MAD-CB

Regressão Linear

Regressão – Historia

- Termo vem de eugenismo “eugenics” de Sir Francis Galton.
- Estudou alturas de famílias
  - Observou que crianças de pais altos tendiam de ser mais baixo de que os pais e crianças de pais baixos tendiam de ser mais altos - Chamou a tendência “regressão à média”
- Usaremos esses dados clássicos

Método de Mínimos Quadrados

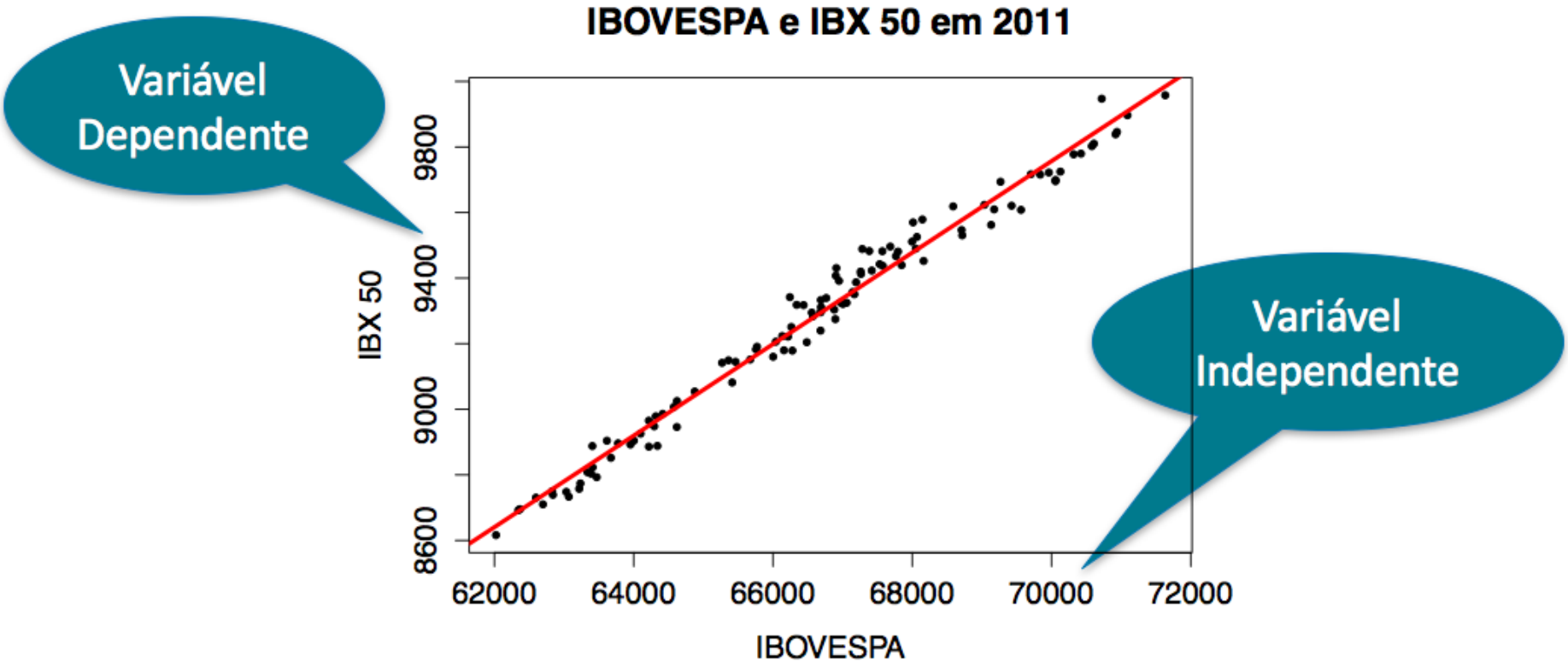
- Solucionamos com o método *Mínimos Quadrados*
  - Inventado por Carl Friedrich Gauss
  - Método minimiza as divergências entre os valores lineares previstos e os valores dos dados
  - Consegue o melhor relação entre a variável de resultado e as variáveis prognosticas
- Por enquanto, vamos restringir o modelo para forma linear
  - Outras formas existem

Proposito

Prever um resulatdo numa variável dependente baseado em uma ou mais variáveis independentes

- Uma – regressão linear *simples*
- Mais – regressão linear *múltipla*

Visualização de Regressão



Linha Reta

$$y = \beta_1 x + \beta_0$$

- β_1 = inclinação da linha (*slope*)
- β_0 = intercepto (onde cruza o eixo y)
- Os dois parâmetros da regressão
- Com estes parâmetros, Mínimos Quadrados acha a reta que melhor prevê o valor da variável dependente dado o valor de independente

“Melhor” Quer Dizer “Bom”?

- Apesar de ser a melhor maneira de prever y, possível que não descreve bem y
- **Bom** depende dos dados
- **Melhor** depende do método

Equação de Regressão

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

- Y_i = valor de variável dependente
- β_0 = intercepto
- β_1 = inclinação da reta de regressão
- X_i = valor da variável independente
- ϵ_i = termo de erro de cada caso

Equação de Regressão - Estimação

$$\hat{Y}_i = b_0 + b_1 X_i + e_i$$

- $\hat{Y}_i$ = valor de variável dependente
- b_0 = intercepto
- b_1 = inclinação da reta de regressão
- X_i = valor da variável independente
- e_i = termo de erro de cada caso

Termo de Erro (ϵ)

- Também chamado **resíduo**
- Responsável pela variabilidade em y que a reta não consegue explicar

Mínimos Quadrados

- Faz o cálculo que minimiza o quadrado da soma dos erros
- Erros = resíduos = diferenças entre o valor *observado* e o valor *esperado*

$$min \sum (y_i - \hat{y}_i)^2$$

- y_i = valor observado da variável dependente
- $\hat{y}_i$ = valor estimado da variável dependente

Basta de Teoria – Exemplo

- A base de dados de Galton sobre altura nas famílias
- Pergunta é se filhos/as são mais altos ou mais baixos de que os pais
- Mediu 898 filhos/as em 197 famílias
- Base de dados originais (em papel) fica na University College, London (UCL)

Variáveis

```
## 'data.frame':    898 obs. of  6 variables:
##  $ family: Factor w/ 197 levels "1","10","100",..: 1 1 1 1 108 108 108 108 123 1
##  $ father: num  78.5 78.5 78.5 78.5 75.5 75.5 75.5 75.5 75 75 ...
##  $ mother: num  67 67 67 67 66.5 66.5 66.5 66.5 64 64 ...
##  $ sex   : Factor w/ 2 levels "F","M": 2 1 1 1 2 2 1 1 2 1 ...
##  $ height: num  73.2 69.2 69 69 73.5 72.5 65.5 65.5 71 68 ...
##  $ nkids : int  4 4 4 4 4 4 4 4 2 2 ...
```

- `height`, `father`, `mother` todos medem altura em polegadas

Foco em Pais e Filhos

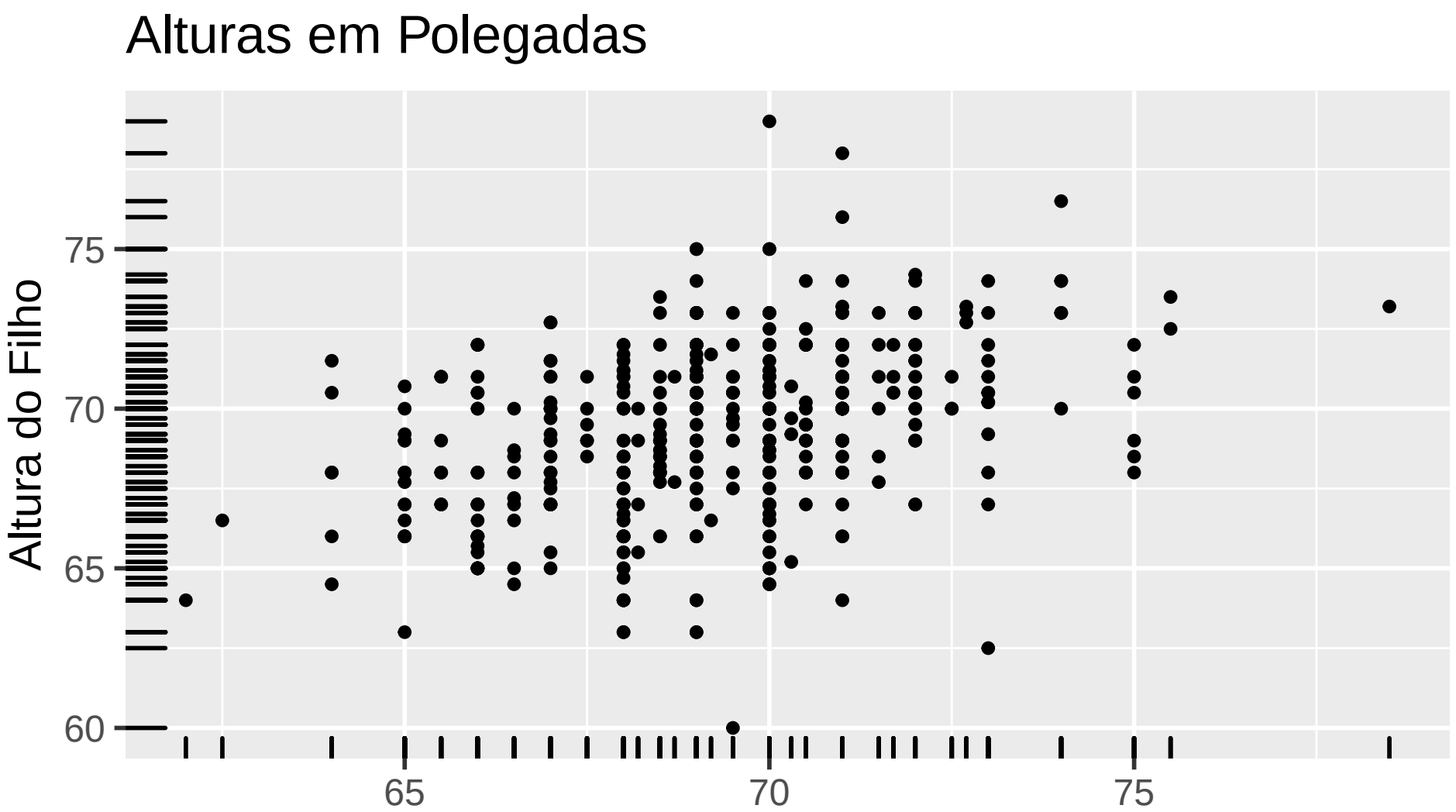
```
boys <- Galton %>% filter(sex == "M") %>% select(-family, -mother, -sex, -nkids)
glimpse(boys)
```

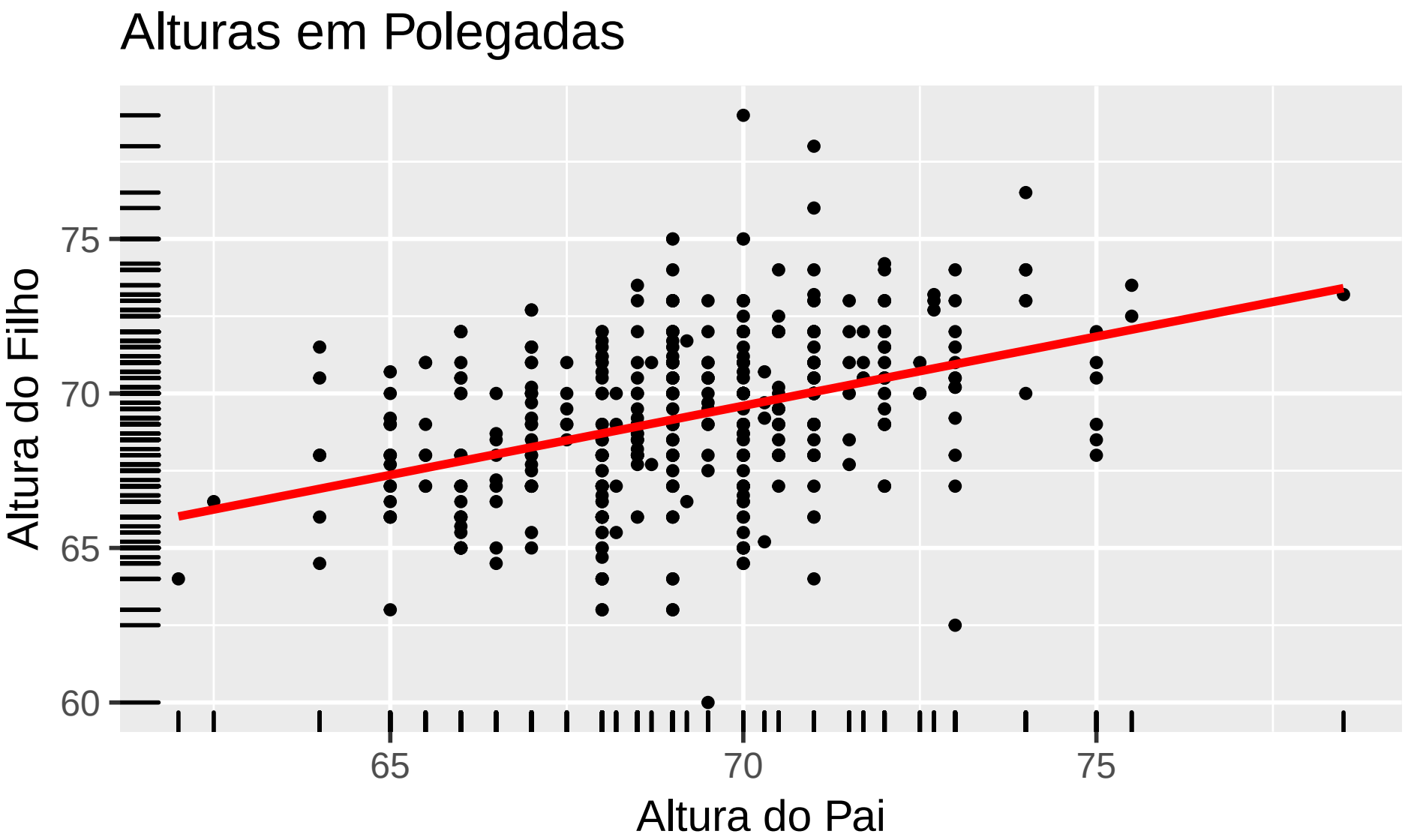
```
## Observations: 465
## Variables: 2
## $ father <dbl> 78.5, 75.5, 75.5, 75.0, 75.0, 75.0, 75.0, 75.0, 75.0, 7...
## $ height <dbl> 73.2, 73.5, 72.5, 71.0, 70.5, 68.5, 72.0, 69.0, 68.0, 7...
```

- `father` é a variável independente
- `height` é a variável dependente
- Queremos ver se a altura do pai prevê a altura do filho

Pai/Filho – Gráfico de Dispersão

```
grpf <- ggplot(data = boys, aes(x = father, y = height)) + geom_point(shape = 20) +
grpf <- grpf + labs(x = "Altura do Pai", y = "Altura do Filho", title = "Alturas em
grpf
```




Alturas em Polegadas
Altura do Filho
75
70
65
60
65
70
75
Altura do Pai

O Que Podemos Dizer Agora?

- **Parece** que mais altos os pais, mais altos os filhos
- Vamos olhar nas estatísticas descritivas das 2 variáveis
  - mais correlação

```
##        vars   n  mean   sd median trimmed  mad min  max range  skew
## father    1 465 69.17 2.30   69.0   69.16 1.93  62 78.5  16.5  0.11
## height    2 465 69.23 2.63   69.2   69.25 2.67  60 79.0  19.0 -0.03
##        kurtosis   se
## father     0.55 0.11
## height     0.29 0.12
```

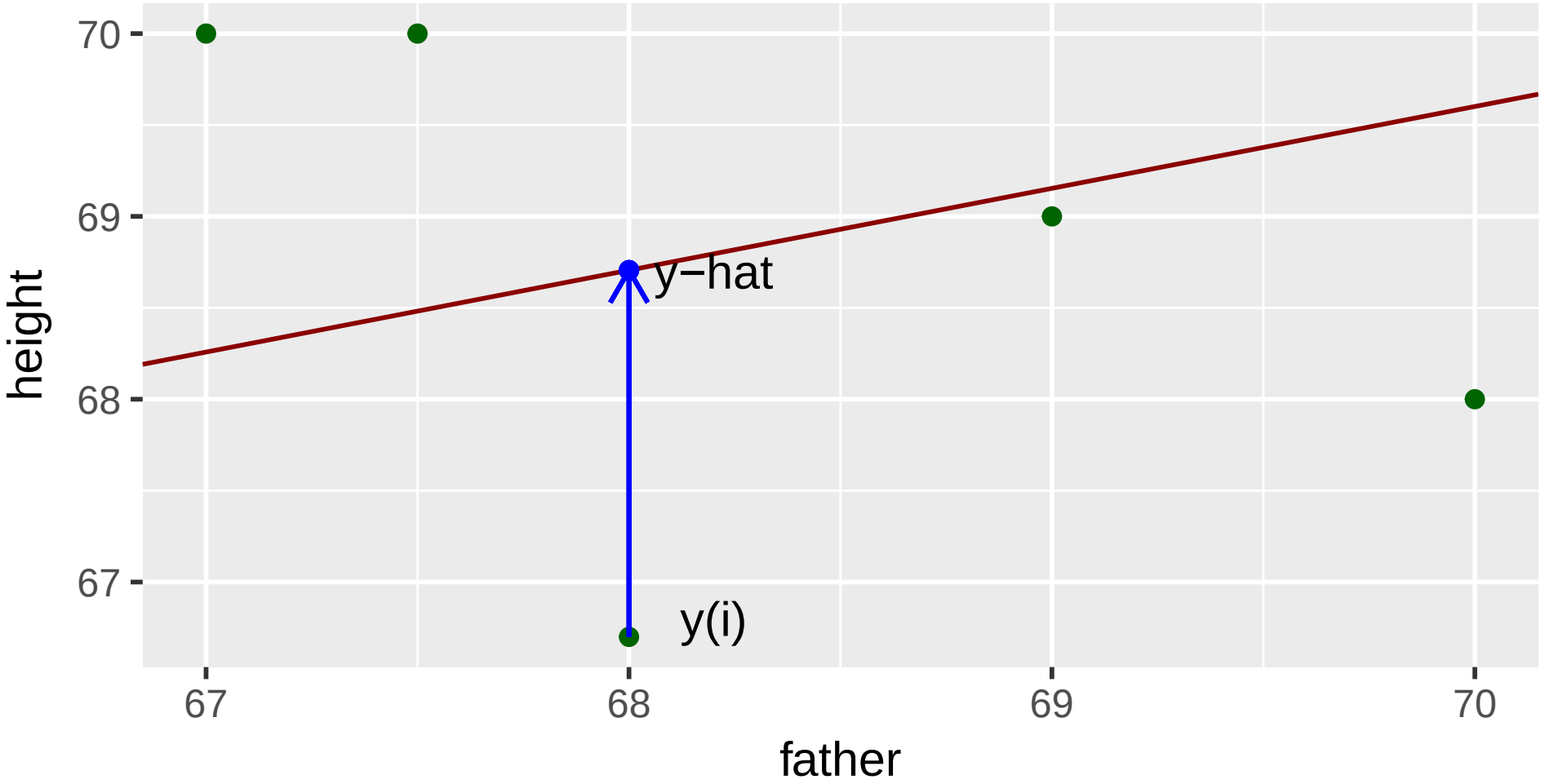
```
## [1] "Coeficiente de Correlação: 0.391"
```

O Que É a "Correlação"?

- *Coeficiente de Correlação* mede o grau da associação linear entre 2 variáveis
- Sempre cai entre -1 e $+1$
  - -1 significa uma relação perfeitamente inversa (quando x sobe, y desce pela mesma proporção)
  - 0 significa que não existe uma relação linear entre as 2 variáveis
  - $+1$ significa uma relação perfeitamente positiva (quando x sobe, y sobe pela mesma proporção)
- V.S.S: quando tem correlação positiva, tem inclinação da linha de tendência positiva, e vice versa

Para Calcular a Linha de Regressão – O Que Queremos?

- Uma linha que minimiza a diferença entre y_i e $\hat{y}$
- Precisamos trabalhar com o quadrado da diferença
  - para não ter uma soma de 0

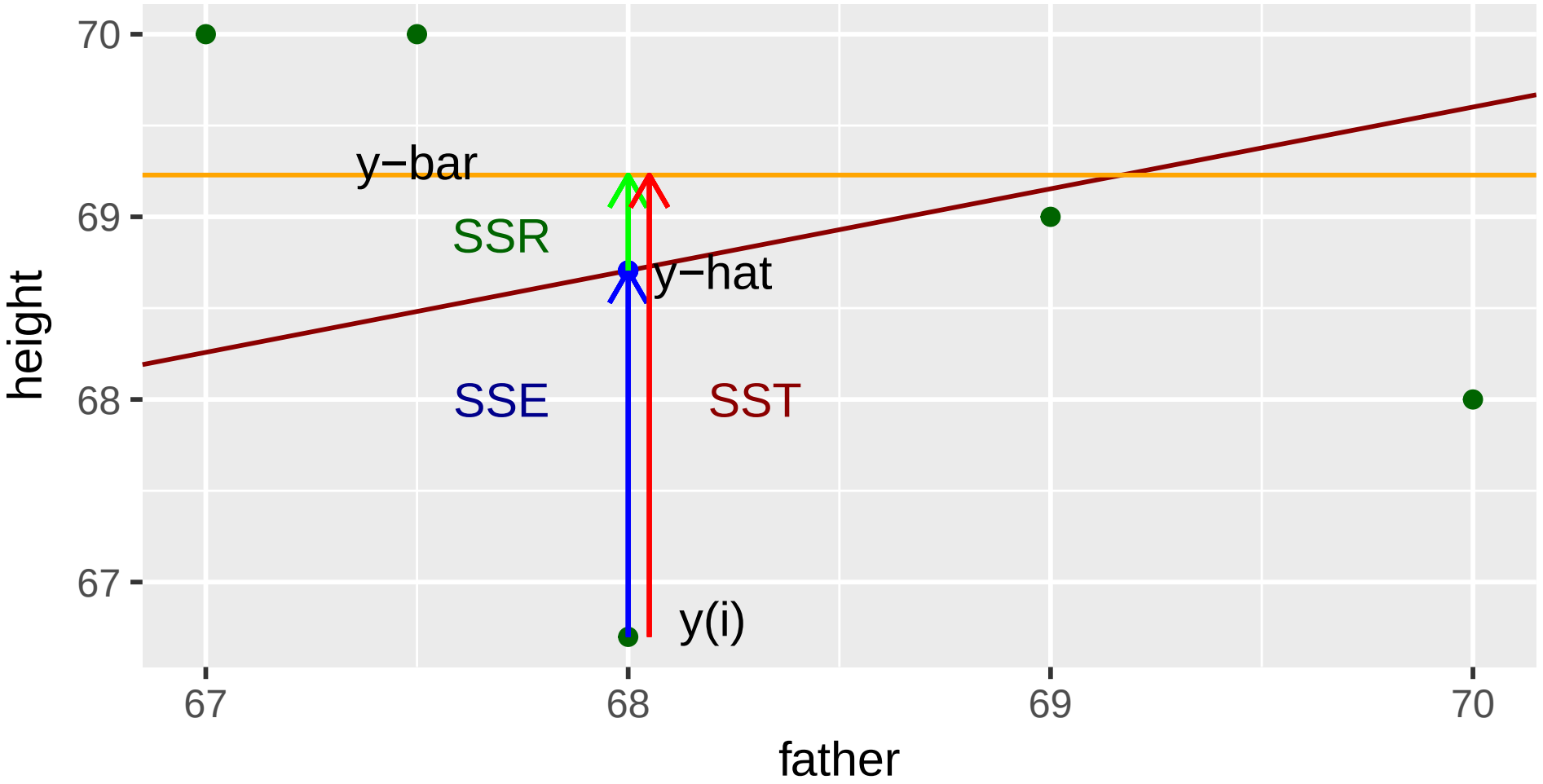


SSE – Um Componente do Soma de Quadrados (SST)

- $SST = SSE + SSR$
- SST – Total
- SSE – Relacionados aos Erros/Resíduos
- SSR – Relacionados/Explicados pela regressão

SST – O Que Representa?

- A variância total é a diferença entre o valor do modelo para cada valor de X e a média dos valores da variável dependente ($\hat{y}$)



Soma dos Quadrados

- Referimos a esse soma dos quadrados que queremos minimizar como **SSE**
  - Error sum of squares
- SSE como componente da soma dos quadrados total
  - SSE — soma dos quadrados relacionados ao resíduo
  - SSR — soma dos quadrados relacionados a regressão
- Expressão de SSE

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y})^2$$

$$SSE = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

Para Determinar a Formula para β_0 e β_1

- Para minimizar a SSE (determinar a linha mais eficiente), precisamos usar cálculo
- Fazer a derivativo parcial com respeito a β_0 e β_1

$$\frac{\partial}{\partial \beta_0} SSE = \frac{\partial}{\partial \beta_1} SSE = 0$$

- Chamadas as equações normais
- Confiamos nos softwares para calcular os parâmetros da equação

Função em R

- Função `lm` ("linear model")
- `lm(formula, data, subset, weights, na.action, method = "qr", model = TRUE, x = FALSE, y = FALSE, qr = TRUE, singular.ok = TRUE, contrasts = NULL, offset, ...)`
- Os importantes são `formula`, `data`, `subset`, `weights`, `na.action`
- `formula`: onde mostra quais variáveis você está modelando
  - Variável dependente vem primeiro
  - Separada da independente(s) por " ~ "
  - Para os `boys`: `height ~ father`
  - `data`: data frame ou tibble que contem as variáveis
  - `subset`, `weights`: parâmetros que permitem que você customizar tratamento das variáveis
  - `na.action`: como vai tratar os dados missing na base de dados

Função Aplicada aos Pais e Filhos

- Função `lm` produz uma lista de 12 itens em um formato especial

```
fit1 <- lm(height ~ father, data = boys)
summary(fit1)
```

```
##
## Call:
## lm(formula = height ~ father, data = boys)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.3774 -1.4968  0.0181  1.6375  9.3987
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 38.25891    3.38663   11.30   <2e-16 ***
## father       0.44775    0.04894    9.15   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.424 on 463 degrees of freedom
## Multiple R-squared:  0.1531, Adjusted R-squared:  0.1513
## F-statistic: 83.72 on 1 and 463 DF,  p-value: < 2.2e-16
```

O Que Diz o Modelo

$$\hat{y} = 38.259 + 0.448x$$

- Se o pai tivesse 0 altura, o filho teria 38.259 polegadas de altura
  - Não faz sentido prático, mas estabelece a base para calculo de altura
  - Para cada polegada incremental da altura do pai, o filho seria 0.448 polegadas mais alto

Extrair os Valores dos Coeficientes

1. Usar `broom::tidy`

```
broom::tidy(fit1) %>% kable()
```

term	estimate	std.error	statistic	p.value
(Intercept)	38.2589122	3.3866340	11.297032	0
father	0.4477479	0.0489353	9.149788	0

2. Usar `coef`

```
coef(fit1)
```

```
## (Intercept)      father
##  38.2589122   0.4477479
```

Previsões de Novos Valores

- Pode usar o modelo para prever novos valores da altura dos filhos
- Usar `broom::augment`

```
fit1 %>% broom::augment(newdata = data_frame(father = 72))
```

```
##   father  .fitted   .se.fit
## 1     72 70.49676 0.1784466
```

O Que Significa o Modelo? Como Interpretar Ele?

Existe Relação Entre Variáveis Independente e Dependentes?

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

- Se β_1 (inclinação da linha) for 0, o que seria a equação?

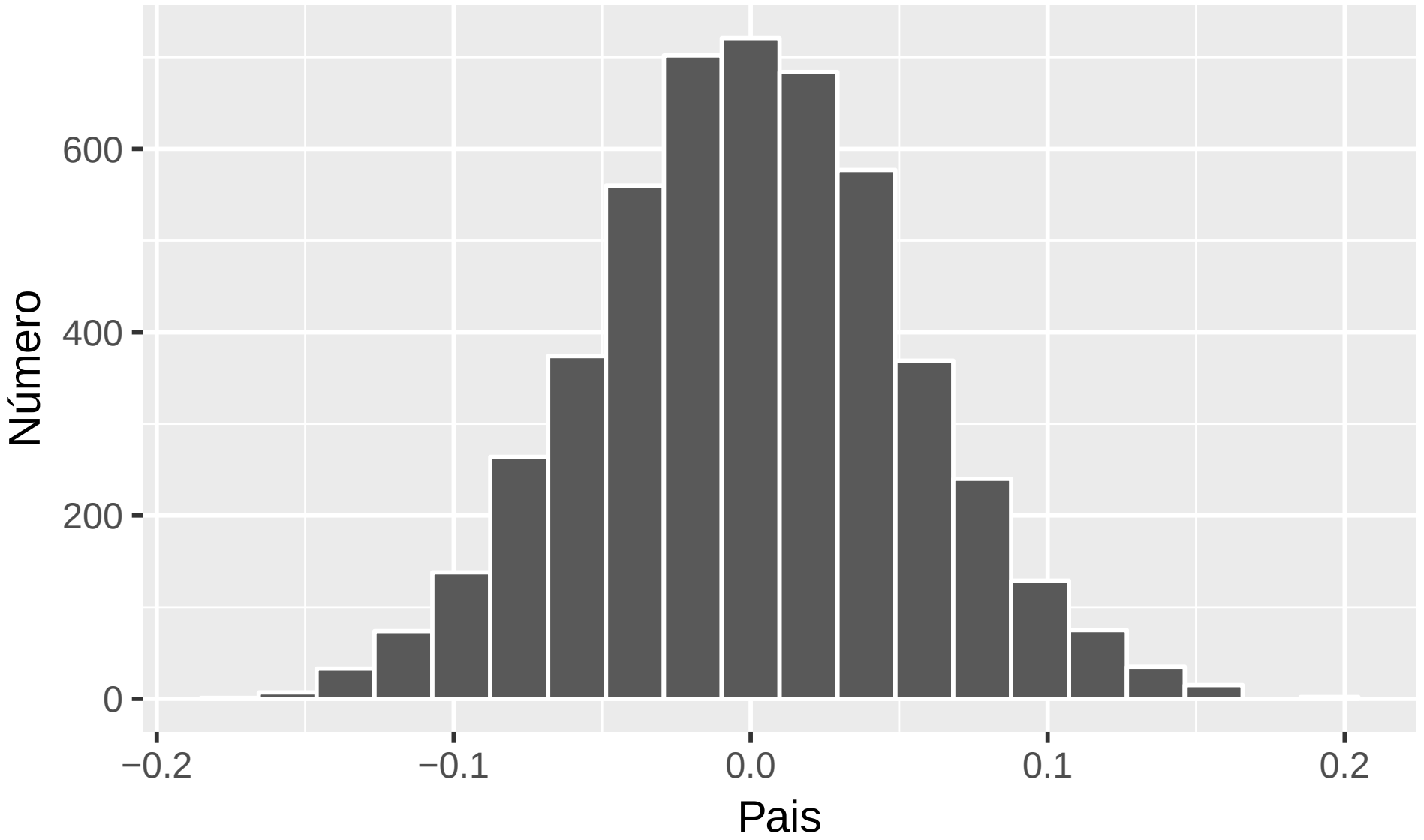
$$Y_i = \beta_0 + \epsilon_i$$

- X desaparece
- Relação entre Y e X não existe
  - Só tem intercepto e erro
- Faz possível teste eficiente de existência ou não de uma relação entre X e Y
- Cria uma hipótese nula de $H_0 : \beta_1 = 0$

Teste de Hipótese Nula

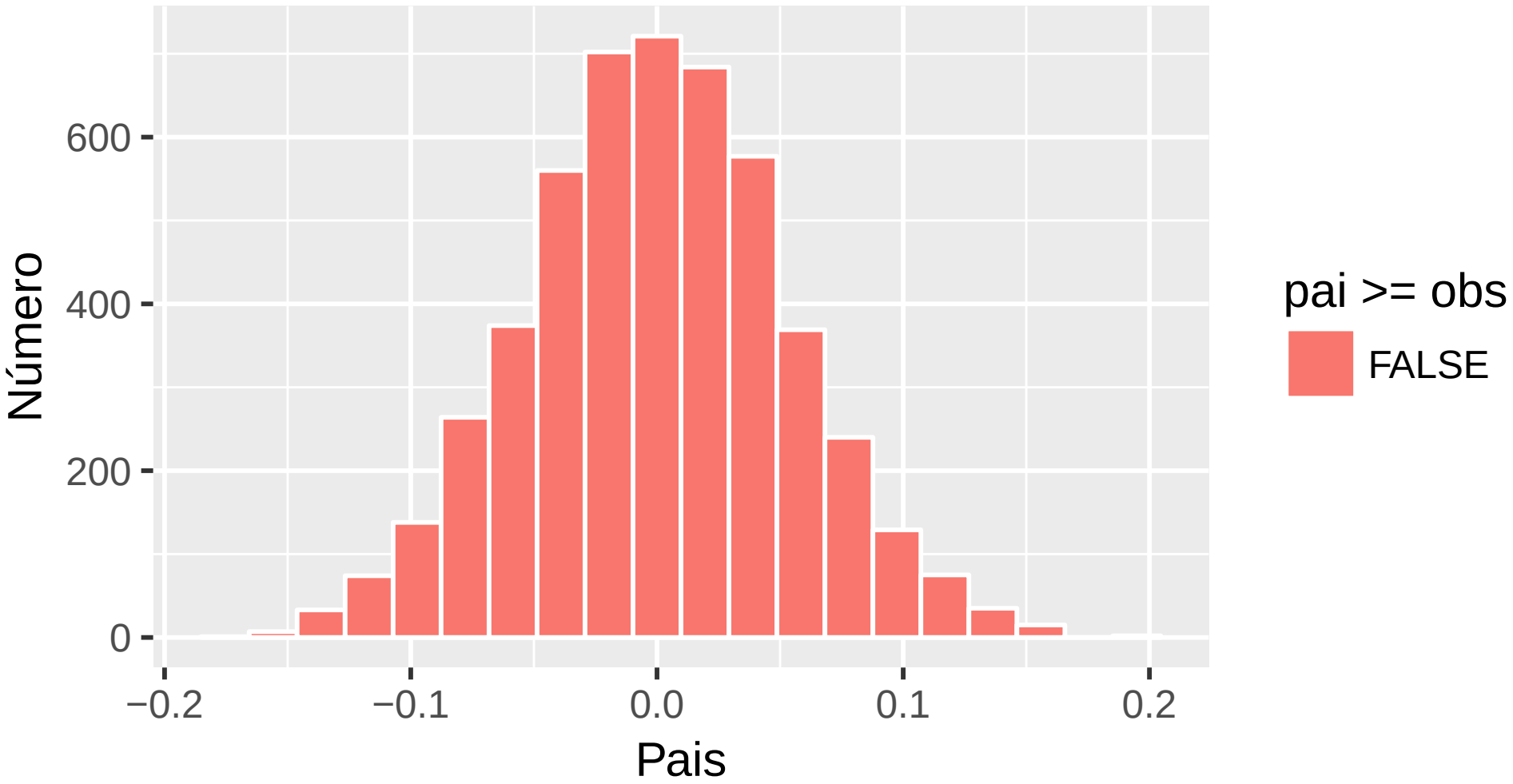
- Vamos fazer uma simulação de hipótese nula
- Se a nula é correta, qualquer altura do filho podia ter ocorrido com qualquer altura do pai.
- Podemos calcular o modelo de regressão 5.000 vezes com valores de todo a base de alturas dos filhos
- Como resultado, vamos focar nos valores da inclinação, β_1
- Depois, nós vamos comparar nosso valor de β_1 observado e ver onde cai na distribuição dos valores simulados

Histograma dos Modelos



Histograma com Valores Abaixo/Acima do Valor da Amostra

```
## [1] "Número de simulações com beta1 >= obs: 0"
```



O Valor-p da Inclinação (β_1)

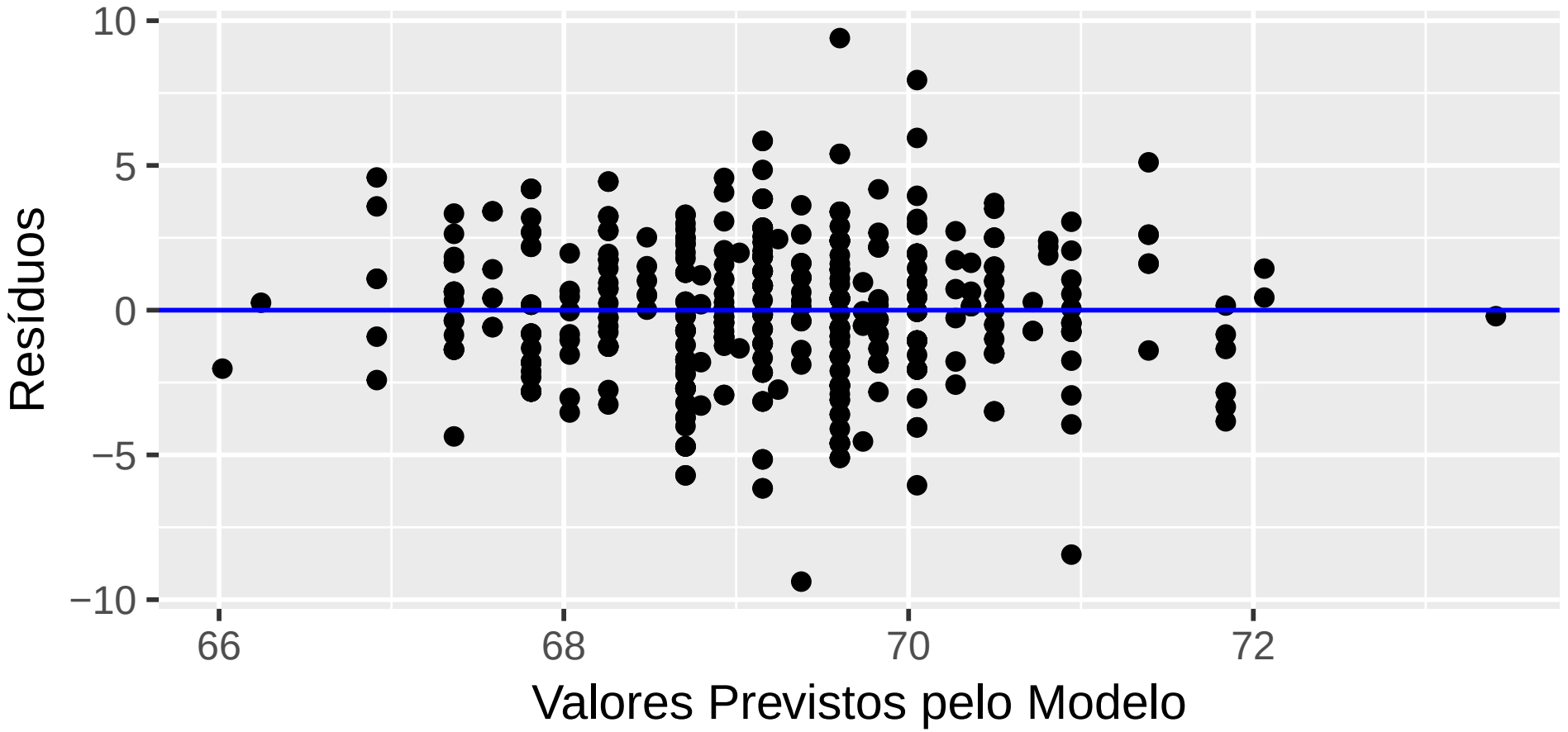
- Porque **nenhuma** das simulações produziu um valor superior ao observado (0.448)
  - Pode concluir que o valor-p deste teste é 0.
  - Não parece existir nenhuma chance que a inclinação $= 0$
- Assim, rejeitamos a hipótese nula e concluir que uma relação linear entre as alturas dos pais e filhos realmente existe.

Premissas de Regressão Linear

1. Todas as variáveis devem ter a mesma variância
   - Gráfico de resíduo deve evitar padrões indo de esquerda até direta
2. Todas as observações, resíduos e variáveis independentes: todos devem ser independentes
   - Gráfico de resíduo não deve mostrar um padrão sinuoso
3. Resíduos têm uma distribuição perto a normal
   - Gráfico "qq" dos resíduos padronizados

Gráfico de Resíduos

- Gráfico que mostra o valor previsto pelo modelo (“fitted value”) vs. o resíduo
- Uso da função `broom::augment()`
  - Eficiente para extrair os valores utilizados nos testes dos modelos



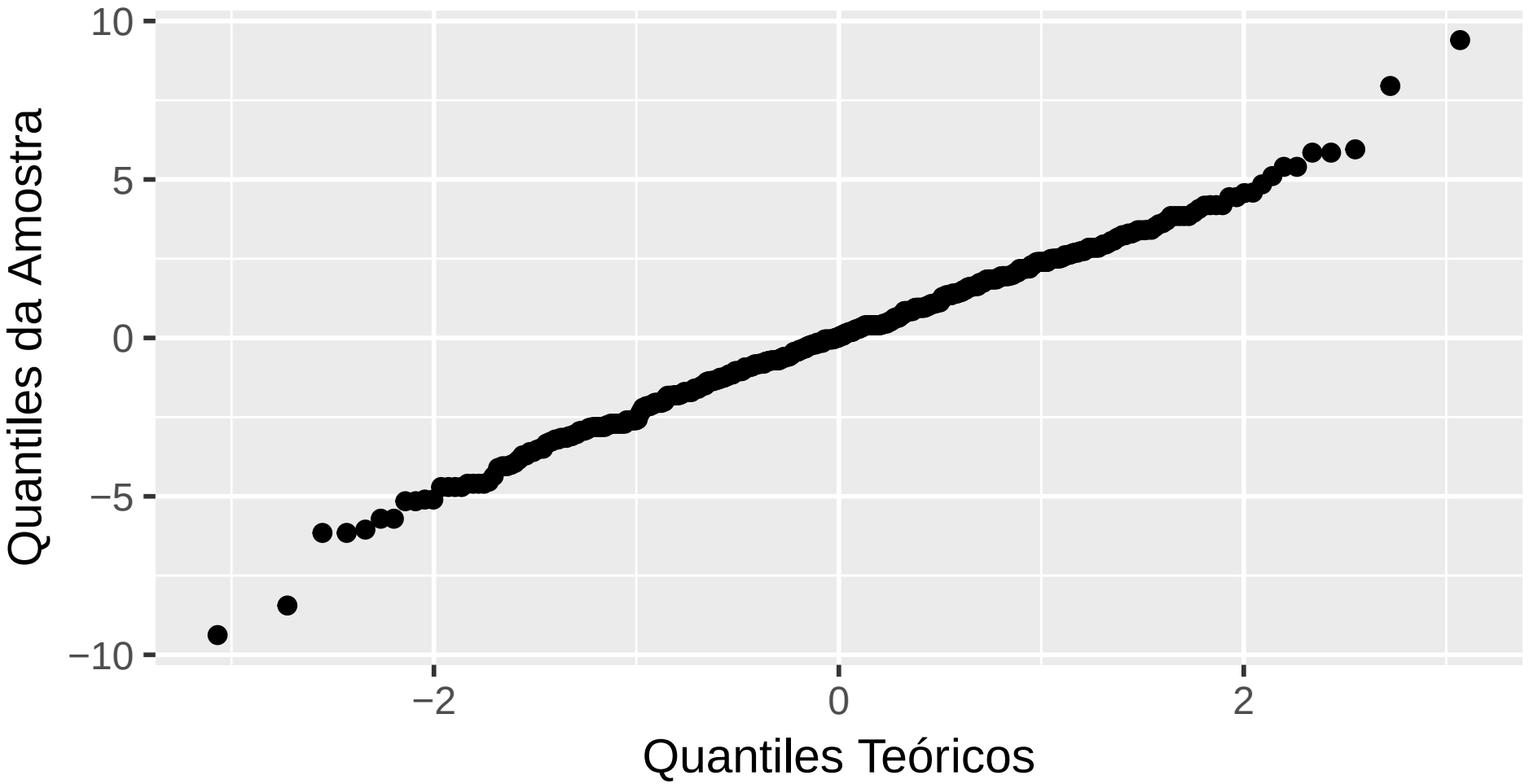
Importância dos Resíduos

- Pode usar os erros/resíduos para verificar se as premissas da regressão foram respeitadas
- Não devem mostrar um padrão linear

Gráfico Q-Q

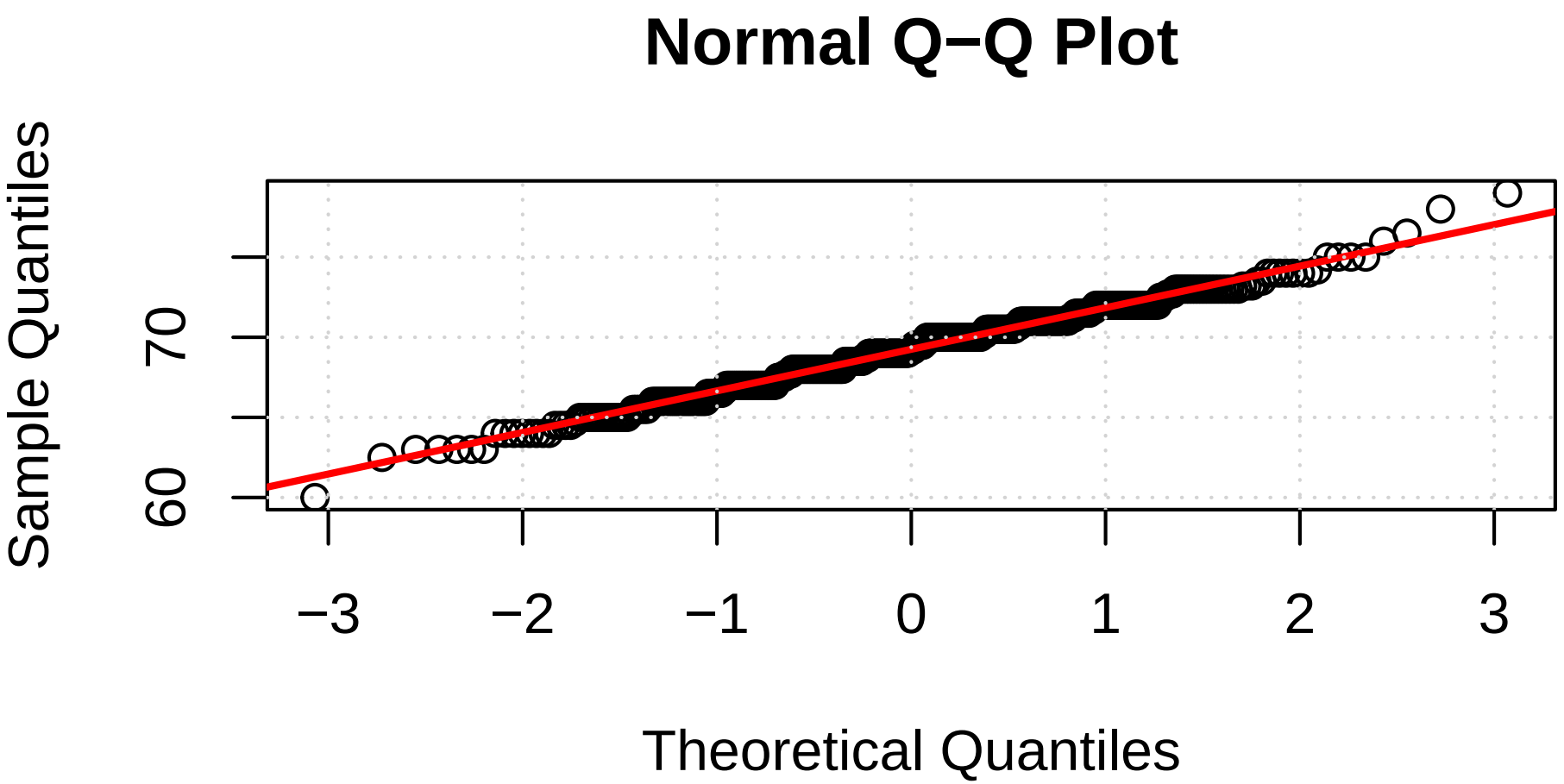
- Verifica a normalidade dos resíduos
  - Mais perto a uma linha reta, melhor o "fit" com uma distribuição normal

```
grqq <- ggplot(data = mods, aes(sample = .resid))
grqq <- grqq + stat_qq()
grqq <- grqq + labs(x = "Quantiles Teóricos",
                    y = "Quantiles da Amostra")
```


10
5
0
−5
−10
Quantiles da Amostra
−2
0
2
Quantiles Teóricos

Gráficos Q-Q Também Disponível Diretamente em Base R

```
qqnorm(boys$height)
qqline(boys$height, col = 2, lwd = 2)
grid()
```


Normal Q−Q Plot
Sample Quantiles
60
70
−3
−2
−1
0
1
2
3
Theoretical Quantiles

Teste-F das Variâncias do Modelo

- Teste-F é um teste que verifica que as variâncias das variáveis são perto de iguais
- Utiliza a Distribuição F
  - Tem 2 graus de libera de como parâmetros
- Serve como um teste de significância total de um modelo
- Produzido pelo função `Summary` da função `lm`

Teste-F do Modelo das Alturas Pai-Filho

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 38.25891    3.38663   11.30   <2e-16 ***
father       0.44775    0.04894    9.15   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.424 on 463 degrees of freedom
Multiple R-squared:  0.1531,    Adjusted R-squared:  0.1513
F-statistic: 83.72 on 1 and 463 DF,  p-value: < 2.2e-16
```

Resumo de Soma dos Quadrados

- Soma Total de Quadrados

$$SST = \sum (y_i - \bar{y})^2$$

- Soma dos Quadrados dos Erros

$$SSE = \sum (y_i - \hat{y})^2$$

- Soma dos Quadrados de Regressão

$$SSR = \sum (\hat{y}_i - \bar{y})^2 = SST - SSE$$

R^2 – Coeficiente de Determinação

- Medida de quanto a linha de regressão explica a variância em Y
- Relação entre a SSR e a SST

$$R^2 = \frac{SSR}{SST}$$

- Calculado pelo `lm`
  - visível em `Summary`
- Varia entre 0 e 1
- $\sqrt{R^2} = r$ (coeficiente de correlação)

R^2

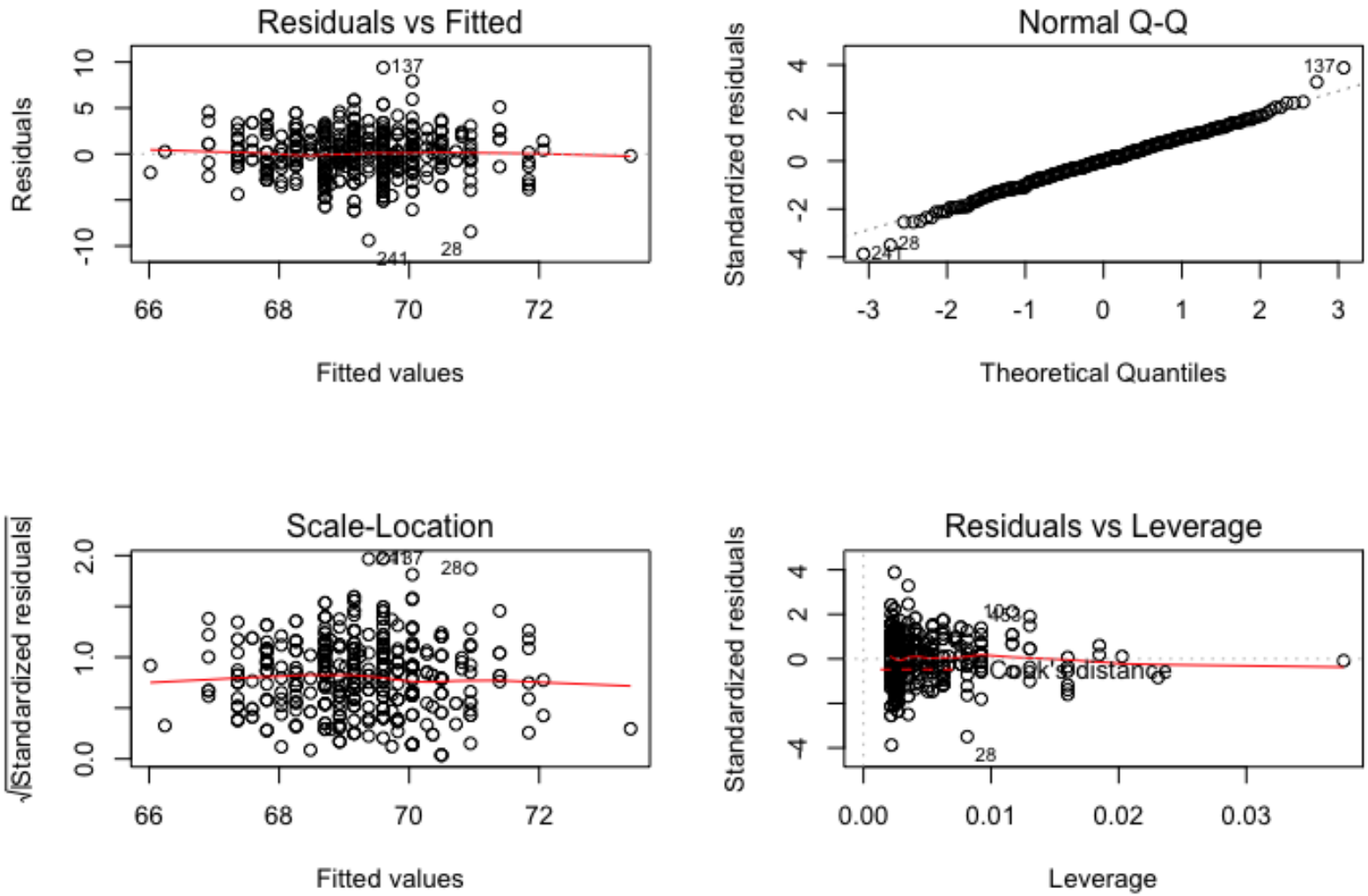
```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 38.25891    3.38663   11.30   <2e-16 ***
father       0.44775    0.04894    9.15   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.424 on 463 degrees of freedom
Multiple R-squared:  0.1531,    Adjusted R-squared:  0.1513
F-statistic: 83.72 on 1 and 463 DF,  p-value: < 2.2e-16
```

Significância de R^2

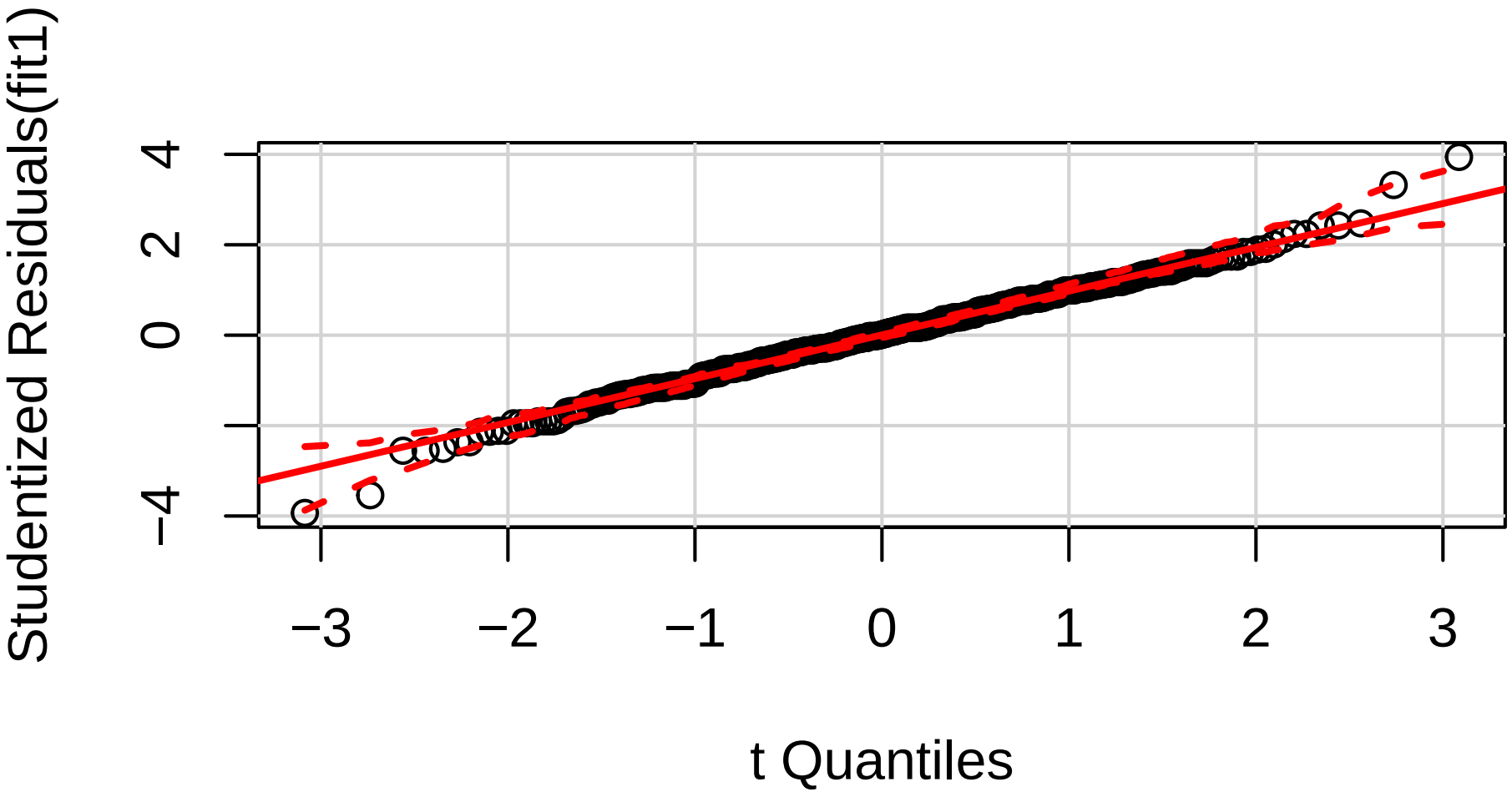
- Se 100% da variância ser explicado pela regressão
- $SSR = SST$
- $\therefore\ R^2 = SST/SST = 1$
- Variância completamente explicado pela regressão
- Em geral, o grau em que a regressão explica a variância no modelo

Dois Gráficos Mais Avançados

Função `plot` para Objetos `lm`



Função qqPlot() do Pacote car



Próxima Aula

- Análise mais profundas de nossos modelos de regressão
- Regressão com múltiplas variáveis independentes
- Regressão como modelo de machine learning