 0 ou 1, através de uma análise de regressão, onde a $Pr(y = 0)$ ou $Pr(y = 1)$ é afetada pelos regressores. Consideremos a função de ligação denotada por $G(x, \beta)$, onde:
$Pr(y = 1|x) = G(x, \beta)$,
$Pr(y = 0|x) = 1 - G(x, \beta)$,
e x: vetor $p \times n$,
β: vetor $p \times 1$ de parâmetros.
Precisamos encontrar uma estrutura para regredir.

51 Modelos de probabilidade linear

Nossa função de ligação para este modelo é $G(x, \beta) = x'\beta$, para este modelos temos que $E(y|x) = 1 * G(x, \beta) + 0 * (1 - G(x, \beta)) = G(x, \beta)$, cai em nosso modelo linear de regressão, o qual apresenta os seguintes problemas: $E(\epsilon) = 0, V(\epsilon) = x'\beta(1 - x'\beta)$, por tanto o erro (ϵ) é heterocedástico, Não há restrições para $\epsilon = Pr(y = 1|x)$, podendo obter estimativas de ϵ fora de $[0, 1]$.

Uma outra estrutura para regredir é usar modelos para respostas binarias, onde $G(x, \beta)$ esta restrito ao intervalo inteiro padrão($[0, 1]$), onde

$$\lim x'\beta \to -\infty, \qquad Pr(y = 1|x) = 0 \tag{66}$$

$$\lim x'\beta \to \infty, \qquad Pr(y = 1|x) = 1 \tag{67}$$

Para a escolha de G, usamos os seguintes modelos:

51.1 Modelo Probit

A função de ligação G tem distribuição normal padrão,

$$Pr(y = 1|x) = \int_{-\infty}^{x'\beta} \phi(t)dt = \Phi(x'\beta) \tag{68}$$

Onde ϕ é a função de densidade da normal padrão, e Φ é a função de distribuição acumulada da normal padrão.

51.2 Modelo Logit

A ligação de G é a função de distribuição logística, dada por:

$$Pr(y = 1|x) = \frac{\exp x'\beta}{1 + \exp x'\beta} = \Lambda(x'\beta) \tag{69}$$

Modelos alternativos sem simetria são apresentados a seguir

52 Modelo Weibull

Usamos a função da distribuição:

$$Pr(y = 1|x) = \exp - \exp x'\beta \tag{70}$$

53 Modelo C Log-log

Usamos a função da distribuição:

$$Pr(y = 1|x) = 1 - \exp - \exp -x'\beta \tag{71}$$

Dentre os modelos apresentados acima o mais usado é o logit, a distribuição logit e probit são semelhante na parte central, já nas caudas o modelo logit fornece maiores probabilidades de $y = 0$, quando $x'\beta$ é muito pequeno, e $y = 1$, quando $x'\beta$ é muito grande.

53.0.1 Estimação

A estimação dos parâmetros do modelo logit é feita por máxima verosimilhança, consideramos cada observação como um ensaio de Bernoulli, onde:

$$Pr(Y_1 = y_1, Y_2 = y_2, \cdots, Y_t = y_t) = \prod_{y_t=0} [1 - G(X'\beta)] \prod_{y_t=1} [G(x'\beta)] \tag{72}$$

A função de verosimilhança é

$$L(\beta|y, x) = \prod_{t=1}^{T} [1 - G(X'\beta)]^{1-y_t} \prod [G(x'\beta)]^{y_t} \tag{73}$$

Onde a função de log-verosimilhança é dada por:

$$l(\beta|x,y) = \log L = \sum^{T} (1 - y_t) * \log[1 - G(X'\beta)] + y_t * [G(x'\beta)] \qquad (74)$$

Apresentamos um tabela para algumas distribuição de probabilidade com sus respectivas funções de ligação .

Distribuição	Lig. Canônica	Função de variância
Normal	μ	1
Poisson	$\log(\mu)$	μ
Binomial	$\log \frac{\mu}{1-\mu}$	$\mu(1-\mu)$
Gamma	μ^{-1}	μ^2
Gausiana inversa	μ^{-2}	μ^3

Para implementar modelos de regressão para respostas binárias usamos a função **glm()**

53.1 Exemplo 1

Considere o conjunto de dados (dados1), referente Xi = renda em milhares de dólares, Ni = número de famílias com renda Xi, e ni1 = número de famílias que possuem uma casa. (Pág 562. Gujarati)

```
> attach(dados1)
> dados1
    x  Ni   ni1
1   6    40    8
2   8    50   12
3  10    60   18
4  13    80   28
5  15   100   45
6  20    70   36
7  25    65   39
8  30    50   33
9  35    40   30
10 40    25   20
#######################
Logit

> lgm=glm(formula=cbind(ni1, Ni - ni1)~x, family=binomial(logit))
> summary(lgm)
```

```
Call:   glm(formula = cbind(ni, Ni - ni1) ~ x, family = binomial(logit))

Deviance Residuals:
     Min        1Q     Median        3Q       Max
-0.67026  -0.35105  -0.18558   0.06073   1.06817

Coefficients:
             estimate  Std. Error  z value  Pr(>|z|)
(Intercept) -1.60234    0.20403     -7.853  4.05e-15 ***
x            0.07907    0.01011      7.819  5.34e-15 ***

Null deviance: 72.7581  on 9  degrees of freedom
Residual deviance:  2.3542  on 8  degrees of freedom
AIC: 49.01
> fitlgm=fitted(lgm)
######################
Modelo Probit

> Pgm=glm(formula=cbind(ni1, Ni - ni1)~x, family=binomial(probit))
> summary(Pgm)
fitpgm=fitted(Pgm)
######################
Modelo cloglog

> cloglogm=glm(formula=cbind(ni1, Ni - ni1)~x,
family=binomial(cloglog))
> summary(cloglogm)
> fitcloglogm=fitted(cloglogm)

> fitclog=fitcloglogm*Ni
> fitlglm=fitlgm*Ni
> fitpglm=fitpgm*Ni
> comparacao=data.frame(ni1,fitlglm, fitpglm, fitclog)
> comparacao
   ni1   fitlglm     fitpglm    fitclog
1    8   9.781599    9.720856  10.65496
2   12  13.745853  13.721940  14.57061
3   18  18.451892  18.464347  19.09529
4   28  28.816154  28.858391  28.95371
5   45  39.738898  39.768847  39.32643
```

```
6   36  34.632031  34.542022  33.51850
7   39  38.512363  38.324448  37.21732
8   33  34.172042  34.031625  33.49879
9   30  30.489303  30.475689  30.57855
10  20  20.659865  20.757408  21.20739
```

Interpretação: A cada dólar que aumenta na renda das famílias a razão de chances (entre ter uma casa e não ter) aumenta em 7.9 por cento. Para explicar em termos lineares podemos elevar este valor na potencia exponencial para obter a razão de chances, da seguinte forma: $\exp\{0.07907\} = 1.08228$, mostrando que a razão de chances entre ter uma casa e não ter é de 8.22 por cento

53.2 Exemplo 2

Uma empresa esta buscando uma regra para discriminar a qualidade dos funcionários. Em seus arquivo de dados dispor de 109 registros com as seguintes informações dos funcionários $x1$: idade dos funcionários, $x2$: sexo, $x3$: anos de experiência na função, $x4$: estado civil, $x5$: número de filhos, $x6$: nota de 0 - 10 de um teste de avaliação de conhecimento na sua área, $x7$: teste psicotécnico de 0 a 50, $x8$: satisfação com a vida pessoal. Y, definido a priori como grupo1 (desempenho médio) e grupo 2 (melhor desempenho), 0 se faz parte do grupo 1 e 1 se faz parte do grupo 2. ver dados no apêndice.

```
> func=read.table("func.txt",header=T)
> attach(func)
> glm1=glm(grupo ~ tconhec +tpsico, family=binomial(logit))

> summary(glm1)

Call:

glm(formula = grupo ~ tconhec + tpsico, family = binomial(logit))

Deviance Residuals:

    Min       1Q   Median       3Q      Max
-3.1987  -0.0997   0.0130   0.1078   1.3017

Coefficients:
```

```
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -44.2310    11.2683  -3.925 8.66e-05 ***
tconhec      -0.4924     0.2662  -1.850 0.064336 .
tpsico        7.7609     2.1575   3.597 0.000322 ***

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 149.552  on 108  degrees of freedom
Residual deviance:  38.339  on 106  degrees of freedom
AIC: 44.339

Number of Fisher Scoring iterations: 8
> fit=fitted(glm1)
> com=data.frame(grupo,fit) #compara os ajustes do modelo
> anova(glm1)
> glm2=glm(grupo ~ tconhec + tpsico +aexper +ecivil +idade +
nfilhos +satisf +sexo, family=binomial(logit), data=func)
> fit=fitted(glm1) #ajuste do segundo modelo
```

Para discriminar a sensibilidade do modelo usamos o teste a seguir

53.3 Teste de Hosmer Lemeshow

Este teste foi implementado para avaliar o ajuste de um modelo logístico, a função implementada no R é dada por:

```
> hosmer=function (y, yhat, g = 10)
+ {
+   cutyhat <- cut(yhat, breaks = quantile(yhat, probs = seq(0,
+         1, 1/g)), include.lowest = T)
+     obs <- xtabs(cbind(1 - y, y) ~ cutyhat)
+    expect <- xtabs(cbind(1 - yhat, yhat) ~ cutyhat)
+     chisq <- sum((obs - expect)^2/expect)
+      P <- 1 - pchisq(chisq, g - 2)
+     c("X^2" = chisq, Df = g - 2, "P(>Chi)" = P)
+ }
> hosmer(grupo, fit2)
      X^2          Df    P(>Chi)
5.1208521   8.0000000   0.7445847
> hosmer(grupo, fit)
      X^2          Df    P(>Chi)
```

```
7.1358400   8.0000000   0.5220501
```

Referências

[1] Ballie, RT., Chung, CF. e Tieslau, MA.(1996):*Analising inflation by the fractionally integrated ARFIMA-GARCH model*. Journal of applied econometrics, 11, 23 – 40.

[2] Beran, J. (1994): *Statistics for long memory process*. New York: Chapman and Hall.

[3] Borde, Arvind (1992). TEX *by example*. A beginner´s guide. Ed. Academic Press Professional. United States of America.

[4] Carneiro, Orlando (1997). *Ecomometria Básica. Teoria e Aplicações*. Editora Atlas. Segunda Edição. São Paulo.

[5] Conover, W. J. (1999) *Practical nonparametric statistics*. 3.ed. New York.

[6] Cribari Neto, F.; Cabral de Araújo Gois Matheus (2002). *Uma Análise de Monte Carlo do Desempenho de Estimadores de Matrizes de Covarância sob Heterocedasticidade de Forma Desconhecida*. RBE, Rio de Janeiro 56(2), p.309-334.

[7] Cribari Neto, F (2002). *C for Econometricians*. Computational Economics 14. p.135-149.

[8] Reisen, V., Cribari-Neto, F., Jensen, Mark. *Long Memory Inflationary Dynamics: The Case of Brazil*. Studies in Nonlinear Dynamics & Econometrics. *Vol. 7, Issue 3*.(2003)

[9] Ehlers,Ricardo.(2007): Análise de séries Temporais. Universidade Federal do Paraná.

[10] Ferreira , M (2006). *Análise da sensibilidade dos testes de normalidade de Jarque-Bera e Lilliefors em modelos de regressão Linear*. Rev. Mat. Estat, v.24, n.4, p.89-98. São Paulo.

[11] Frery, Alejandro; Cribari-Neto, Francisco.(2011).*Elementos de Estatística Computacional Usando Plataformas de Software Livre/Gratuito*. Publicações Matemáticas. Editora IMPA.

[12] Geweke, J. and Porter-Hudak, S. (1983). The estimation and aplication of long memory time series models. *Journal of Time Series Analysis*, **4**, 221 − 238.

[13] Goldreich, O. and Micali, S. (1986). *How to construct random functions.* Journal of the Association for Computing Machinery, 33(4), p.792−807.

[14] Gujarati, Damodar N (2000). *Ecomometria Básica*. Makron Books. Terceira Edição. São Paulo.

[15] Hill, Carter R.; Griffiths, William E.; Judge, George G (2003). *Ecomometria*. Editora Saraiva. Segunda Edição. São Paulo.

[16] Hosking, J. R. M.(1981). Fractional differencing.Biometrica,**68**, 165 − 176.

[17] Knuth, D.E. (1981). *The Art of Computer Programming, Third Edition. Volume 2: Seminumerical. Algorithms*, Addison-Wesley Publishing Company, Massachusetts.

[18] Korgi, Rodrigo de Castro (2003). *El universo LaTeX*. Editora Facultad de Ciencias. Segunda Edição. Bogota.

[19] Krawczyk, H. *How to Predict Congruential Generators*. In:Journal of Algorithms, V. 13, N. 4, 1992.

[20] L'Ecuyer, P (1994). *Uniform random number generation*. Annals of Operations Research, pp. 77 − 120.

[21] L'Ecuyer, P (2001). *Software for uniform random number generation: distinguishing the good and the bad*. In Proceedings of the 2001 Winter Simulation Conference.

[22] Matsumoto, M.; Nishimura, T. (1998). *Mersenne twister: a 623-dimensionally equidistributed uniform pseudo-random number generator*. ACM Transactions on Modeling and Computer Simulation, 8.1.

[23] Marsaglia, G. *The Marsaglia Random Number including the DIEHARD Battery of Tests of Randomness*
(Disponível em: http://stat.fsu.edu/pub/diehard/).

[24] Marsaglia, G. *A Current View of Random Number Generators, Keynote Address, Computer Science and Statistics*. In: 16th Symposium on the Interface, Atlanta. Published by Elsevier Press, 1984.

[25] Marsaglia, G.(1993) *Monkey Tests for Random Number Generators*. In: Computers and Mathematics with Applications, 9, p.1-10, 1993.

[26] McCullough (1998). *Benchmarking Statistical software. American Statistician*, Fortcoming.

[27] Mingoti(2005).*Analise De Dados Atraves De Metodos De Estatistica Multivariada - Uma Abordagem Aplicada*. Editora UFMG. Belo Horizonte.

[28] Morettin, A. P., Clélia, M. C.(2006) *Análise de séries temporais*. Editora Egard Blucher, 2^{da} edição.

[29] Papoulis, A. e Pillai, S. (2002). *Probability, Random Variables and Stochastic Processes*. McGraw-Hill, 4th edition.

[30] R Development Core Team (2010). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

[31] Reisen, V., Lemonte, A.(2006). The fracdiff package Version $1.3 - 1$.

[32] Reisen, V. (2007): Minicurso: MODELO ARFIMA.

[33] Soto, J. *Statistical Testing of Random Number Generators*. In: Proceedings of the 22nd National Information Systems Security Conference, Crystal City, Virginia, 1999.
(Disponível em: http://csrc.nist.gov/rng/nissc-paper.pdf).

[34] VenablesWN, Ripley BD (2000). *S Programming*. Springer-Verlag, New York.

[35] Venables WN, Ripley BD (2002). *Modern Applied Statistics with S*. 4th edition. Springer-Verlag, New York.

[36] Wong, P. C (1990). *Random number generation without multiplication*. In 9th Annual Intl Phoenix Conf on Computers and Communications, p.217-221.

[37] Zafon, G; Manacero Jr, Aleardo (2006). *Construção de Geradores independentes de números aleatórios para diferentes distribuições probabilísticas*. Anais do XXVI Congresso de SBC, p.16-21. Campo Grande. MS