



# ME613 - Análise de Regressão

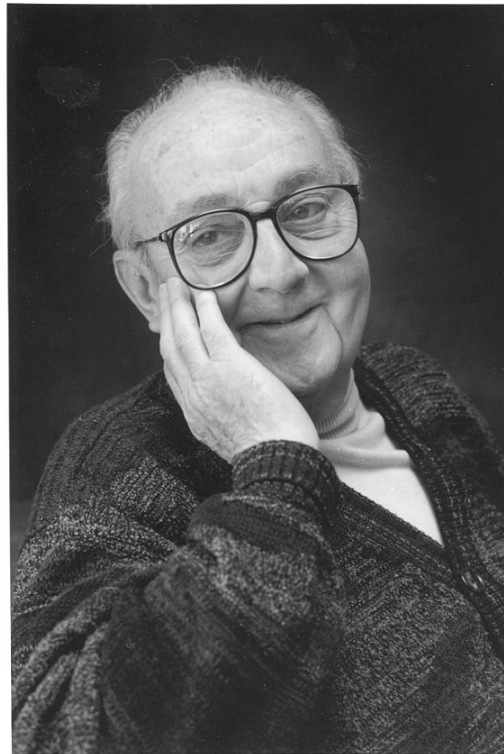
## Parte 3

Benilton S Carvalho - 1S2020

# Diagnóstico do Modelo

# George Box

"All models are wrong but some are useful."



[By DavidMCEddy at en.wikipedia, CC BY-SA 3.0](#)

# Introdução

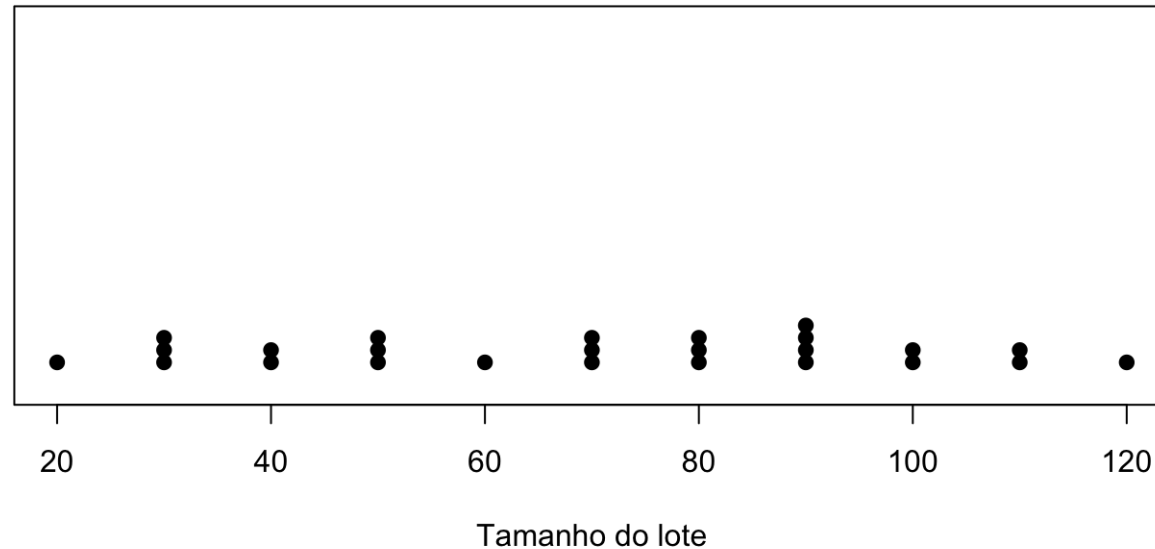
Como saber se a função de regressão é adequada para explicar a relação entre as variáveis observadas?

- Gráficos
- Testes

Caso a função de regressão não seja adequada, o que fazer?

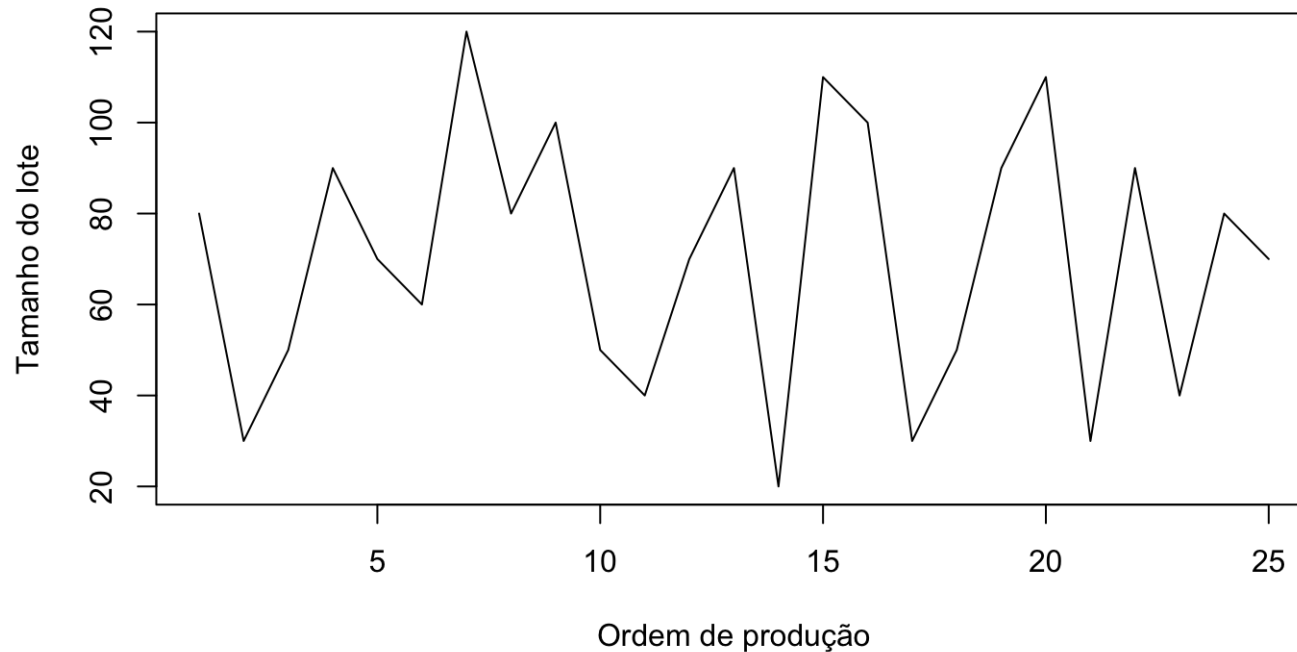
- Transformação de variáveis.
- Outro tipo de modelo.

# Diagnósticos para a variável preditora



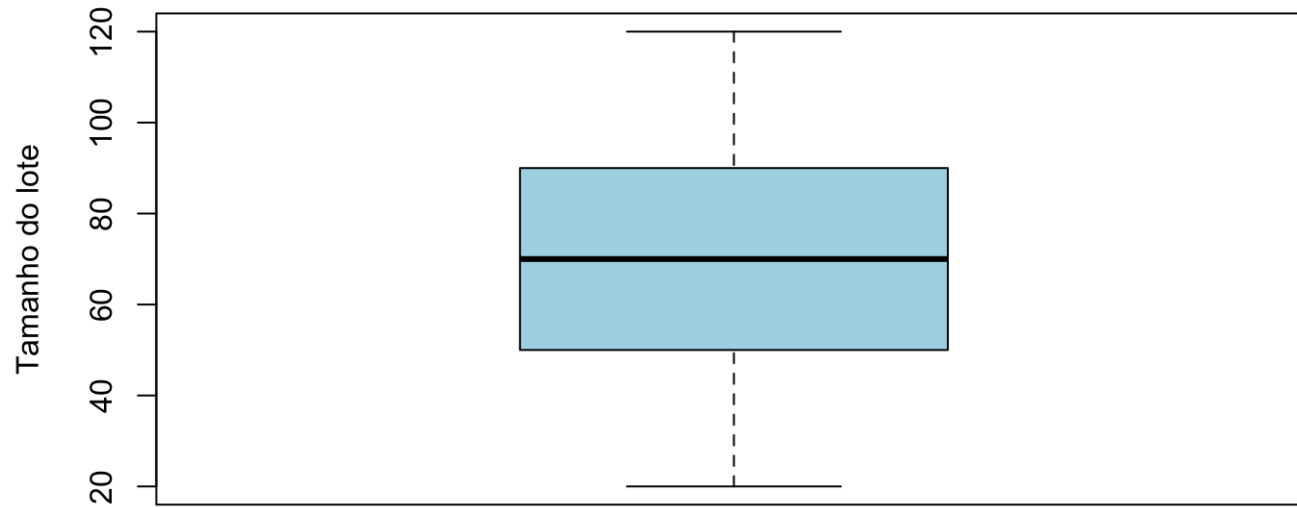
- Quantas observações em cada valor diferente de  $X$ ?
- Amplitude dos valores de  $X$ .

# Diagnósticos para a variável preditora



Se as observações são feitas ao longo do tempo, há correlação entre seus valores e o tempo de observação?

# Diagnósticos para a variável preditora



# Resíduos

Gráficos de diagnósticos para  $Y$  não são úteis em análise de regressão, pois a variável resposta é uma função da preditora.

Os gráficos de diagnósticos são feitos para os resíduos.

- Resíduos:

$$e_i = Y_i - \hat{Y}_i$$

- Erros:

$$\varepsilon_i = Y_i - E(Y_i)$$

Se assumimos  $\varepsilon_i \stackrel{iid}{\sim} \mathcal{N}(0, \sigma^2)$ , esperamos que  $e_i$ 's, os resíduos observados, reflitam as propriedades assumidas para  $\varepsilon_i$ 's.



# Propriedades dos resíduos

- Média dos resíduos:

$$\bar{e} = \frac{\sum_{i=1}^n e_i}{n} = 0$$

Como  $\bar{e}$  é sempre nula, não fornece informação sobre  $E(\varepsilon_i)$ .

- Variância dos resíduos:

$$s^2 = \frac{\sum_{i=1}^n (e_i - \bar{e})^2}{n - 2} = \frac{SQErro}{n - 2} = QMErro$$

Se o modelo é apropriado,  $s^2$  é um estimador não viesado para  $\sigma^2$ .

# Propriedades dos resíduos

- Independência:  $e_i$ 's não são variáveis aleatórias independentes, pois dependem de  $\hat{Y}_i$ , que por sua vez dependem das mesmas estimativas para os parâmetros  $\beta_0$  e  $\beta_1$ .

Quanto maior o número de observações com relação ao número de parâmetros estimados, menor é a dependência dos resíduos (pode ser ignorada).

# Resíduos semi-studentizados

Para detectar outliers, muitas vezes é mais útil observar resíduos padronizados.

Resíduo semi-studentizado:

$$e_i^* = \frac{e_i}{\sqrt{QME}}$$

# Suposições do modelo estudadas através da análise de resíduos

Podemos utilizar análise de resíduos para verificar se:

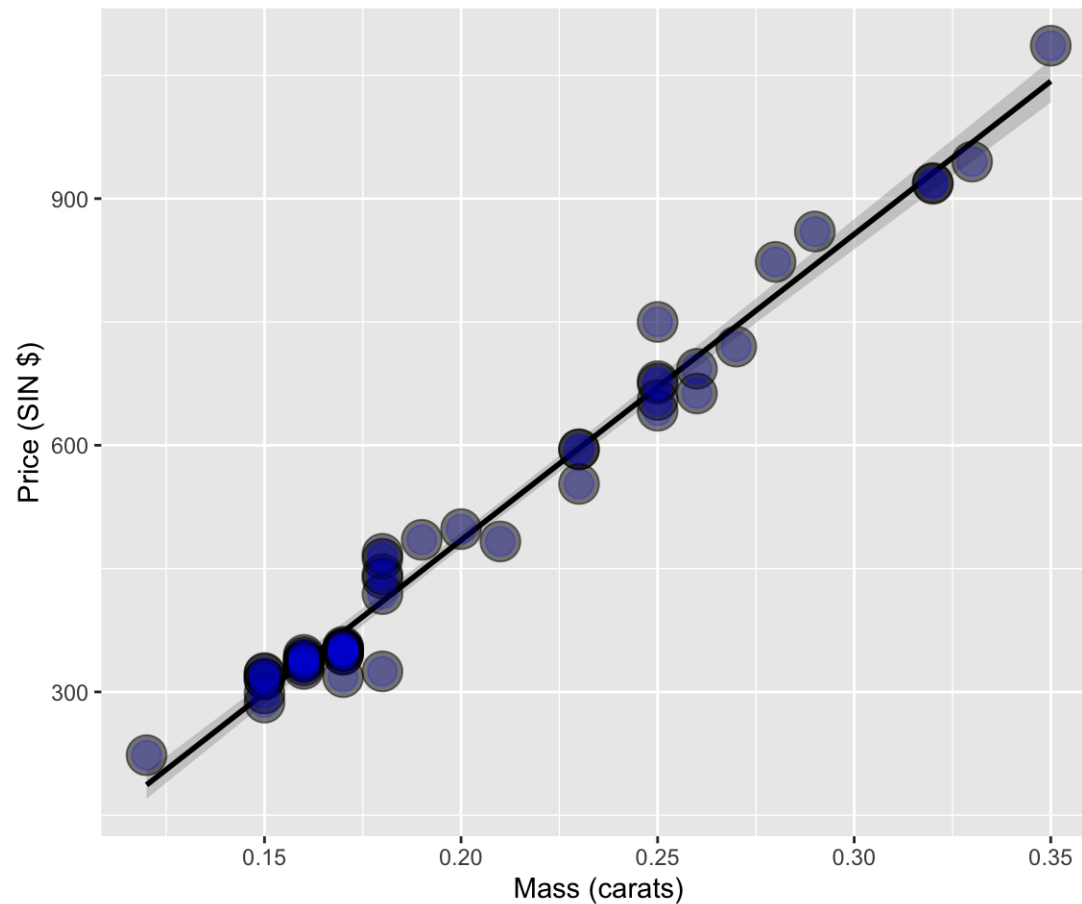
1. a função de regressão não é linear
2. erros não possuem variância constante (heterocedasticidade)
3. erros não são independentes
4. o modelo ajustado é adequado, com exceção de algumas observações outliers
5. erros não seguem distribuição normal
6. uma ou mais variáveis preditoras foram omitidas no modelo

# Análise de Resíduos: gráficos

- Resíduos x Variável preditora
- Resíduos absolutos x Variável preditora
- Resíduos x Valores ajustados ( $\hat{Y}$ )
- Resíduos ao longo do tempo/em sequência
- Resíduos X Variável preditora não incluída no modelo
- Boxplot dos resíduos
- Gráfico de probabilidade normal dos resíduos

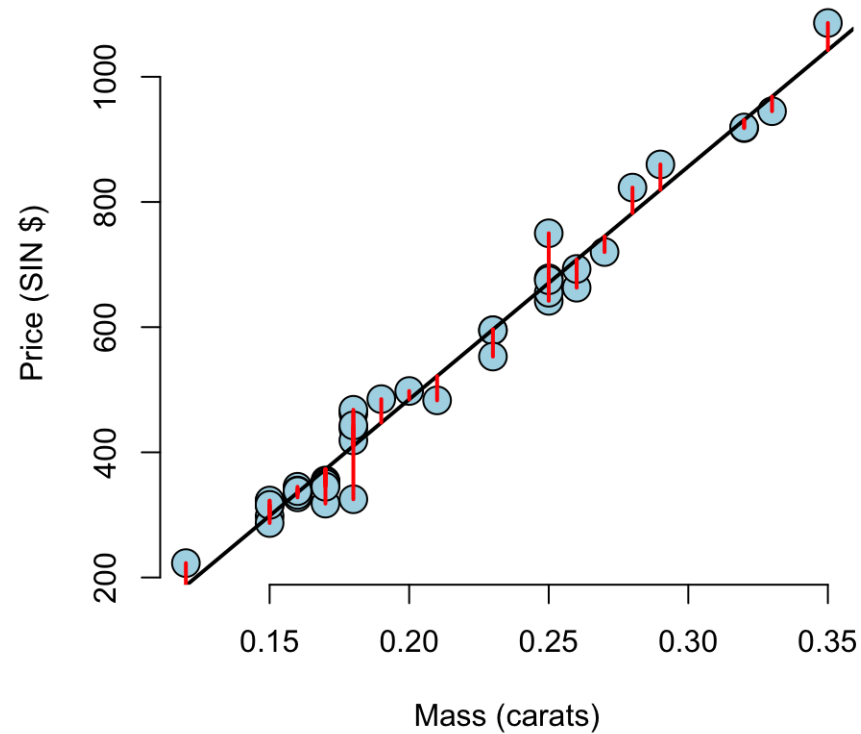
# Exemplo: Preço do diamante

Preços de diamantes (dólares de Singapura) e peso em quilates (1 quilate = 0.2g).



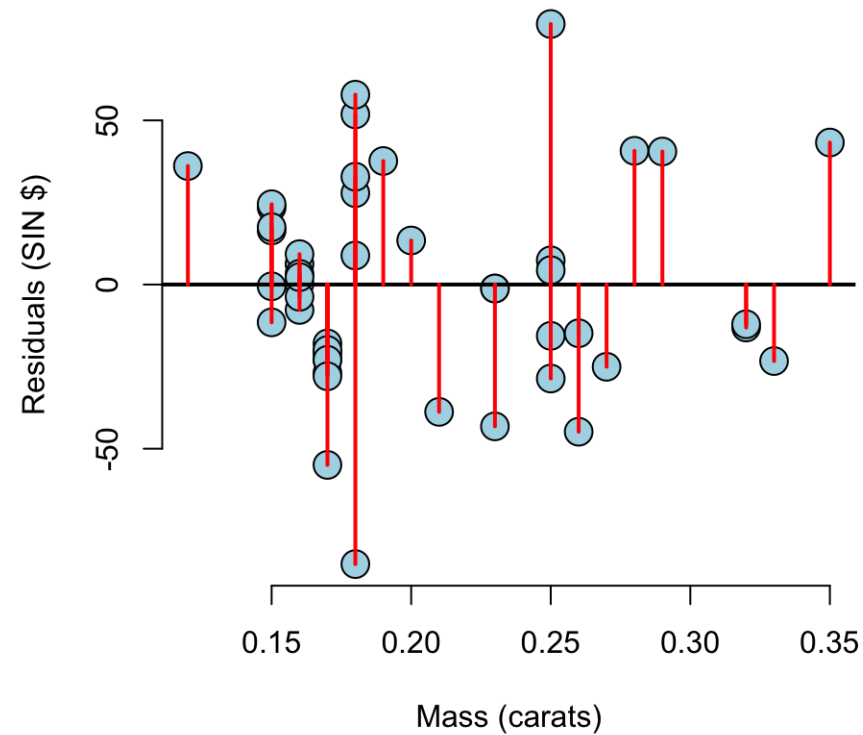
# Exemplo: Preço do diamante

Resíduos: comprimento das linhas vermelhas, mantendo o sinal.



# Exemplo: Preço do diamante

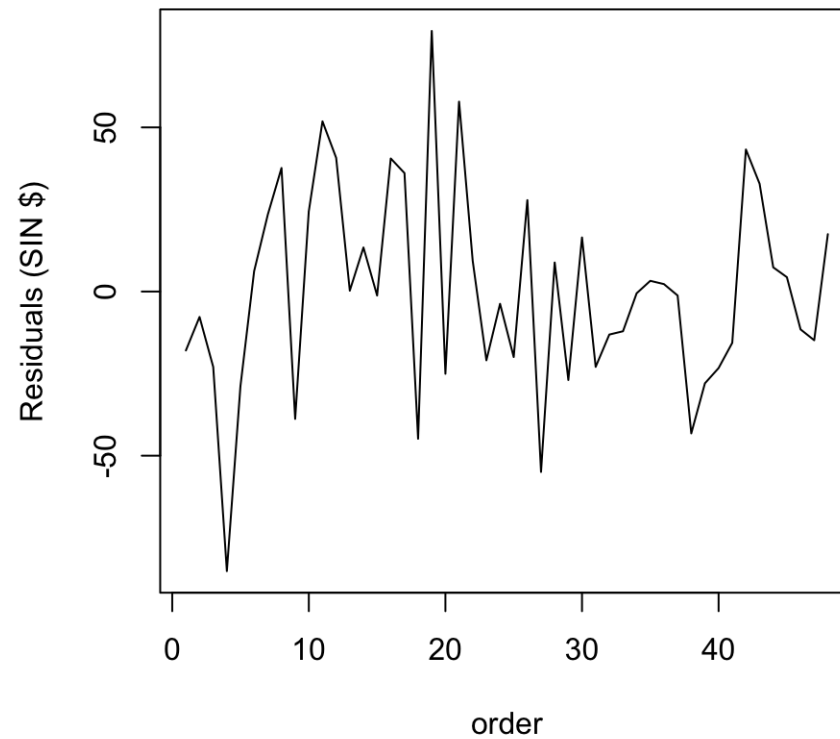
Resíduos X Variável preditora





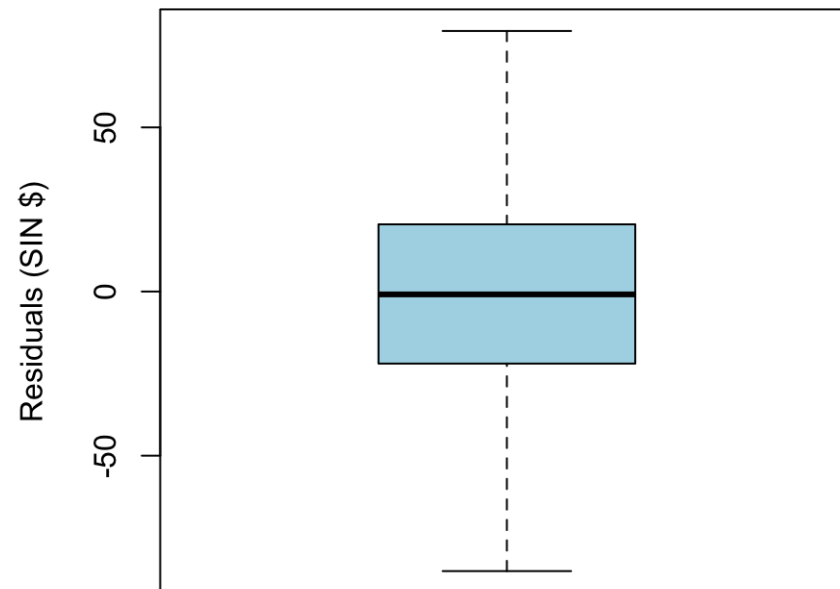
# Exemplo: Preço do diamante

Resíduos X ordem



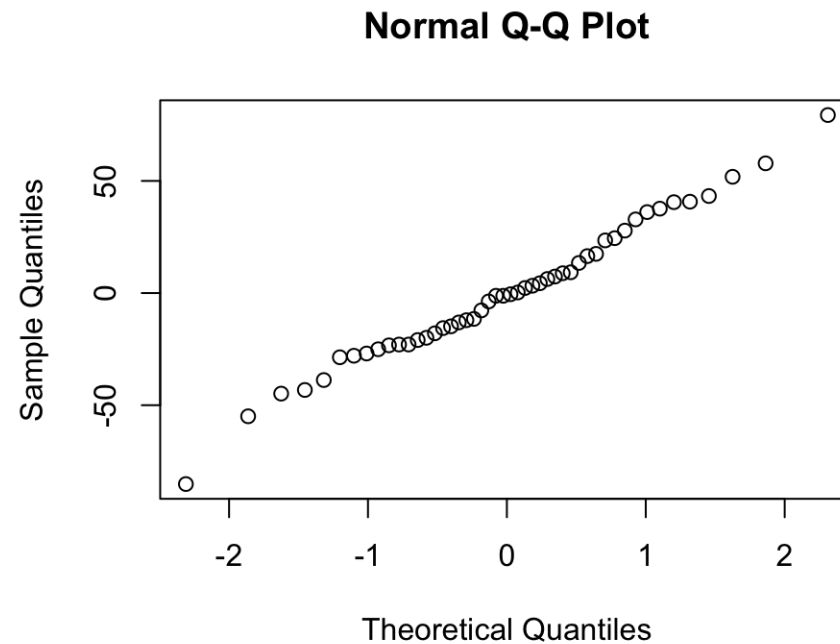
# Exemplo: Preço do diamante

Boxplot dos resíduos



# Exemplo: Preço do diamante

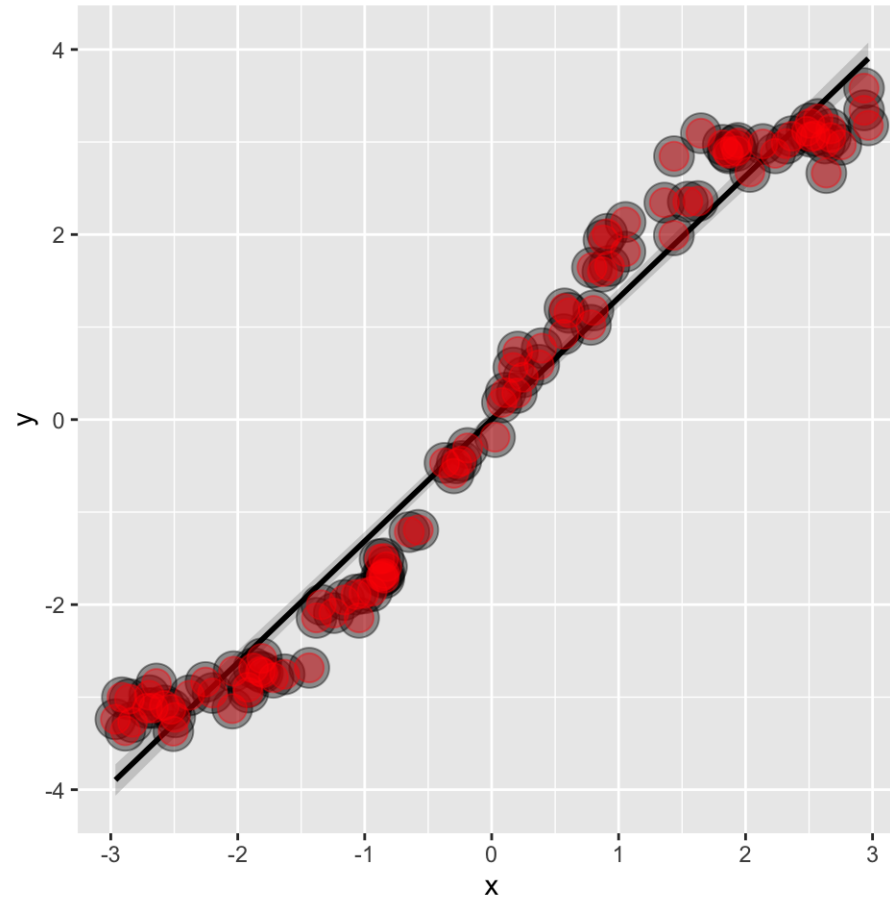
Gráfico de normalidade dos resíduos



(ver Kutner pag 110)

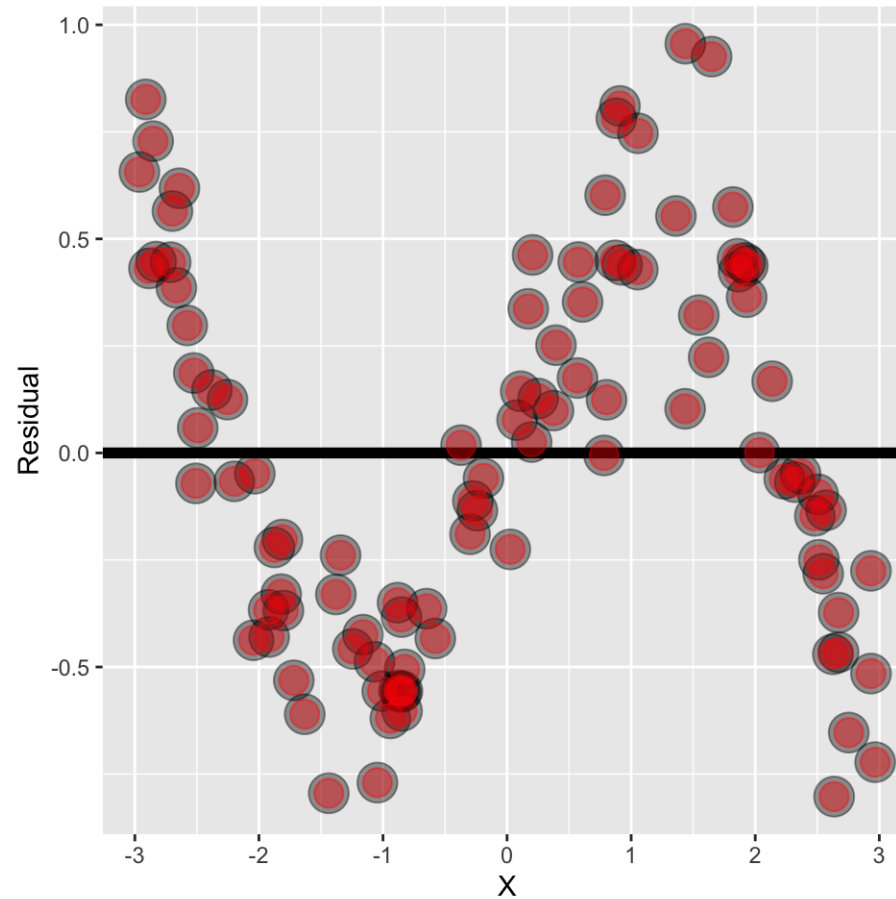
# Exemplo: não-linearidade

```
## `geom_smooth()` using formula 'y ~ x'
```



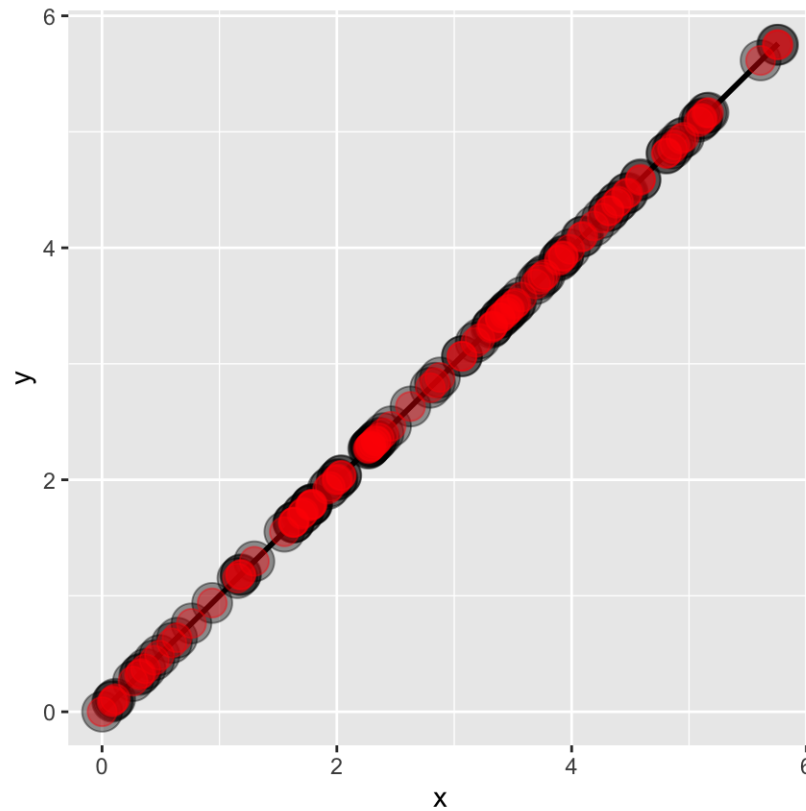
# Exemplo: não-linearidade

Resíduos X Variável preditora

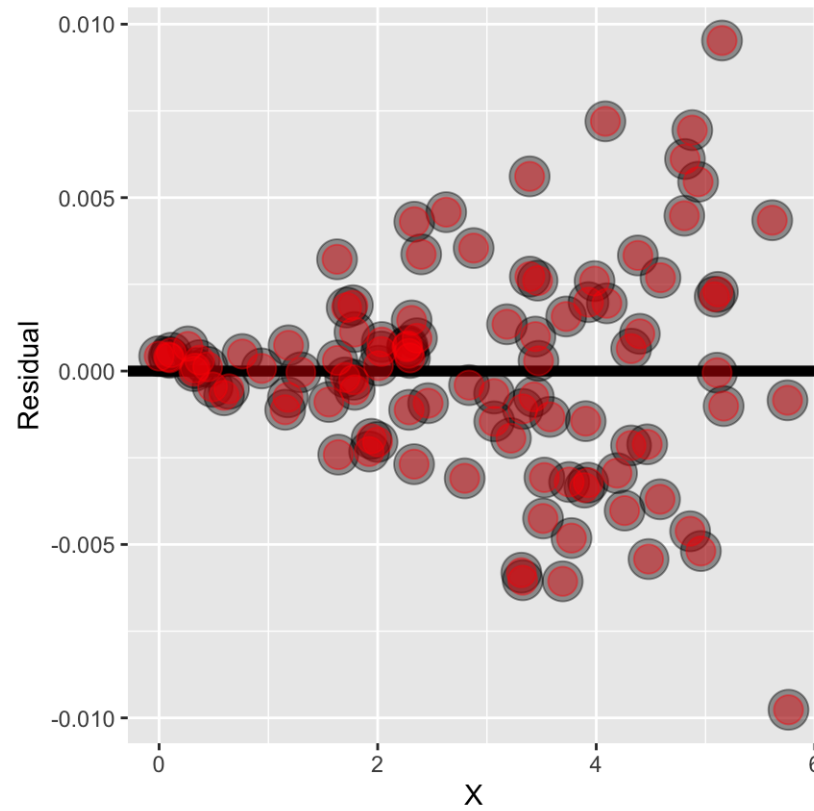


# Exemplo: heterocedasticidade

```
## `geom_smooth()` using formula 'y ~ x'
```



# Exemplo: heterocedasticidade



# Análise de Resíduos: testes de hipóteses

Além de gráficos, podemos também utilizar testes de hipóteses estatísticos para avaliar algumas suposições do modelo.



# Teste de correlação para normalidade

Calcular coeficiente de correlação entre  $e_i$  e seus valores esperados, segundo a distribuição normal

Quanto maior a correlação, maior o indício de normalidade dos resíduos.

Detalhes sobre o teste: [Looney & Gulledge \(1985\) - Use of the Correlation Coefficient with Normal Probability Plots](#)

# Teste de correlação para normalidade

Calcular a correlação entre os resíduos ordenados e

$$r = \{r_1, r_2, \dots, r_k \dots, r_n\}$$

em que  $r_k$  é o valor esperado do  $k$ -ésimo resíduos, segundo a suposição de normalidades.

$$r_k \approx \sqrt{QME} \left[ z \left( \frac{k - 0.375}{n + 0.25} \right) \right]$$

$z(A)$  é o  $A \times 100$  percentil da distribuição normal padrão.

Obs: Outro teste conhecido: [Teste de Shapiro-Wilks](#).

# Teste de homocedasticidade: Brown-Forsythe

- Não depende da suposição de normalidade
- Aplicável no caso de regressão linear simples quando a variância do erro aumenta ou decresce com a variável preditora  $X$ .
- Tamanho amostral grande o suficiente para que possamos assumir que os resíduos são independentes.
- Idéia parecida com um teste-t para duas amostras.

# Teste de homocedasticidade: Brown-Forsythe

- $X$  é dividida em  $X_1$  (valores mais baixos de  $X$ ) e  $X_2$  (valores mais altos de  $X$ ). Isto é, temos dois grupos.
- $e_{i1}$  é o  $i$ -ésimo resíduo para o grupo 1 e  $e_{i2}$  é o  $i$ -ésimo resíduo para o grupo 2.
- $n_j$  é o número de observações no grupo  $j$
- $n = n_1 + n_2$ .
- $\tilde{e}_j$  é a mediana dos resíduos do grupo  $j$ .
- $d_{i1} = |e_{i1} - \tilde{e}_1|$        $d_{i2} = |e_{i2} - \tilde{e}_2|$

# Teste de homocedasticidade: Brown-Forsythe

Estatística do teste:

$$t_{BF}^* = \frac{\bar{d}_1 - \bar{d}_2}{s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$s^2 = \frac{\sum (d_{i1} - \bar{d}_1)^2 + \sum (d_{i2} - \bar{d}_2)^2}{n - 2}$$

Se a variância é constante e  $n_1$  e  $n_2$  grandes o suficiente, então  $t_{BF}^*$  segue distribuição  $t$  com  $n - 2$  graus de liberdade.

Valores altos de  $t_{BF}^*$  indicam heterocedasticidade.

# Exemplo: Toluca

```
data <- read.table("./dados/CH01TA01.txt")
x = data[,1]
y = data[,2]
ind1 = which(x<80)
ind2 = which(x>=80)
fit = lm(y~x)
resi = fit$resi
resi1 = resi[ind1]
resi2 = resi[ind2]
d1 = abs(resi1-median(resi1))
d2 = abs(resi2-median(resi2))
t.test(d1,d2)
```

# Exemplo: Toluca

```
##  
##  Welch Two Sample t-test  
##  
## data:  d1 and d2  
## t = 1.3215, df = 22.999, p-value = 0.1993  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
##  -9.253335 41.982809  
## sample estimates:  
## mean of x mean of y  
## 44.81507 28.45034
```

Não rejeitamos  $H_0$ , isto é, não obtemos evidências para rejeitar a hipótese de homocedasticidade dos erros.

# Teste de homocedasticidade: Breusch-Pagan

Assume que  $\varepsilon_i \sim \mathcal{N}(0, \sigma_i^2)$ .

O seguinte modelo é assumido:

$$\log_e \sigma_i^2 = \gamma_0 + \gamma_1 X_i$$

Se  $\gamma_1 = 0$  temos homocedasticidade.

$H_0: \gamma_1 = 0$  vs  $H_1: \gamma_1 \neq 0$ .



# Teste de homocedasticidade: Breusch-Pagan

Estatística do teste:

$$X_{BP}^2 = \frac{\frac{SQRq}{2}}{\left(\frac{SQE}{n}\right)^2}$$

em que  $SQRq$  é a soma de quadrados da regressão de  $e_i^2$  em  $X$  e  $SQE$  é a soma de quadrados do erro da regressão de  $Y$  em  $X$ .

Sob  $H_0$ ,  $X_{BP}^2 \sim \chi_1^2$ .

# Exemplo: Toluca

```
fit = lm(y~x)
resi = fit$resi
fit2 = lm(resi^2~x)
a1 <- anova(fit)
a2 <- anova(fit2)
estatistica <- (a2$`Sum Sq`[1]/2)/((a1$`Sum Sq`[2]/length(y))^2)
```

# Exemplo: Toluca

```
## Analysis of Variance Table
```

```
##
```

```
## Response: y
```

```
##           Df Sum Sq Mean Sq F value    Pr(>F)
```

```
## x           1 252378  252378   105.88 4.449e-10 ***
```

```
## Residuals 23  54825    2384
```

```
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## Analysis of Variance Table
```

```
##
```

```
## Response: resi^2
```

```
##           Df      Sum Sq Mean Sq F value Pr(>F)
```

```
## x           1   7896142 7896142   1.0914  0.307
```

```
## Residuals 23 166395896 7234604
```

```
## [1] 0.8209192
```

# Teste de homocedasticidade: Breusch-Pagan

Diretamente pelo R:

```
library('car')
```

```
## Loading required package: carData
```

```
ncvTest(fit)
```

```
## Non-constant Variance Score Test
```

```
## Variance formula: ~ fitted.values
```

```
## Chisquare = 0.8209192, Df = 1, p = 0.36491
```

# Teste F para falta de ajuste

Queremos testar se a função de regressão proposta é adequada para os dados.

Suposições: as observações de  $Y|X$  são

- independentes
- seguem distribuição normal
- variância constante  $\sigma^2$

É preciso que tenhamos mais de uma observação para cada nível de  $X$  (replicações).

# Teste F para falta de ajuste

Suponha que tenhamos  $m$  valores distintos para  $X$

Para cada valor diferente de  $X$ ,  $x_j$ , temos  $n_j$  observações.

- $Y_{11}, Y_{12}, \dots, Y_{1n_1}$  são  $n_1$  repetições para  $X = x_1$ .
- $Y_{21}, Y_{22}, \dots, Y_{2n_2}$  são  $n_2$  repetições para  $X = x_2$ .
- $Y_{m1}, Y_{m2}, \dots, Y_{mn_m}$  são  $n_m$  repetições para  $X = x_m$ .

$$n = \sum_{j=1}^m \sum_{u=1}^{n_j} 1 = \sum_{j=1}^m n_j$$

# Teste F para falta de ajuste

Com replicações para cada  $X = x_k$ , temos um estimador para  $Var(Y|X = x_k) = \sigma^2$  para todo  $k$ .

Por exemplo, quando  $X = x_1$ , podemos usar  $Y_{11}, Y_{12}, \dots, Y_{1n_1}$  e o seguinte estimador para  $Var(Y|X = x_1)$ :

$$\frac{1}{n_1 - 1} \sum_{j=1}^{n_1} (Y_{1j} - \bar{Y}_1)^2$$

Combinando todas as observações (não apenas as replicações para um dado nível de  $X$ ), temos o seguinte estimador para  $\sigma^2$ :

$$s_p^2 = \frac{\sum_{j=1}^m \sum_{u=1}^{n_j} (Y_{ju} - \bar{Y}_j)^2}{\sum_{j=1}^m (n_j - 1)}$$

# Teste F para falta de ajuste

$$SQE_p = \sum_{j=1}^m \sum_{u=1}^{n_j} (Y_{ju} - \bar{Y}_j)^2$$

com graus de liberdade  $n_p = \sum_{j=1}^m (n_j - 1) = \sum_{j=1}^m n_j - m$ .

$$QME_{puro} = s_p^2 = \frac{\sum_{j=1}^m \sum_{u=1}^{n_j} (Y_{ju} - \bar{Y}_j)^2}{\sum_{j=1}^m n_j - m}$$



# Teste F para falta de ajuste

Relembrando:

$$SQE = \sum_{j=1}^m \sum_{u=1}^{n_j} (Y_{ju} - \hat{Y}_j)^2$$

Mostre<sup>1</sup> que a  $SQE$  pode ser particionada da seguinte forma:

$$\sum_{j=1}^m \sum_{u=1}^{n_j} (Y_{ju} - \hat{Y}_j)^2 = \sum_{j=1}^m \sum_{u=1}^{n_j} (Y_{ju} - \bar{Y}_j)^2 + \sum_{j=1}^m n_j (\hat{Y}_j - \bar{Y}_j)^2$$

# Teste F para falta de ajuste

$$\underbrace{\sum_{j=1}^m \sum_{u=1}^{n_j} (Y_{ju} - \hat{Y}_j)^2}_{SQE} = \underbrace{\sum_{j=1}^m \sum_{u=1}^{n_j} (Y_{ju} - \bar{Y}_j)^2}_{SQE_p} + \underbrace{\sum_{j=1}^m n_j (\hat{Y}_j - \bar{Y}_j)^2}_{SQF_a}$$

$SQE$  é a soma de quadrados dos resíduos, representando a variação em torno da reta.

$SQE_p$  é a soma de quadrados do erro puro, representando a variação de  $Y$ , para  $X$  fixo, independente do modelo (pois utilizamos as replicações).

$SQF_a$  é a soma de quadrados da falta de ajuste, representando a falta de ajuste do modelo.

# Teste F para falta de ajuste

$SQE$  é o numerador de  $s^2$  e  $SQE_p$  é o numerador de  $s_p^2$ .

Valores relativamente altos de  $SQF_a$  indicam discrepância entre esses dois estimadores e, conseqüentemente, indicam evidência contra a adequação do modelo de regressão linear simples.

Quanto mais alto o valor de  $SQF_a$ , pior é o ajuste do modelo, pois, se o modelo fosse perfeito, teríamos que  $SQF_a = 0$ .

# Teste F para falta de ajuste

$H_0$ : o MRLS é adequado

$H_1$ : o MRLS não é adequado

Estatística do teste:

$$F^* = \frac{SQF_a/(m-2)}{SQE_p/(n-m)} \underset{\sim}{\text{sob } H_0} F_{m-2, n-m}$$

# Teste F para falta de ajuste

Fonte de Variação	gl	SQ
Regressão	1	$SQReg = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$
Erro	$n - 2$	$SQE = \sum_{i=1}^n (Y_i - \hat{Y})^2$
(Falta de ajuste)	$m - 2$	$SQF_a = \sum_{j=1}^m n_j (\hat{Y}_j - \bar{Y}_j)^2$
(Erro puro)	$n - m$	$SQE_p = \sum_{j=1}^m \sum_{u=1}^{n_j} (Y_{ju} - \bar{Y}_j)^2$
Total (ajustada)	$n - 1$	$SQT = \sum_{i=1}^n (Y_i - \bar{Y})^2$

# Exemplo: Agências bancárias

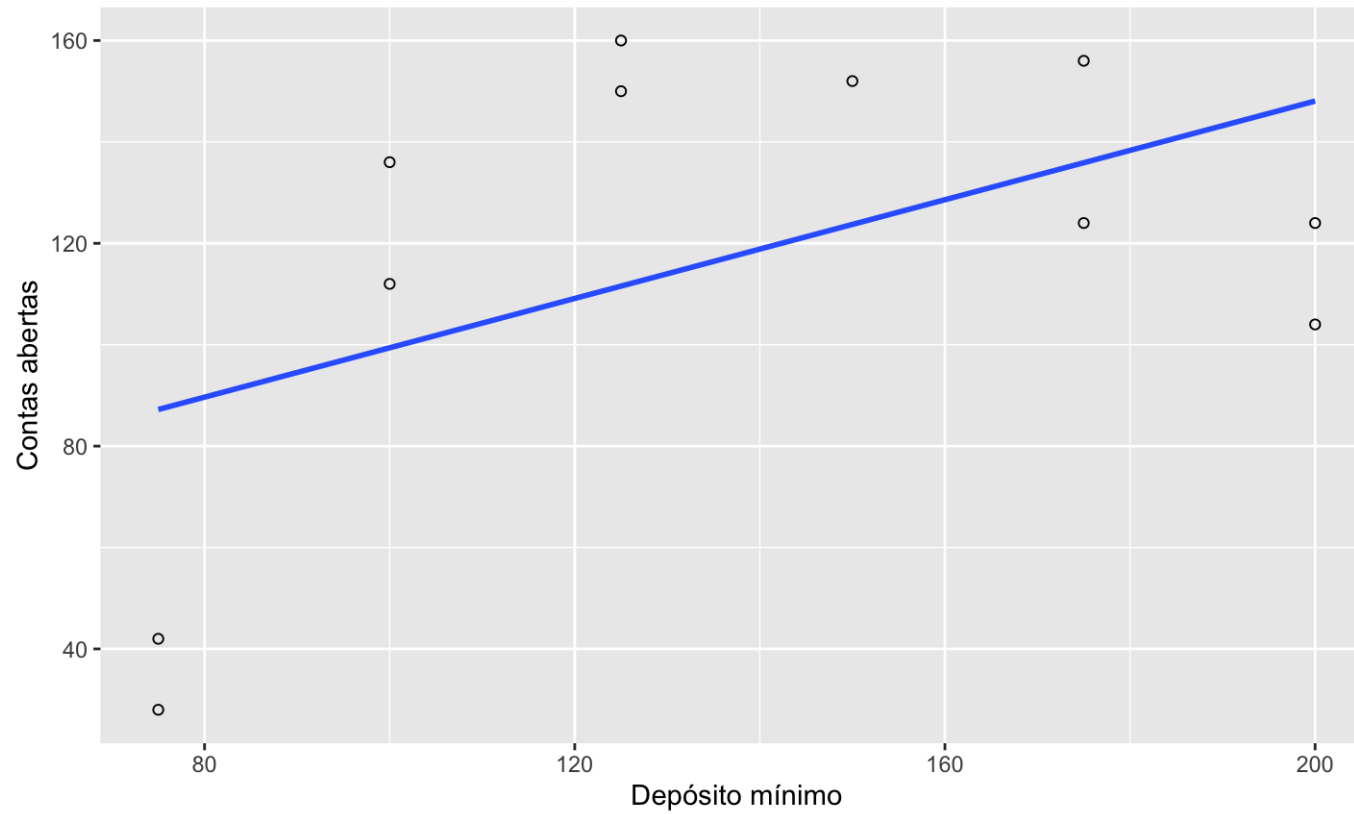
11 agências de um certo banco ofereceram brindes para novos clientes.

Um depósito inicial mínimo era necessário para qualificar para o brinde.

Cada agência especificou o valor do depósito mínimo.

Interesse: valor do depósito mínimo inicial e número de contas que foram abertas na agência.

# Exemplo: Agências bancárias



# Exemplo: Agências bancárias

Modelo completo:

$$Y_{ij} = \mu_j + \varepsilon_{ij}$$

Temos que  $E(Y_{ij}) = \mu_j$ , isto é, temos esperança diferente para cada  $X_j$ .



# Exemplo: Agências bancárias

```
Full=lm(y~as.factor(x)-1)
summary(Full)
```

```
##
## Call:
## lm(formula = y ~ as.factor(x) - 1)
##
## Residuals:
##      1      2      3      4      5      6      7      8      9     10     11
##      5    -12     10     -7      0     16      7    -16     -5    -10     12
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## as.factor(x)75      35.00      10.71   3.267 0.022282 *
## as.factor(x)100     124.00      10.71  11.573 8.45e-05 ***
## as.factor(x)125     155.00      10.71  14.466 2.85e-05 ***
## as.factor(x)150     152.00      15.15  10.031 0.000168 ***
## as.factor(x)175     140.00      10.71  13.066 4.68e-05 ***
## as.factor(x)200     114.00      10.71  10.640 0.000127 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.15 on 5 degrees of freedom
## Multiple R-squared:  0.9933, Adjusted R-squared:  0.9852
## F-statistic: 123.1 on 6 and 5 DF,  p-value: 2.893e-05
```

# Exemplo: Agências bancárias

Modelo reduzido: reta de regressão.

```
Reduced=lm(y~x)
summary(Reduced)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -59.23 -34.06  12.61  32.44  48.44
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  50.7225     39.3979   1.287    0.23
## x            0.4867      0.2747   1.772    0.11
##
## Residual standard error: 40.47 on 9 degrees of freedom
## Multiple R-squared:  0.2586, Adjusted R-squared:  0.1762
## F-statistic: 3.139 on 1 and 9 DF,  p-value: 0.1102
```

$s = 40.47$  enquanto que, usando o modelo anterior, temos que  $s_e = 15.15$ .

# Exemplo: Agências bancárias

```
Reduced=lm(y~x)
Full=lm(y~as.factor(x)-1)
anova(Reduced, Full)
```

```
## Analysis of Variance Table
##
## Model 1: y ~ x
## Model 2: y ~ as.factor(x) - 1
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1         9 14742
## 2         5  1148  4      13594 14.801 0.005594 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Estatística do teste:

$$F^* = \frac{SQF_a/(m-2)}{SQE_p/(n-m)} \underset{\sim}{\text{sob } H_0} F_{m-2, n-m} = F_{4,5}$$

# Exemplo: Agências bancárias

$$F_{obs}^* = \frac{[SQE(Reduced) - SQE(Full)]/4}{SQE(Full)/5} = \frac{1.3593571 \times 10^4/4}{1148/5} = 14.8$$

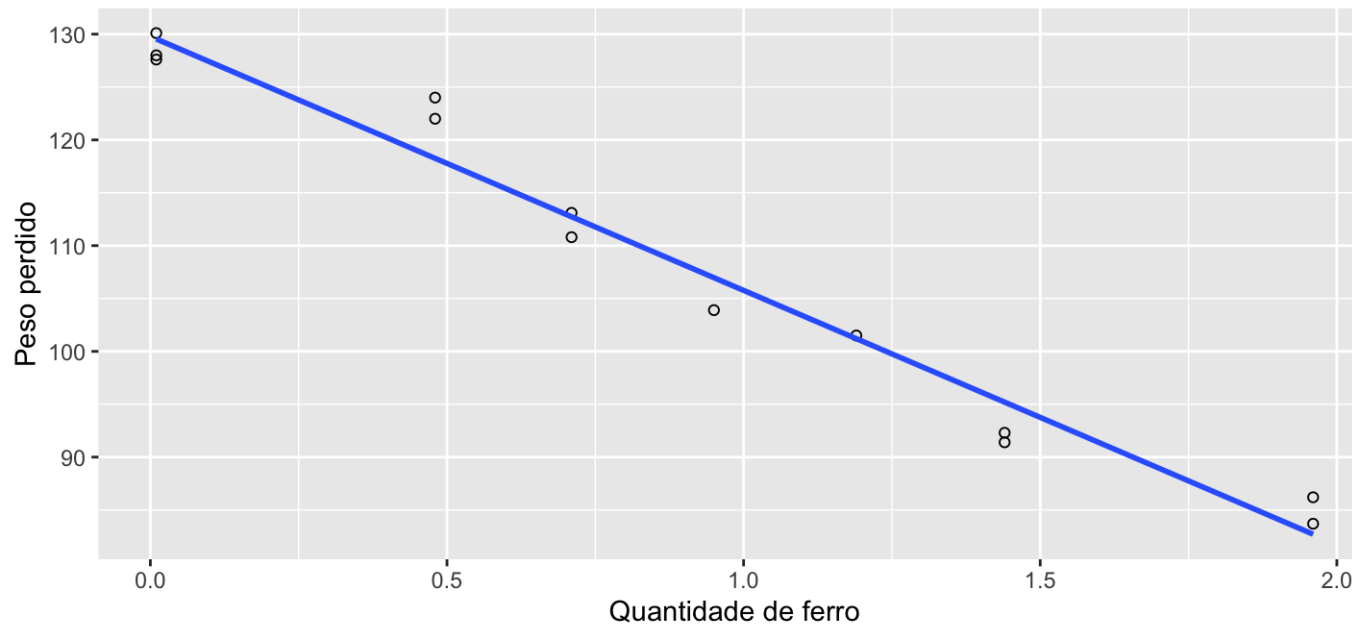
valor de p: 0.0055938.

Valor crítico para  $\alpha = 0.05$ : 5.19

# Exemplo: Ligas de cobre e níquel

13 ligas de cobre e níquel, mas quantidades diferentes de ferro.

As ligas foram submersas em água do mar por 60 dias e a perda de peso devido à corrosão foi anotada em miligramas por decímetro quadrado por dia.



# Exemplo: Ligas de cobre e níquel

```
##  
## Call:  
## lm(formula = loss ~ Fe, data = corrosion)  
##  
## Residuals:  
##      Min       1Q   Median       3Q      Max   
## -3.7980 -1.9464  0.2971  0.9924  5.7429   
##  
## Coefficients:  
##              Estimate Std. Error t value Pr(>|t|)      
## (Intercept)  129.787      1.403   92.52  < 2e-16 ***  
## Fe          -24.020      1.280  -18.77 1.06e-09 ***  
## ---  
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
##  
## Residual standard error: 3.058 on 11 degrees of freedom  
## Multiple R-squared:  0.9697, Adjusted R-squared:  0.967  
## F-statistic: 352.3 on 1 and 11 DF,  p-value: 1.055e-09
```

# Exemplo: Ligas de cobre e níquel

```
##
## Call:
## lm(formula = loss ~ factor(Fe) - 1, data = corrosion)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.2500 -0.9667  0.0000  1.0000  1.5333
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## factor(Fe)0.01 128.5667     0.8090  158.91 4.19e-12 ***
## factor(Fe)0.48 123.0000     0.9909  124.13 1.84e-11 ***
## factor(Fe)0.71 111.9500     0.9909  112.98 3.24e-11 ***
## factor(Fe)0.95 103.9000     1.4013   74.15 4.05e-10 ***
## factor(Fe)1.19 101.5000     1.4013   72.43 4.66e-10 ***
## factor(Fe)1.44  91.8500     0.9909   92.70 1.06e-10 ***
## factor(Fe)1.96  84.9500     0.9909   85.73 1.70e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.401 on 6 degrees of freedom
## Multiple R-squared:  0.9999, Adjusted R-squared:  0.9998
## F-statistic: 1.145e+04 on 7 and 6 DF,  p-value: 6.062e-12
```

$s_e = 1.4$  enquanto que, usando o modelo anterior, temos que  $s = 3.06$ .

# Exemplo: Ligas de cobre e níquel

```
anova(modeloreduzido,modelocompleto)
```

```
## Analysis of Variance Table
##
## Model 1: loss ~ Fe
## Model 2: loss ~ factor(Fe) - 1
##   Res.Df    RSS Df Sum of Sq    F    Pr(>F)
## 1      11 102.850
## 2       6  11.782  5    91.069 9.2756 0.008623 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Estatística do teste:

$$F^* = \frac{SQF_a/(m-2)}{SQE_p/(n-m)} = \frac{91.069/5}{11.782/6} = 9.28$$



# Exemplo: Ligas de cobre e níquel

```
library(alr3)
pureErrorAnova(modeloreduzido)
```

```
## Analysis of Variance Table
```

```
##
```

```
## Response: loss
```

##		Df	Sum Sq	Mean Sq	F value	Pr(>F)	
##	Fe	1	3293.8	3293.8	1677.4028	1.417e-08	***
##	Residuals	11	102.9	9.4			
##	Lack of fit	5	91.1	18.2	9.2756	0.008623	**
##	Pure Error	6	11.8	2.0			

```
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

# Como melhorar o modelo

Se nos testes e/ou gráficos de diagnóstico do modelo determinarmos que o mesmo não é adequado, o que fazer?

- Abandonar o modelo e procurar outro mais adequado: regressão logística, regressão não-paramétrica, regressão robusta...
- Aplicar alguma transformação nos dados, de maneira que a regressão linear simples fique adequada.

# Agradecimento

- Slides criados por Samara F Kiihl / IMECC / UNICAMP

# Leitura

- Applied Linear Statistical Models: 3.1-3.7.
- Caffo - [Regression Models for Data Science in R](#): Residuals.
- Draper & Smith - [Applied Regression Analysis](#): Capítulo 2.
- Weisberg - [Applied Linear Regression](#): Capítulo 2.
- Faraway - [Linear Models with R](#): 8.3.

