Introdução ao software R

Correlação e Regressão

www.de.ufpb.br

https://www.youtube.com/estatisticalivre

ESTATÍSTICA APLICADA EM SOFTWARE LIVRE

UFPB Departamento de ESTATÍSTICA

Introdução

Quando existe o interesse em analisar a relação linear entre duas variáveis quantitativa (X e Y, por exemplo) duas técnicas podem ser consideradas:

- **Correlação:** Quantifica a força dessa relação, com uma medida que resume o grau de relacionamento entre duas variáveis

- **Regressão:** Modela a forma dessa relação, tendo como resultado uma equação matemática que descreve o relacionamento entre variáveis.

Motivação

Como um exemplo de um problema no qual a análise de correlação e regressão pode ser útil, suponha que um engenheiro está estudando sobre o sistema de abastecimento de máquinas de venda automática de refrigerantes. Ele está interessado em entender como o número de refrigerantes estocados na máquina, se relaciona com o tempo necessário para o funcionário abastecer e fazer a manutenção de rotina das máquinas.

- *Time:* Tempo de execução (minutos)
- *Cases:* Número de refrigerantes

Análise Gráfica

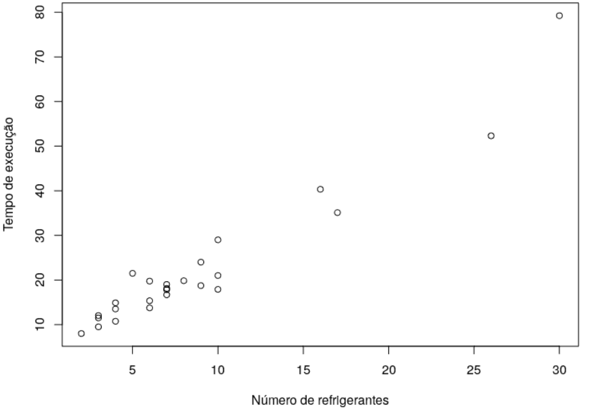
A análise gráfica é um dos primeiros passos para poder observar se existe e de se ter alguma ideia do tipo de relação estatística entre duas variáveis.

O gráfico utilizado para analisar o comportamento conjunto de duas variáveis quantitativas é chamado **gráfico de dispersão**.

Este gráfico é feito no R com a função *plot()*

Gráfico de Dispersão no R

```
> plot(Cases, Time, xlab = "Número de refrigerantes",
ylab = "Tempo de execução")
```



O gráfico de dispersão sugere claramente uma relação crescente entre o tempo de execução e o número de refrigerantes estocados.

Coeficiente de correlação linear de Pearson

Uma medida do grau e do sinal da correlação linear entre duas variáveis (X, Y) é dado pelo **Coeficiente de Correlação Linear de Pearson**, definido por:

$$r = \frac{\text{Cov}(X, Y)}{S_X S_Y},$$

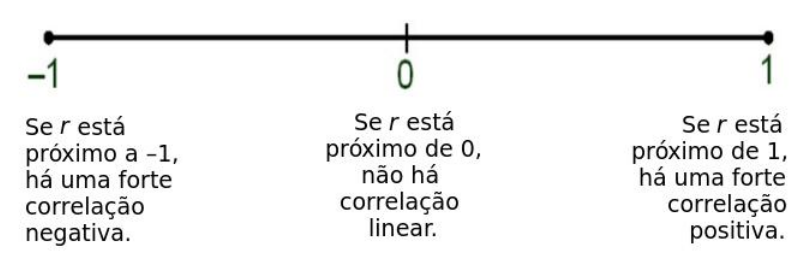
em que S_X e S_Y representam o desvio padrão amostral das variáveis X e Y, respectivamente, e $\text{Cov}(X, Y)$ é a covariância entre elas, definida por:

$$\text{Cov}(X, Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

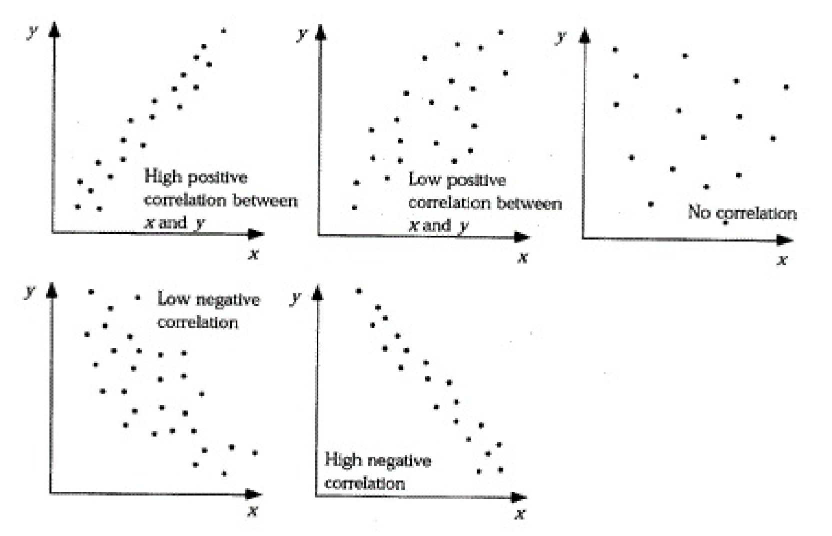
Coeficiente de Correlação Linear

Este coeficiente é adimensional, logo não é afetado pelas unidades de medidas das variáveis X e Y.

Temos que $-1 \leq r \leq 1$. O sinal **positivo** indica que a relação entre as variáveis é diretamente proporcional, enquanto que o sinal **negativo** indica relação inversamente proporcional.



Alguns exemplos



Já observamos que havia uma relação crescente e linear entre as variáveis *Time* e *Cases*.

Para medir o grau dessa relação, calculamos o coeficiente de correlação linear de Pearson entre as variáveis, que é obtido por:

```
> cor(Cases,Time)
[1] 0.9646146
```

Este valor indica uma correlação forte e positiva entre o número de refigerantes estocados e o tempo de reparo da máquina.

Teste de Hipóteses para o Coeficiente de Correlação

- [1] Definição das hipóteses:

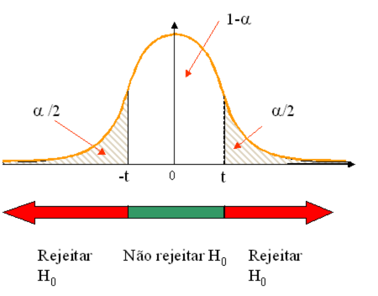
$$\begin{aligned} H_0 &: \rho = 0 \quad \text{(não existe correlação linear)} \\ H_1 &: \rho \neq 0 \quad \text{(existe correlação linear)} \end{aligned}$$

- [2] Fixar o nível de significância α;
- [3] Definir a estatística do teste:

$$T = r\sqrt{\frac{n-2}{1-r^2}} \sim t_{(n-2)}$$

Teste de Hipóteses para o Coeficiente de Correlação

- [4] Definir a região crítica do teste (RC):



- [5] Calcular a estatística do teste T_c.
- [6] Se T_c pertence a RC $\Rightarrow$ rejeitar H_0. Se T_c não pertence a RC $\Rightarrow$ não rejeitar H_0.
- [7] Concluir sobre a decisão tomada no passo 6.

Teste de Hipóteses para a Correlação no R

```
> cor.test(Cases,Time)

Pearson's product-moment correlation

data:  Cases and Time
t = 17.546, df = 23, p-value = 8.22e-15
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
0.9202275 0.9845031
sample estimates:
cor
0.9646146
```

Com p-value$= 8.22e - 15$, rejeitamos a hipóteses nula, que a correlação é igual a zero, e concluímos que a relação linear entre o número de refigerantes estocados e o tempo de reparo da máquina é estatisticamente significante.

Exercício

Agora faça a analise do relacionamento entre as variáveis *Time* e *Distance*.

Inicie com o gráfico de dispersão, calcule o coeficiente de correlação linear de Pearson e realize o teste para saber se a correlação linear entre essas variáveis é estatisticamente significante.

Regressão Linear Simples

Iniciaremos o estudo de regressão com a formulação mais simples, relacionando uma **variável Y**, chamada de **variável resposta** ou **dependente**, com uma **variável X**, denominada de **variável explicativa** ou **independente**.

O modelo em que busca explicar uma variável Y como uma função linear de apenas uma variável X é denominado de modelo de regressão linear simples.

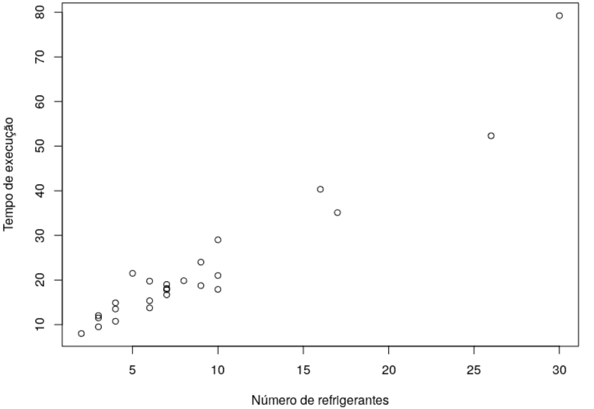
A aplicação da regressão é geralmente feita sob um referencial teórico, que justifique uma relação matemática de causalidade.

Motivação

Como um exemplo de um problema no qual a análise de regressão pode ser útil, suponha que um engenheiro está estudando sobre o sistema de abastecimento de máquinas de venda automática de refrigerantes. Ele está interessado em desenvolver um método para prever do tempo necessário para o funcionário abastecer e fazer a manutenção de rotina das máquinas como uma função do o número de refrigerantes que serão estocados.

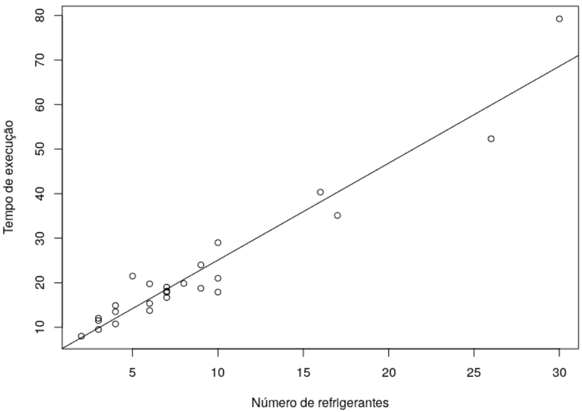
- $\boldsymbol{Y}$: Tempo de execução (minutos) $\rightarrow$ Variável resposta.
- $\boldsymbol{X}$: Número de refrigerantes $\rightarrow$ Variável explicativa.

O gráfico de dispersão sugere claramente uma relação crescente entre o tempo de execução e o número de refrigerantes estocados.



Como o objetivo é modelar o relacionamento das variáveis por meio de uma equação matemática, uma forma simples e razoável é através da relação linear.

Analisando o gráfico com a ilustração desse relacionamento em linha reta, observamos que os pontos dos dados não caem exatamente em linha reta.



Mas, como Y é uma variável aleatória, para o conceito de regressão, é incluido no modelo linear um *componente aleatório*, chamado de erro.

Regressão Linear Simples

A equação que relaciona a variável resposta Y com a variável independente X pode ser escrita da seguinte maneira:

$$Y_i = \alpha + \beta x_i + \epsilon_i,$$

- Y_i é a variável aleatória associada à i-ésima observação de Y;
- x_i é a i-ésima observação do valor fixado para a variável independente X;
- ϵ_i é o **erro aleatório** da i-ésima observação;
- α e β são parâmetros (intercepto e inclinação) que precisam ser estimados.

Ampliando a percepção sobre o modelo

É importante perceber que, na **análise de regressão**, que o regressor X é uma variável controlada (fixa) e que para cada possível valor de x existe uma distribuição de probabilidade para a variável resposta y.

Além disso, devemos pensar em ϵ como um erro estatístico, ou seja, uma variável aleatória que explica a falha do modelo no ajuste dos dados.

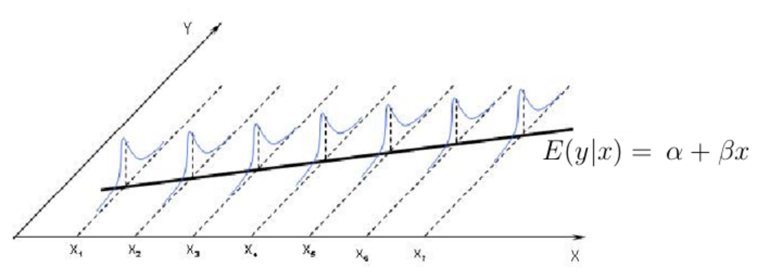
E, adicionalmente, fazemos a suposição que

$$E(\epsilon) = 0 \qquad Var(\epsilon) = \sigma^2$$

Interpretando o modelo

Assim, interpretamos a reta sendo a linha de valores médios (ou esperado) da variável resposta y para um dado valor de x.

Além disso, para cada valor de x, a variabilidade de y se mantém constante e é σ^2, a variância do erro.

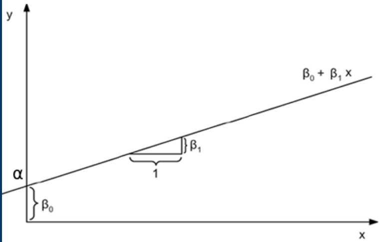


Interpretando os parâmetros

Os parâmetros α e β são geralmente chamados de coeficientes de regressão. Eles têm uma interpretação simples e bastante útil.

- O intercepto α é a média ($\mu_{Y|x}$) quando $x = 0$. Se o intervalo de x não inclui zero, então α não possui interpretação prática.

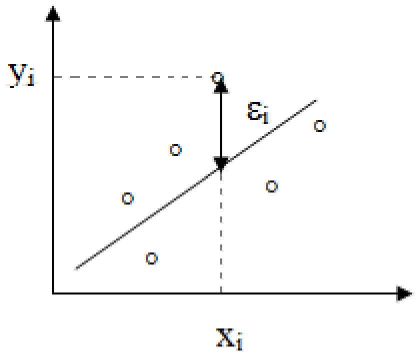
- A inclinação β é a alteração na média da distribuição de y produzida por uma mudança de unidade em x.



Estimando os Parâmetros do Modelo

Como escolher a reta que "MELHOR" se ajusta aos dados?

Queremos encontrar a reta que passe o mais próximo possível dos pontos observados.



- Uma ideia inicial: nosso modelo envolve erros, podemos tentar minimizá-los!

Método de Mínimos Quadrados

Aplicando-se derivadas parciais à expressão anterior, e igualando-se a zero, acharemos as seguintes estimativas para α e β, as quais chamaremos de a e b, respectivamente:

$$b = \frac{n\sum_{i=1}^{n} x_i Y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} Y_i\right)}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

e

$$a = \frac{\sum_{i=1}^{n} Y_i - b\sum_{i=1}^{n} x_i}{n}$$

O modelo ajustado

A chamada equação (reta) de regressão é dada por:

$$\hat{y} = b + ax$$

e para cada valor x_i ($i = 1, \ldots, n$) temos, pela equação de regressão, o valor predito:

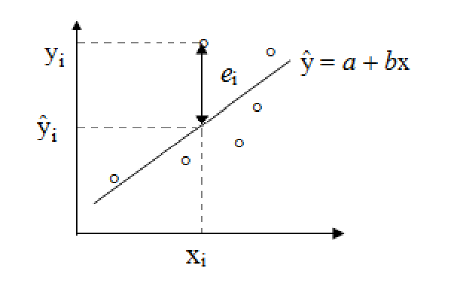
$$\hat{y}_i = b + ax_i$$

A diferença entre os valores observados e os preditos é chamada de resíduo:

$$e_i = y_i - \hat{y}_i$$

Método de Mínimos Quadrados

O resíduo relativo à *i*-ésima observação (e_i) pode ser considerado uma estimativa do erro aleatório (ϵ_i) desta observação (veja ilustração abaixo).



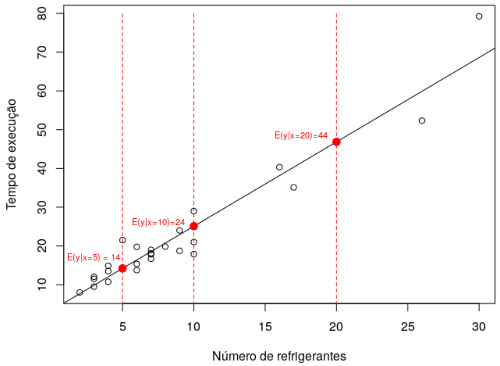
Como medir a “qualidade” do modelo?

Exemplo: Ajustando o modelo

```
> ajuste<-lm(salario~exp)
> summary(ajuste)

Call:
lm(formula = salario ~ exp)

Residuals:
    Min      1Q  Median      3Q     Max
-875.32 -137.49   87.12  237.04  407.18

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 2708.606    122.296   22.15   <2e-16 ***
exp          151.111      7.514   20.11   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1

Residual standard error: 316.7 on 25 degrees of freedom
Multiple R-squared:  0.9418,Adjusted R-squared:  0.9395
F-statistic: 404.5 on 1 and 25 DF,  p-value: < 2.2e-16
```

Interpretando o modelo do exemplo

A reta representa os valores médios do tempo de abastecimento e manutenção das máquinas (em minutos) para cada número especifíco de refrigerantes estocados:



Seja a reta:

$$E(Y|x) = 4 + 2X$$

- 4 é o intercepto
- 2 é a inclinação

Temos que, o tempo médio para manutenção da máquina quando não há a reposição de refrigerante é de 4 minutos e, para cada unidade de refrigerante estocado, haverá um aumento de 2 minutos neste tempo médio.

O Coeficiente de Determinação (R^2)

O coeficiente de determinação é uma medida descritiva da proporção da variação de Y que pode ser explicada por variações em X, segundo o modelo de regressão especificado. Ele é dado pela seguinte razão:

$$R^2 = \frac{\sum_{i=1}^{n} (\hat{y} - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} = \frac{\text{variação explicada pelo modelo}}{\text{variação total}}$$

onde $\bar{y} = \sum_{i=1}^{n} \frac{y_i}{n}$.

Note que $0 \leq R^2 \leq 1$. Se $R^2 = 0$, o modelo não tem nenhum poder explicativo. Se $R^2 = 1$, o poder explicativo do modelo é total.

Teste de Hipóteses para o Coeficiente β

- [1] Definição das hipóteses:

$$\begin{aligned} H_0 &: \beta = 0 \\ H_1 &: \beta \neq 0 \end{aligned}$$

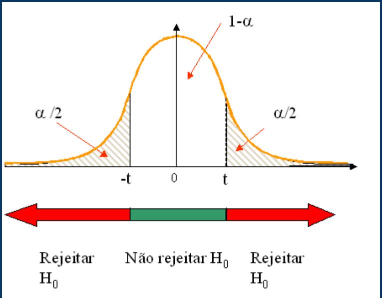
- [2] Fixar o nível de significância α;
- [3] Determinar a estatística do teste:

$$T = \frac{|b|}{S_b} \sim t_{(n-2)}$$

em que $S_b^2 = \dfrac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{(n-2) \sum_{i=1}^{n} (x_i - \bar{x})^2}$.

Teste de Hipóteses para o Coeficiente β

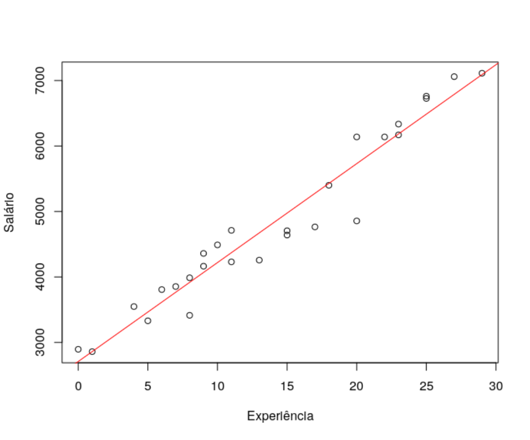
- [4] Definir a região crítica do teste (RC):



- [5] Calcular a estatística do teste T_c.
- [6] Se T_c pertence a RC $\Rightarrow$ rejeitar H_0. Se T_c não pertence a RC $\Rightarrow$ não rejeitar H_0.
- [7] Concluir sobre a decisão tomada no passo 6.

Exemplo: gráfico com reta ajustada

```
> plot(exp,salario, xlab = "Experiência", ylab = "Salário")
> abline(ajuste, col="red")
```



Exemplo: predição

```
> ajuste

Call:
lm(formula = salario ~ exp)

Coefficients:
(Intercept)          exp
     2708.6        151.1

> predict(ajuste, newdata=data.frame(exp=c(10,11)),
interval="prediction")
       fit      lwr      upr
1 4219.712 3552.440 4886.984
2 4370.823 3704.848 5036.797
```

Teste de hipóteses

Uma vez que estimamos os parâmetros do modelo, enfrentamos duas perguntas imediatas:

- Qual a significância da regressão?
- Qual o poder explicativo desse modelo para variável resposta?
- O modelo se adequa bem aos dados?
- As suposições do modelo são satisfeitas?

Teste de hipóteses, medidas descritivas e gráficos podem ser úteis para abordar essas questões.

Teste de Hipóteses para o Coeficiente β

Este teste é considerado para verificar a significância da regressão.

- [1] Definição das hipóteses:

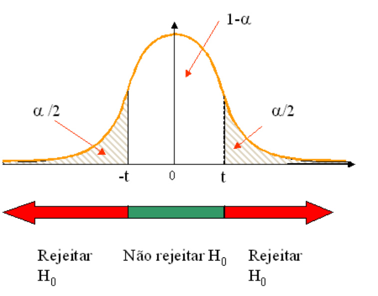
$$\begin{aligned} \mathrm{H}_0 &: \beta = 0 \\ \mathrm{H}_1 &: \beta \neq 0 \end{aligned}$$

- [2] Fixar o nível de significância α;
- [3] Determinar a estatística do teste:

$$T = \frac{|b|}{S_b} \sim t_{(n-2)}$$

em que $S_b^2 = \dfrac{\displaystyle\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{(n-2)\displaystyle\sum_{i=1}^{n}(x_i - \bar{x})^2}$.

- [4] Definir a região crítica do teste (RC):



- [5] Calcular a estatística do teste T_c.
- [6] Se T_c pertence a RC $\Rightarrow$ rejeitar H_0. Se T_c não pertence a RC $\Rightarrow$ não rejeitar H_0.
- [7] Concluir sobre a decisão tomada no passo 6.

Estes teste considera que os erros são independentes e têm distribuição $\mathcal{N}(0, \sigma^2)$.

O Coeficiente de Determinação (R^2)

O coeficiente de determinação é uma medida descritiva da proporção da variação de Y que pode ser explicada por variações em X, segundo o modelo de regressão especificado. Ele é dado pela seguinte razão:

$$R^2 = \frac{\displaystyle\sum_{i=1}^{n}(\hat{y} - \bar{y})^2}{\displaystyle\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{\text{variação explicada pelo modelo}}{\text{variação total}}$$

onde $\bar{y} = \displaystyle\sum_{i=1}^{n} \frac{y_i}{n}$.

Note que $0 \leq R^2 \leq 1$. Se $R^2 = 0$, o modelo não tem nenhum poder explicativo. Se $R^2 = 1$, o poder explicativo do modelo é total.

Obtendo estas estatísticas no R

```
> summary(ajuste)

Call:
lm(formula = Time ~ Cases)

Residuals:
Min       1Q   Median       3Q      Max
-7.5811 -1.8739 -0.3493  2.1807 10.6342

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)     3.321       1.371   2.422   0.0237 *
Cases           2.176       0.124  17.546 8.22e-15 ***
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1

Residual standard error: 4.181 on 23 degrees of freedom
Multiple R-squared:  0.9305,Adjusted R-squared:  0.9275
F-statistic: 307.8 on 1 and 23 DF,  p-value: 8.22e-15
```

Verificação das Suposições do Modelo

Devemos sempre checar a validade dessas suposições no modelo ajustado, ao passo que a violação desses pressupostos podem ter consequências sérias, como a obtenção de um modelo instável.

Normalmente, as inadequações do modelo não podem ser detectadas examinando as estatísticas t ou R^2. Visto que, estas estatísticas são utilizadas para analisar a estimação "geral" do modelo.

Discutiremos a seguir os tipos de inadequações do modelo e quais os métodos úteis para diagnosticar as violações das supoisçoes básicas de regressão.

Suposições do Modelo

S1. A relação entre Y e X é linear.

S2. $\mathrm{E}(\epsilon) = 0$.

S3. $\mathrm{Var}(\epsilon) = \sigma^2$ e constante em todo o modelo.
S4. Os erros são não-correlacionados
S5. Os erros são normalmente distribuídos.

Análise de Resíduos

Os métodos para detectar problemas com o ajuste e/ou quebra das hipóteses primárias de um modelo de regressão normal linear baseiam-se, principalmente, no estudo dos resíduos do modelo.

O resíduo é a diferença entre os valores observados e os preditos é chamada de resíduo:

$$e_i = y_i - \hat{y}_i$$

O resíduo relativo à i-ésima observação (e_i) pode ser considerado uma estimativa do erro aleatório (ϵ_i) desta observação. Assim, qualquer afastamento das suposições dos erros devem aparecer nos resíduos.

Análise Gráfica dos Resíduos

A análise gráfica dos resíduos é uma maneira muito eficaz de investigar o quão bem o modelo de regressão se adequa aos dados.

Apresentaremos alguns gráficos que devem ser sempre examinados quando um modelo de regressão está sendo ajustado. No R são obtidos da forma

```
> plot(ajuste)
Aperte <Enter> para ver o próximo gráfico:
```

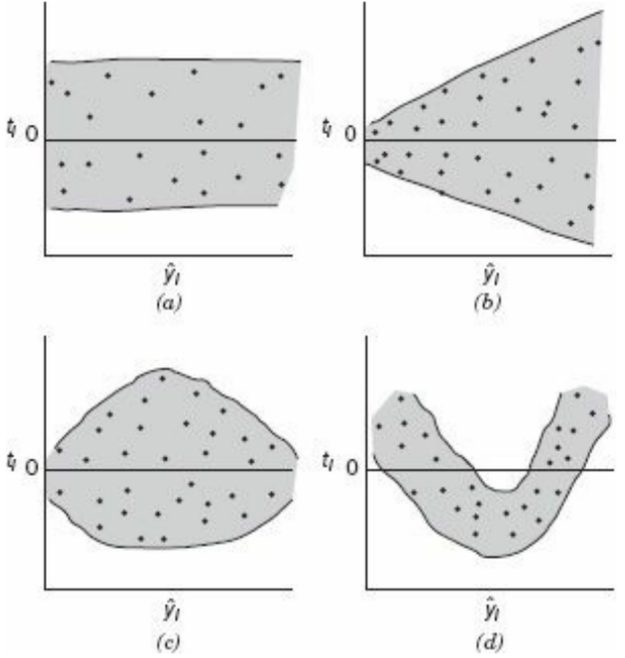
e como retorno, teremos quatro gráficos para a análise de diagnóstico dos resíduos.

Gráfico de Resíduos *versus* Valores Ajustados

O primeiro gráfico é dos resíduos *versus* os correspondentes valores ajustados, e é útil para detectar vários tipos comuns de inadequações do modelo, tais como

- a presença de ***outliers*, pontos de alavanca ou de influência** (observações que se diferem do resto dos dados e que podem prejudicar o ajuste e a adequação do modelo)
- heterocedasticidade (variação dos erros não é constante);
- o modelo pode não ser linear;

Se neste gráfico os resíduos estiverem contidos, de forma aleatória, em uma faixa horizontal, então não há indícios desses problemas no modelo.



t_i 0
$\hat{y}_i$
(a)
t_i 0
$\hat{y}_i$
(b)
t_i 0
$\hat{y}_i$
(c)
t_i 0
$\hat{y}_i$
(d)

Técnica Gráfica para Normalidade

O segundo gráfico, é conhecido como *q-q plot* normal e é utilizado para verificar a suposição de normalidade.

Este gráfico indicará normalidade quando os pontos estiverem aproximadamente em linha reta de 45°.

Pequenos desvios da suposição de normalidade não afetam muito a estimação do modelo, mas a falta de normalidade pode acarretar em testes e os intervalos de confiança pouco confiáveis, visto que são construídos sob esta suposição.

Análise Gráfica dos Resíduos

O terceiro gráfico, que é a raiz quadrada dos resíduos padronizados *versus* os valores ajustados. e utilizado com a mesma finalidade do primeiro gráfico.

E quarto gráfico é utilizado para fazer a detecção de pontos de alavanca.