Aula 3 – Programming in R

Simple Linear Regression

James R. Hunter, Ph.D.

2022-10-27

Simple Linear Regression

Regression – History

- Term comes from eugenics (*eugenismo*) proposed by Sir Francis Galton.
- Studied heights on individuals within families
- Observed that children of
  - Children of tall parents tended to be shorter than the parents
  - Children of shorter parents tended to be taller than the parents
- Called this trend **regression to the mean**

Method of Least Squares

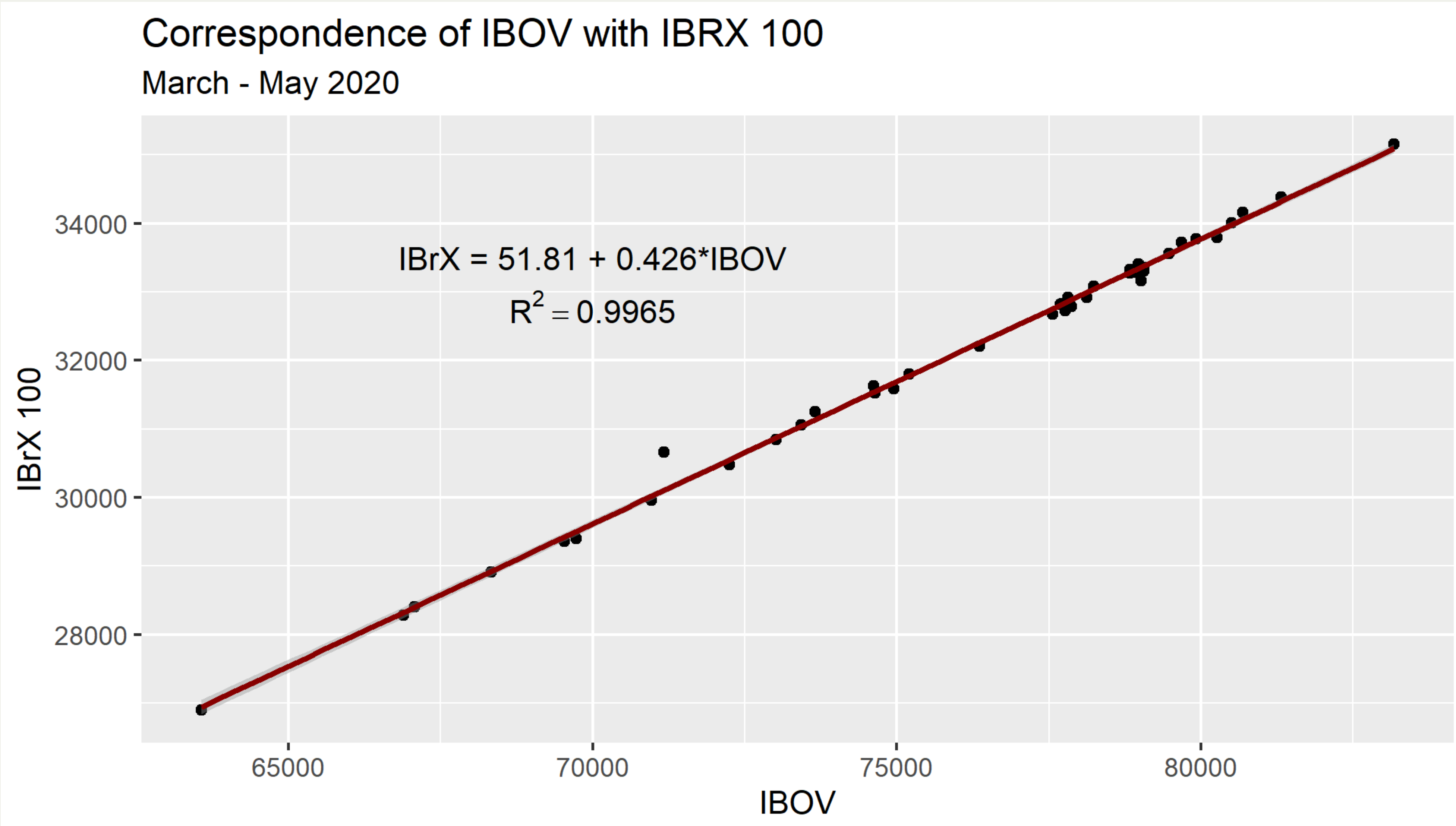
- Solve problems of regression with the *Least Squares* method
- Invented by Carl Friedrich Gauss (1777 - 1855)
- Method minimizes the differences between predicted linear values and the values based on the data
- Achieves the best relation between the real dependent variable and the predicted values of the variable

Purpose

> Predict a result on a dependent variable based on one or more indepedent variables

- One – *simple* linear regression
- More – *multiple* linear regression

Visualization of Regression



Straight Line

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

- β_1 = **Slope** of the line
- β_0 = **Intercept** of the line (where it crosses the y axis)
- Two parameters of regression
- Optimizing these parameters, Least Squares finds the straight line
- *Best* predicts the value of the dependent variable (y) based on the value of the independent variable (x)

Does "Best" Mean "Good"?

- Despite being the best way to predict y ,
  - Possible that it does **not** describe y well
- **Good** depends on the data
- **Best** depends on the algorithm

Regression Equation

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

- Y_i = value of the dependent variable
- β_0 = intercept
- β_1 = slope of the regression line
- X_i = value of the independent variable
- ϵ_i = error term for each case

Regression Equation - Estimation

$$\hat{Y_i} = b_0 + b_1 X_i + e_i$$

- $\hat{Y_i}$ = value of the dependent variable (estimated)
- b_0 = intercept (estimated)
- b_1 = slope of the regression line (estimated)
- X_i = value of the independent variable
- e_i = error term for each case

“Error” Term ϵ

- Also called **residual**
- Responsible for variability in y the the line cannot explain
- Does not mean “wrong”
- Only means “difference from a mean”
- Similar to what you learned with hypothesis tests

Least Squares

- Makes the calculation that minimizes the *error sum of squares*
- Errors = residuals = differences between the *observed* value and the *expected* value

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

- y_i = observed value of the dependent variable
- $\hat{y_i}$ = estimated value of the dependent variable

Example

- Data set of Galton about height in families
- Question is if children are taller or shorter than their parents
- He measured 898 sons/daughters in 197 families
- Original data records are in University College, London (UCL)

Variables

```r
galton <- readRDS(here::here("galton.rds"))

str(galton)
```

```
'data.frame':	898 obs. of  6 variables:
 $ family: Factor w/ 197 levels "1","10","100",..: 1 1 1 1 108 108 108 108 123
123 ...
 $ 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` – all are height in inches

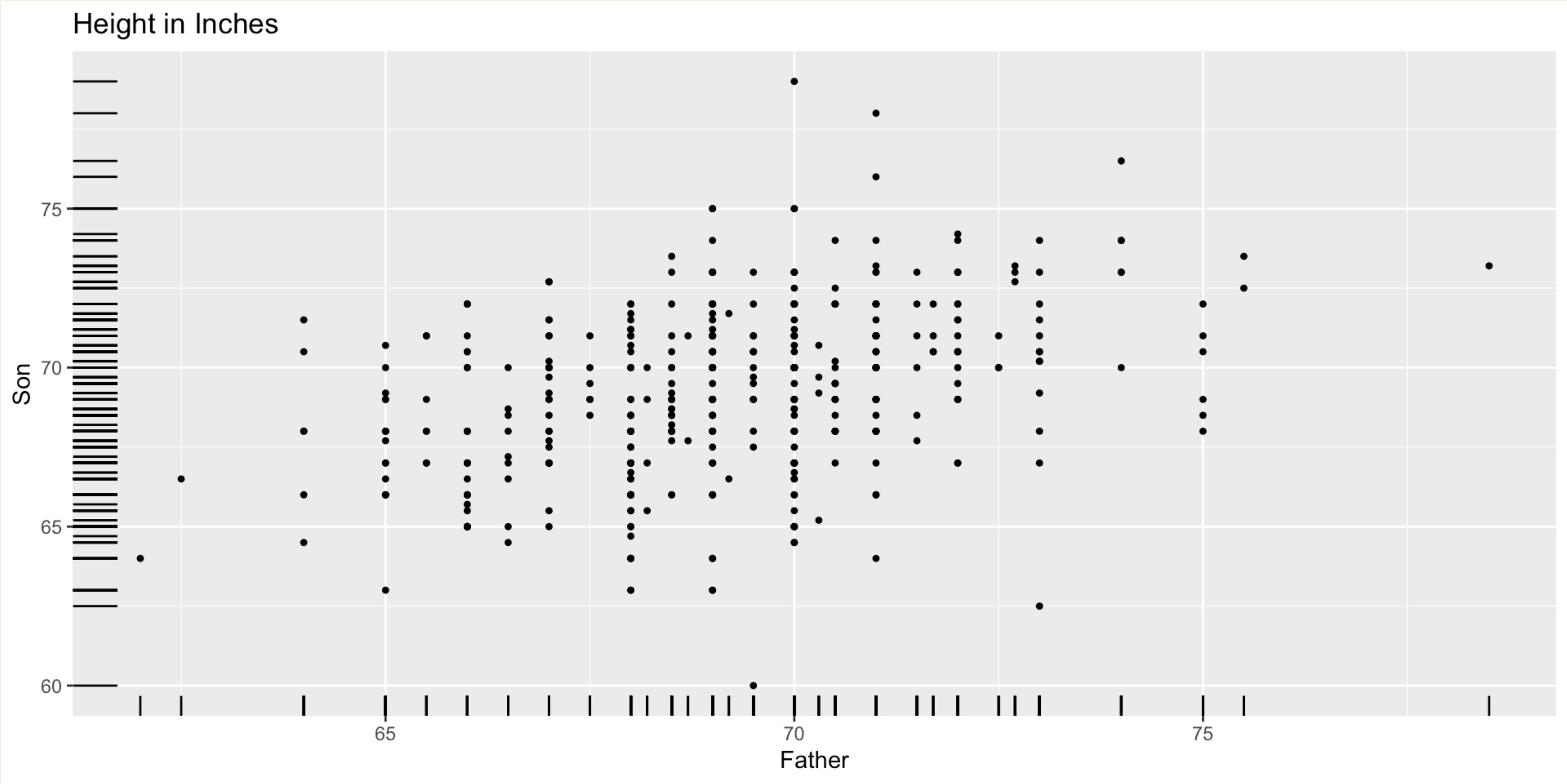
Focus on Fathers and Sons

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

glimpse(boys)
```

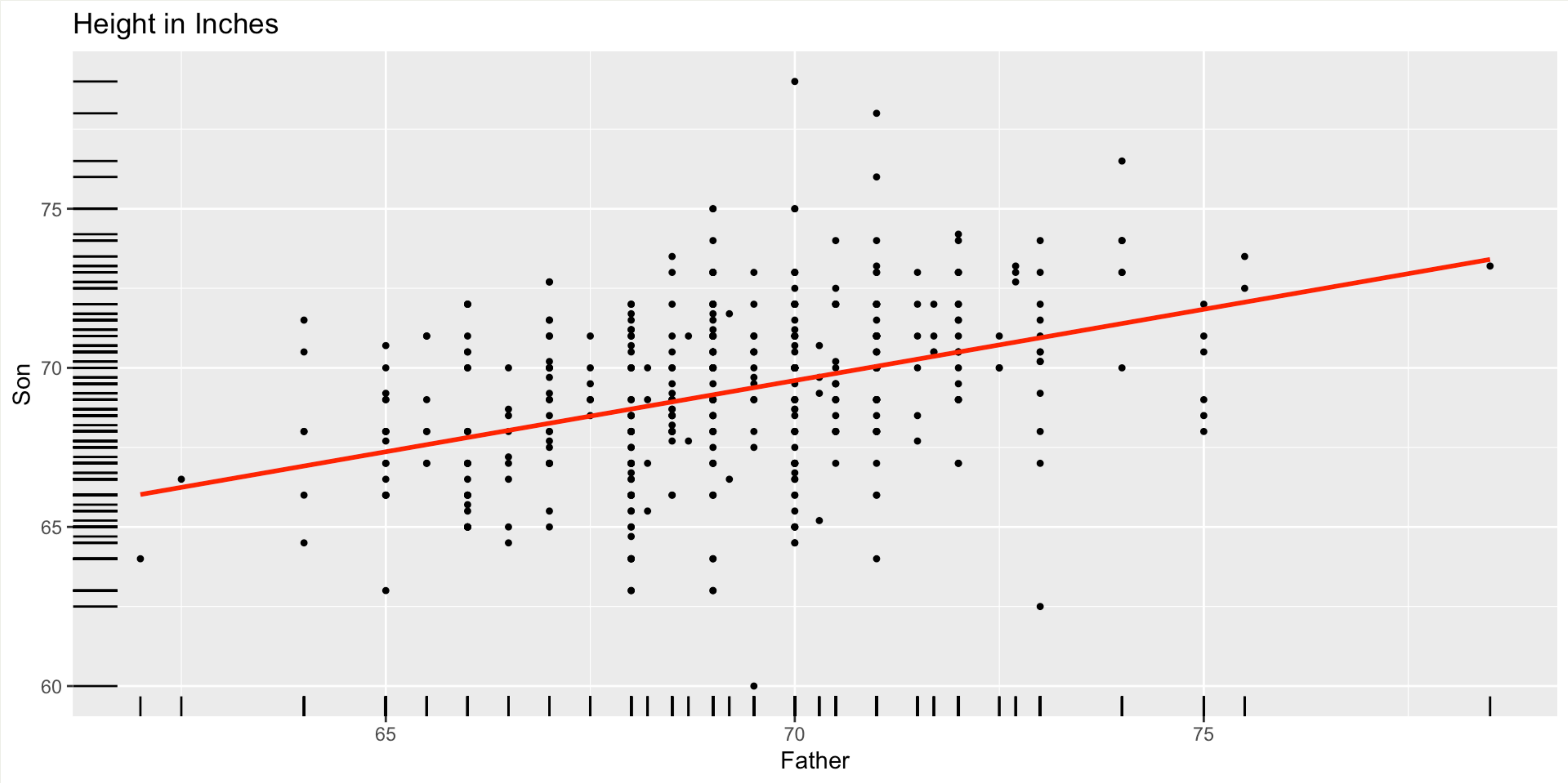
```
Rows: 465
Columns: 2
$ father <dbl> 78.5, 75.5, 75.5, 75.0, 75.0, 75.0, 75.0, 75.0, 75.0, 74.0, 74.…
$ height <dbl> 73.2, 73.5, 72.5, 71.0, 70.5, 68.5, 72.0, 69.0, 68.0, 76.5, 74.…
```

- `father` is the independent variable
- `height` is the dependent variable
- We want to see if the height of the father predicts the height of the son

Father/Son – Scatterplot



With Regression Line



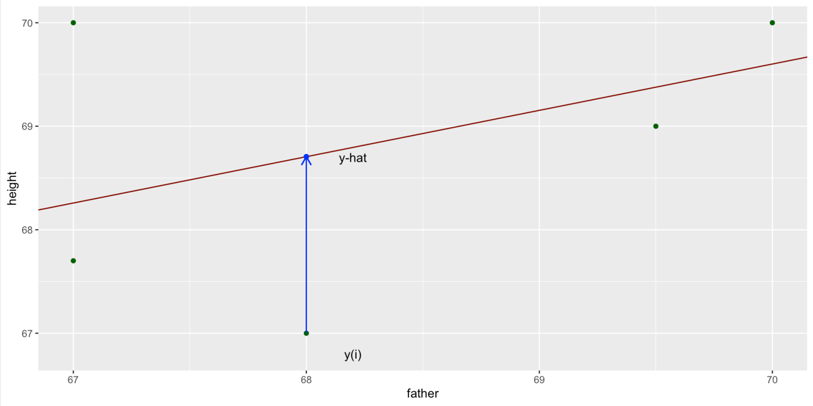
What We Learned from the Plot?

- **Seems** that taller the fathers, taller the sons
- Descriptive statistics of the 2 variables
  - And, correlation

```
boys

 2  Variables      465  Observations
--------------------------------------------------------------------------------
father
       n  missing distinct     Info     Mean  pMedian      Gmd      .05
     465        0       30    0.991    69.17    69.25    2.552     65.0
     .10      .25      .50      .75      .90      .95
    66.0     68.0     69.0     70.5     72.0     73.0

lowest : 62   62.5 64   65   65.5, highest: 73   74   75   75.5 78.5
--------------------------------------------------------------------------------
height
       n  missing distinct     Info     Mean  pMedian      Gmd      .05
     465        0       46    0.996    69.23    69.25    2.952     65.0
     .10      .25      .50      .75      .90      .95
    66.0     67.5     69.2     71.0     72.3     73.0

lowest : 60   62.5 63   64   64.5, highest: 75   76   76.5 78   79
--------------------------------------------------------------------------------
[1] "Correlation Coefficient: 0.391"
```

How Do We Calculate the Regression Line?

- A line that minimizes the difference between y_i and $\hat{y}$
- Need to work with squared differences
  - To not end up with a sum of 0
- SSE - Error Sum of Squares



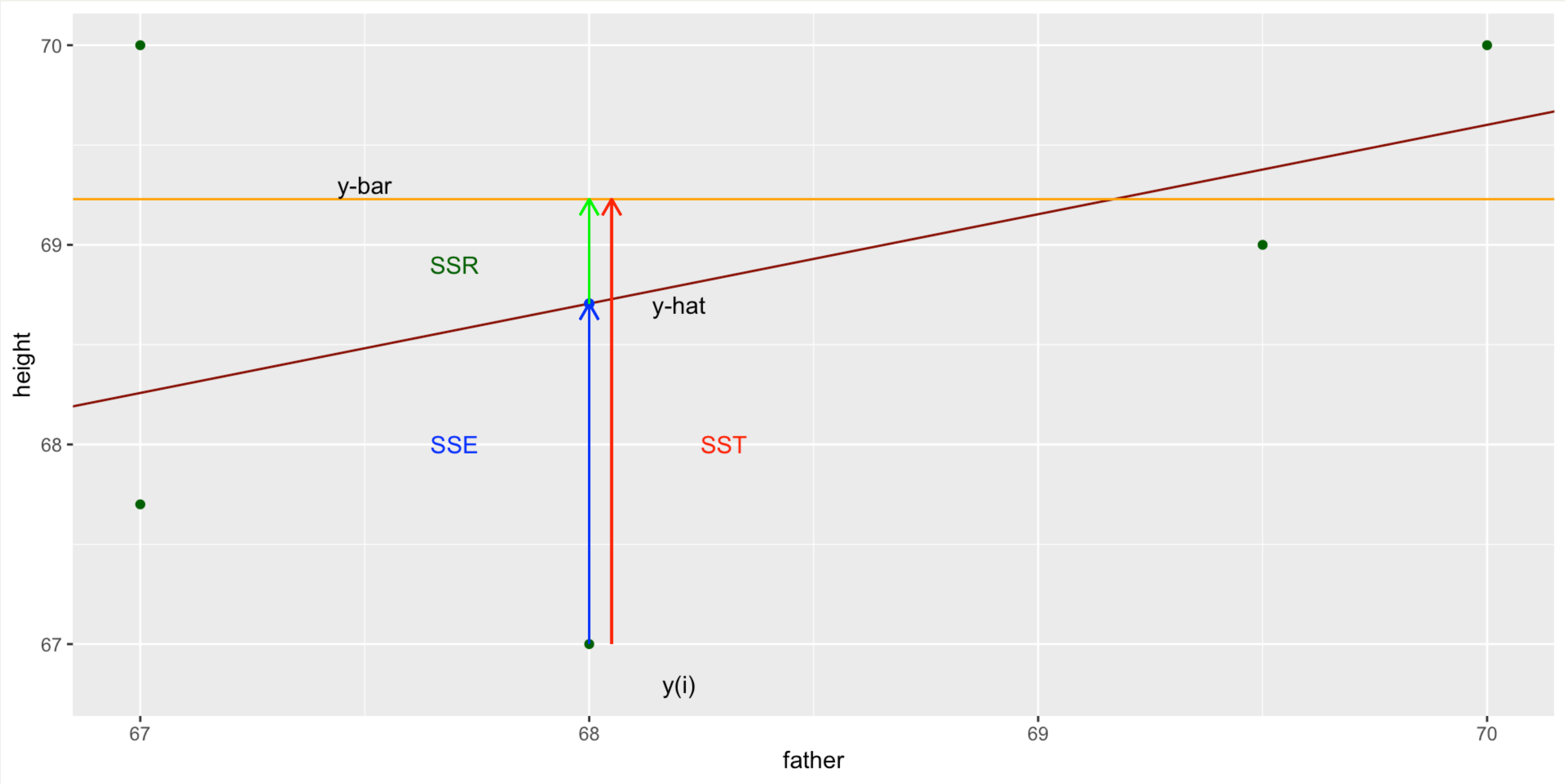
SSE – Part of Total Sum of Squares (SST)

$$SST = SSE + SSR$$

- SST – Total
- SSE – Related to errors/residuals
- SSR – Related to/Explained by regression

SST – What Does It Represent?

The total variance is the difference between the model value for each value of X and the mean of the values of the dependent variable ($\hat{y}$)



Sum of Squares

- Refer to the sum of squares we want to minimize as the **SSE**
  - Error sum of squares
- SSE is a component of the total sum of squares (SST)
- SSE — the of the squares related to the residuals
- SSR — sum of squares related to the regression
- Expression for the SSE

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

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

To Determine the Formula for β_0 & β_1

- To minimize the SSE (determine the most efficient line), we need to use calculus
- Set the partial derivatives of the SSE with respect to β_0 and β_1

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

- Called the normal equations
- We let the software calculate the parameters of the equation

Function in R

- Function `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, …)`
- Important arguments are `formula`, `data`, `subset`, `weights`, `na.action`
  - `formula`: where you show which variables you are modelling
  - Dependent variable comes first
  - Separated from the independent by " ~ "
- For the `boys`: `height ~ father`
  - `data`: data frame or tibble that contains the variables
  - `subset`, `weights`: parameters that permit customization of the variables
  - `na.action`: how you will deal with missing data in the model variables

Function Applied to Fathers and Sons

- Function `lm` produces a *list* of 12 items in a special format

```r
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
```

What Does This Model Say?

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

- If a father had 0 height, the son would be 38.259 inches tall
  - Doesn't make practical sense
  - Establishes a base for the height calculation
- For each incremental inch on the father's height, the son would be 0.448 inches taller

Extract the Coefficient Values

- Option 1: use `broom::tidy`
  - Automatically extracts the key information and puts in a tibble

```
broom::tidy(fit1)  |> knitr::kable()
```

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

- Option 2: use `coef`

```
coef(fit1)
```

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

Predictions of New Values

- You can use the model parameters to predict new values of the heights of sons
  - Use `broom::augment`
- How tall would the son of a 72 inch father be?

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

```
# A tibble: 1 × 2
  father .fitted
   <dbl>   <dbl>
1     72    70.5
```

What Does the Model Mean? How to Interpret It?

Relationship between the Independent and Dependent Variables?

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

- If β_1 (slope of the line) were 0, what would be the equation?

$$Y_i = \beta_0 + \epsilon_i$$

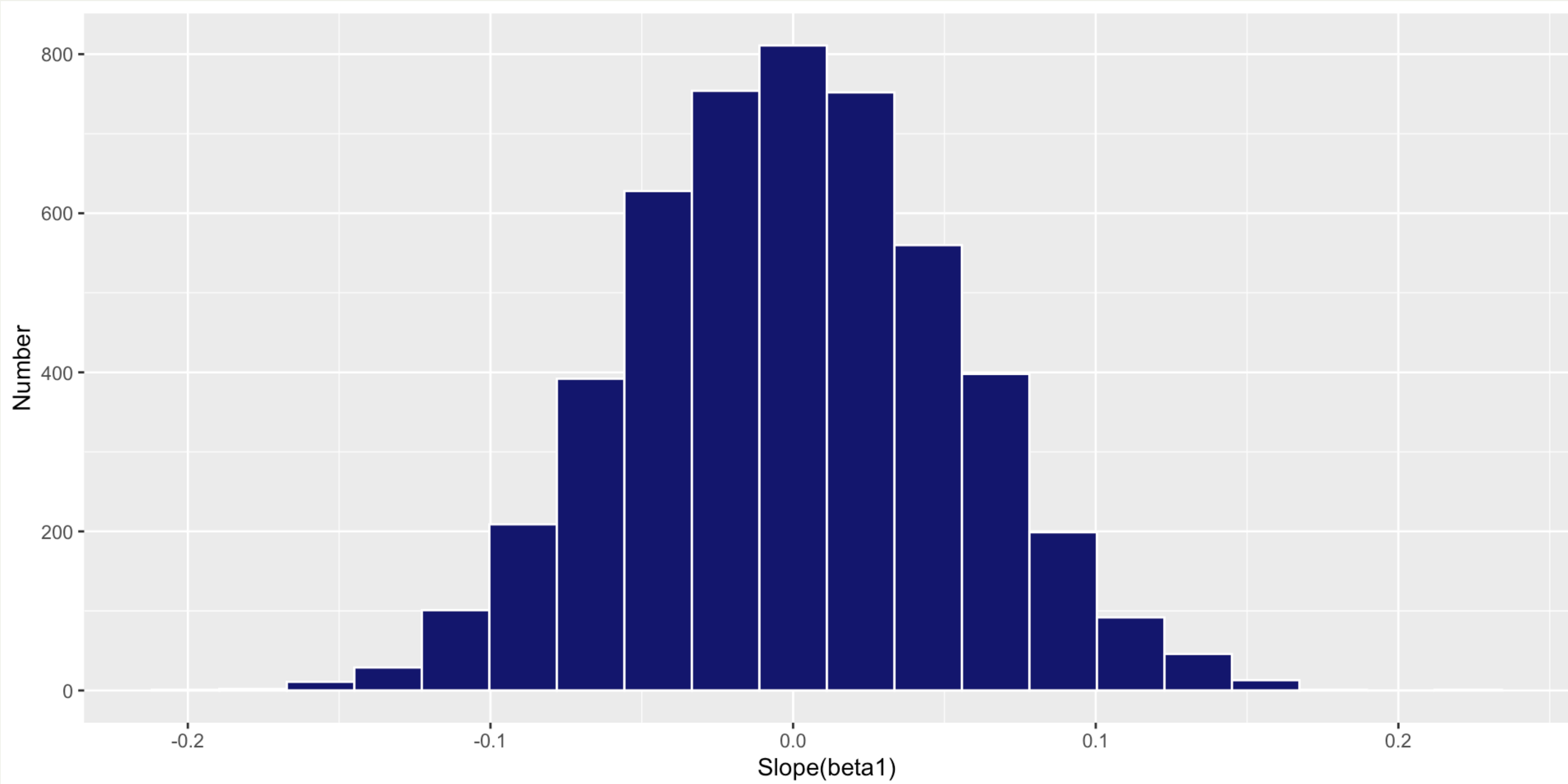
Test of the Null Hypothesis

- We will make a simulation of the null hypothesis
- If we do not reject the null, any son's height could have occurred for any father's height
- We can calculate the regression model 5,000 times shuffling around the son's heights
  - Application of Monte Carlo simulation
- As a result, we can focus on the values of the slope, β_1
- 2nd, we will compare our observed value of β_1 (0.4477479) to see where it falls in the simulated values

Histogram of Slopes of the Simulated Models – Code

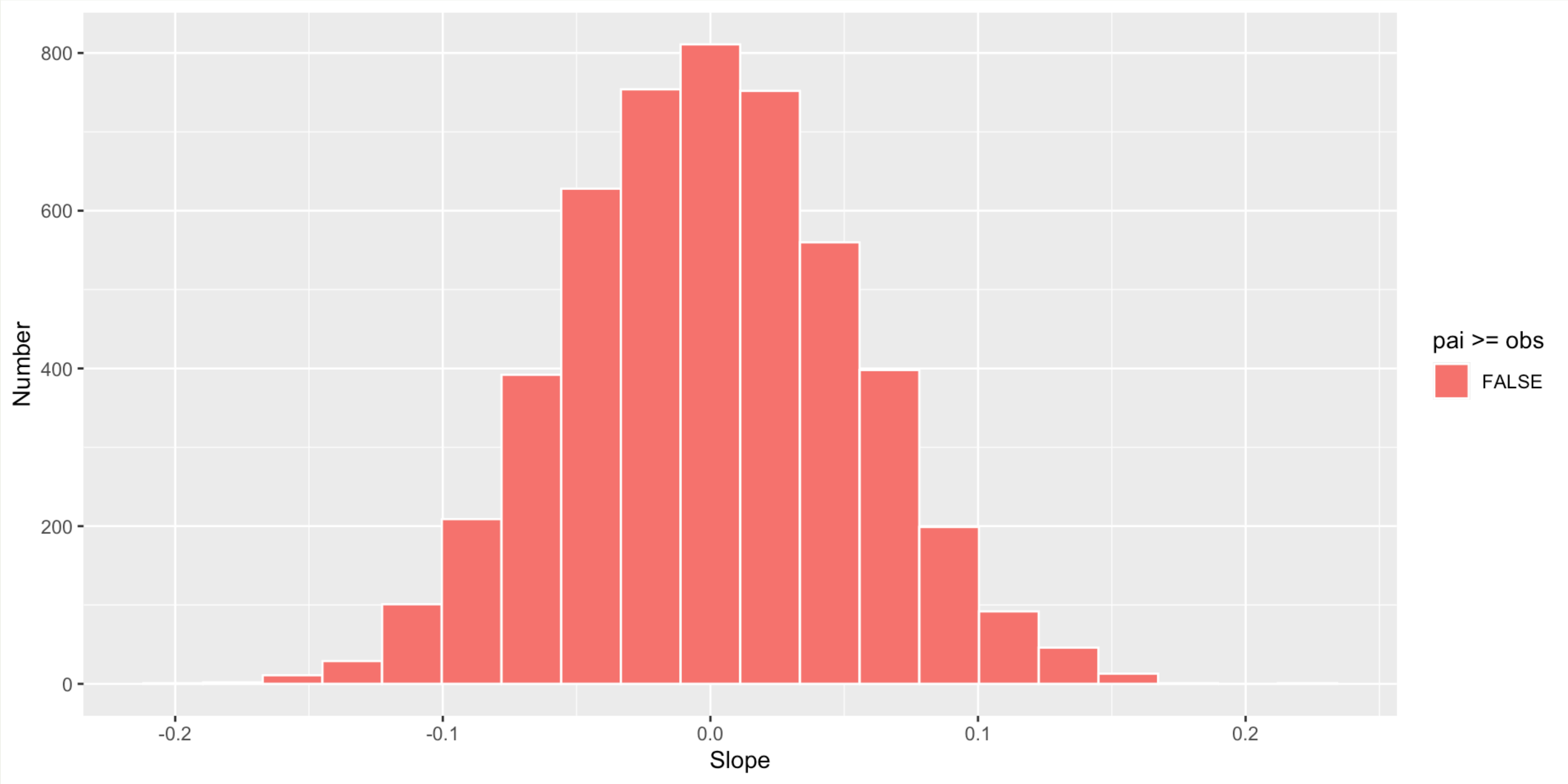
```r
set.seed(1946)
homodelos <- replicate(5000, (lm(mosaic::shuffle(height) ~ father, data = b
homodelos <- tibble(homodelos[2,])
colnames(homodelos) <- "father"
modgr1 <- ggplot(homodelos, aes(x = father))
modgr1 <- modgr1 + geom_histogram(color = "white", fill = "midnightblue", b
modgr1 <- modgr1 + labs(x = "Slope(beta1)", y = "Number")
modgr1
```

Histogram



Histogram with Values Above and Below Observed Slope

```
Number of simulations with beta1 >= obs:  0
```



The p-value of the Slope (β_1)

- Because **none** of the simulations produced a value higher than our observed value (0.448)
- We can conclude that the p-value of this test is 0
- There is **no** chance that the slope = 0
- Thus, we reject the null hypothesis and conclude that a linear relationship does exist between the heights of fathers and sons

Assumptions of Linear Regression and How to Test Them

Assumptions of Linear Regression

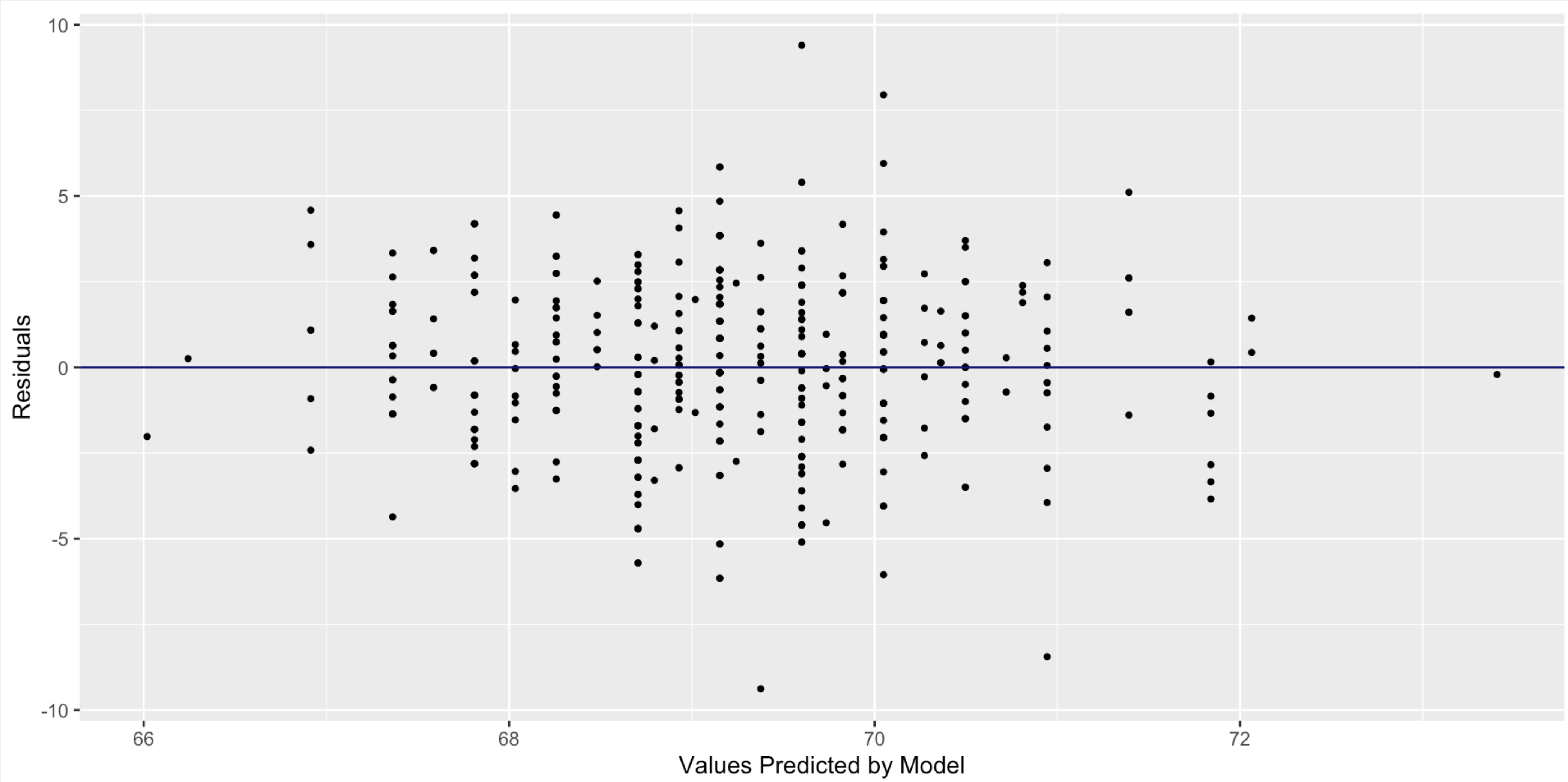
1. All independent variables must have the same variance
    - Graph of residuals should avoid patterns when looking from left to right
2. All the observations, residuals and independent variables must be independent of each other
    - Graph of residuals should not show a sinuous pattern
3. Residuals should have a near-normal distribution
    - Q-Q graph of the standardized residuals should be a straight line
    - Shows that the variables have a multivarite normal distribution
4. Independent variables should avoid *multicollinearity*
    - They should not have high correlations between them

Residuals Graph

- Graph that shows the value predicted by the model ("fitted value") vs. the residual
- Use the function `broom::augment()`
  - Extracts efficiently the values used in the model tests

```r
mods <- broom::augment(fit1)  #<<

residgr <- ggplot(data = mods, mapping = aes(x = .fitted, y = .resid))
residgr <- residgr + geom_point(shape = 20)
residgr <- residgr + geom_hline(yintercept = 0, color = "midnightblue")
residgr <- residgr + labs(x = "Values Predicted by Model",
                          y = "Residuals")
```

Graph



Importance of Residuals

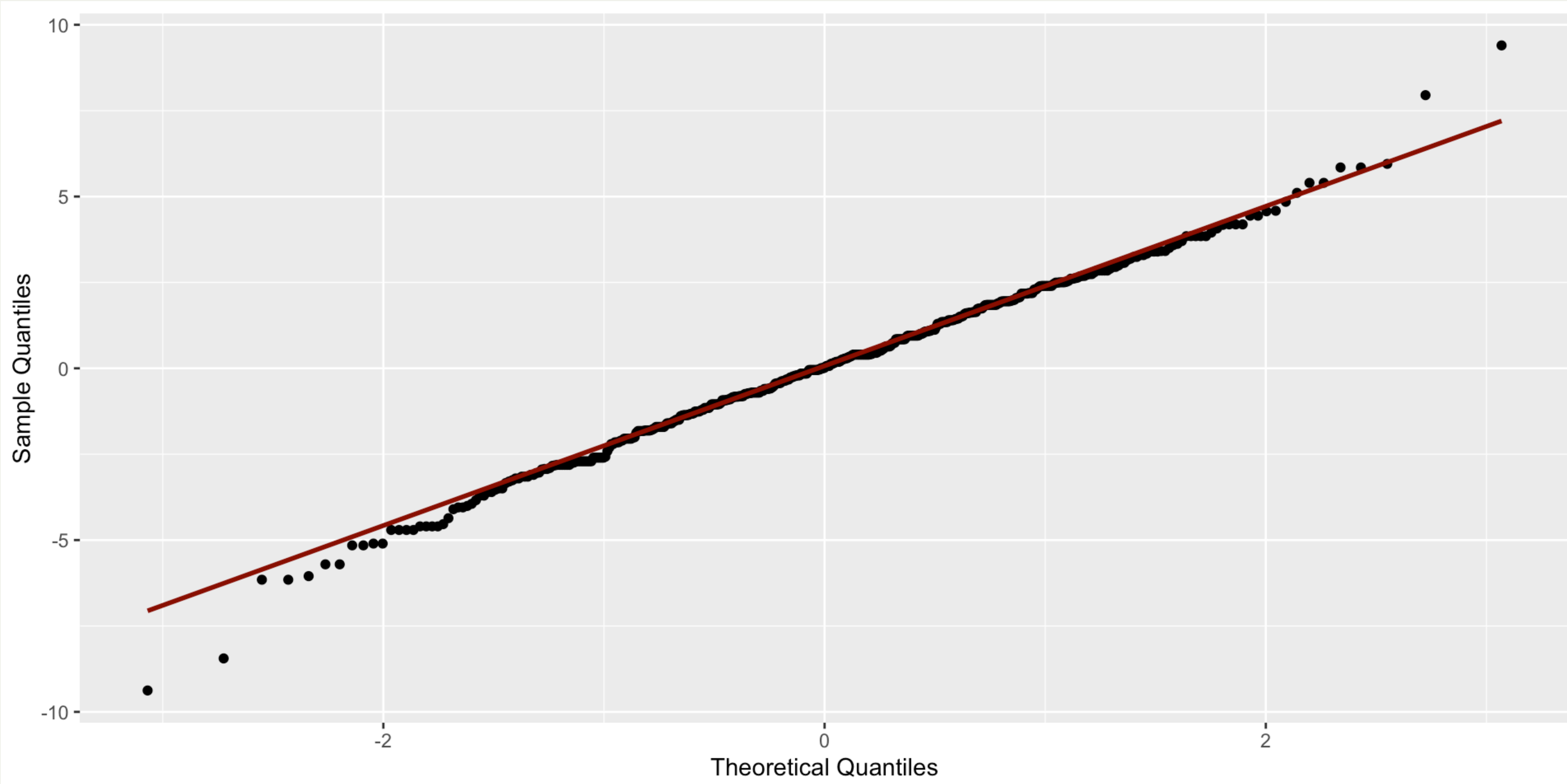
- Can use the residuals to verify if the model respects the assumptions of regression
- Should not show any linear trend

Q-Q Graph

- Verifies the normality of the residuals`
- Closer the curve to a straight line, the better the “fit” with a normal distribution
- This version in `ggplot2` with the `stat_qq` geometry

```r
grqq <- ggplot(data = mods, aes(sample = .resid))
grqq <- grqq + stat_qq()
grqq <- grqq + stat_qq_line(color = "darkred", size = 1)
grqq <- grqq + labs(x = "Theoretical Quantiles",
                    y = "Sample Quantiles")
```

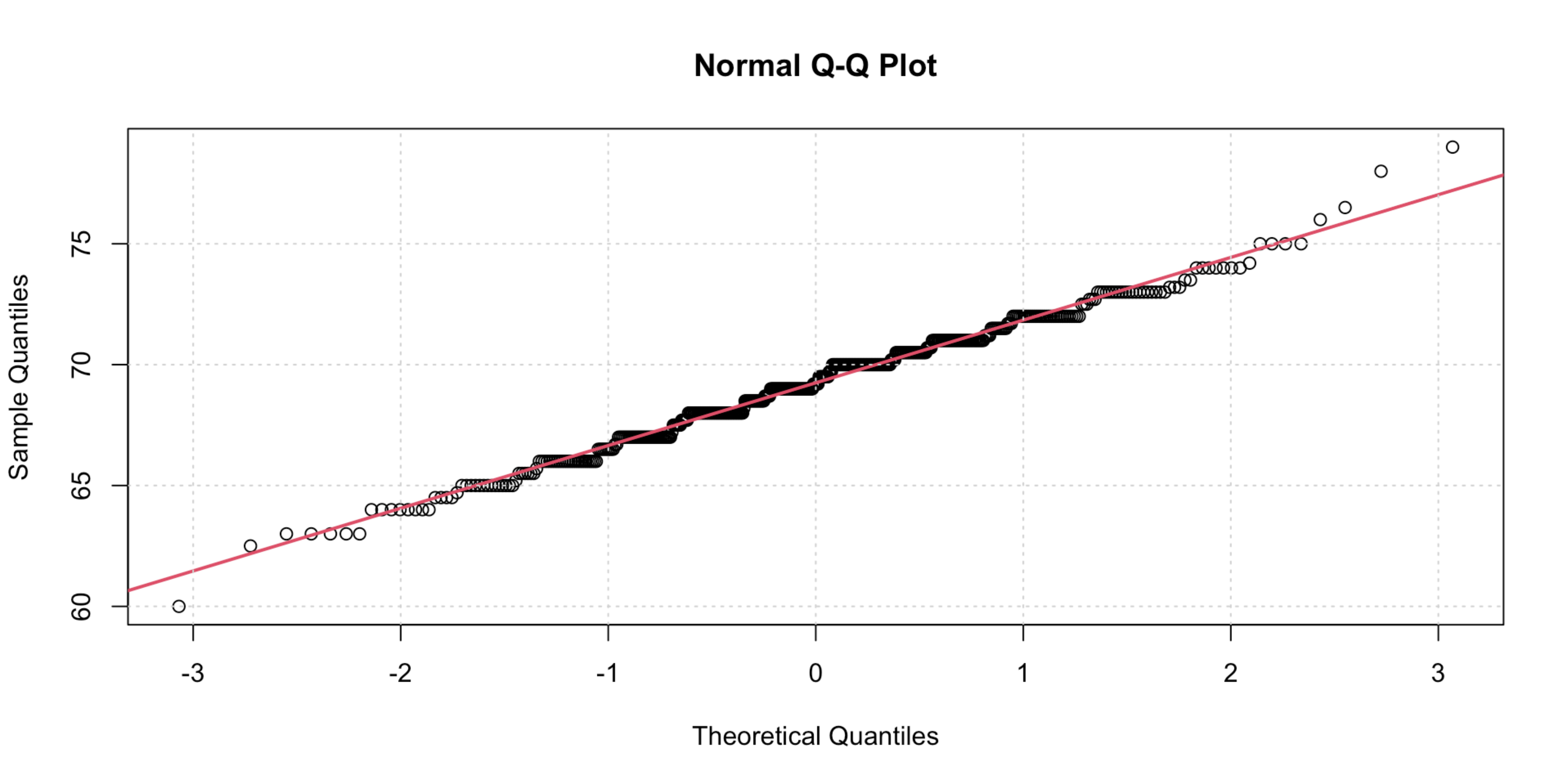
Graph



Q-Q Graphs Also Available in Base R

```r
qqnorm(boys$height)

qqline(boys$height, col = 2, lwd = 2)

grid()
```

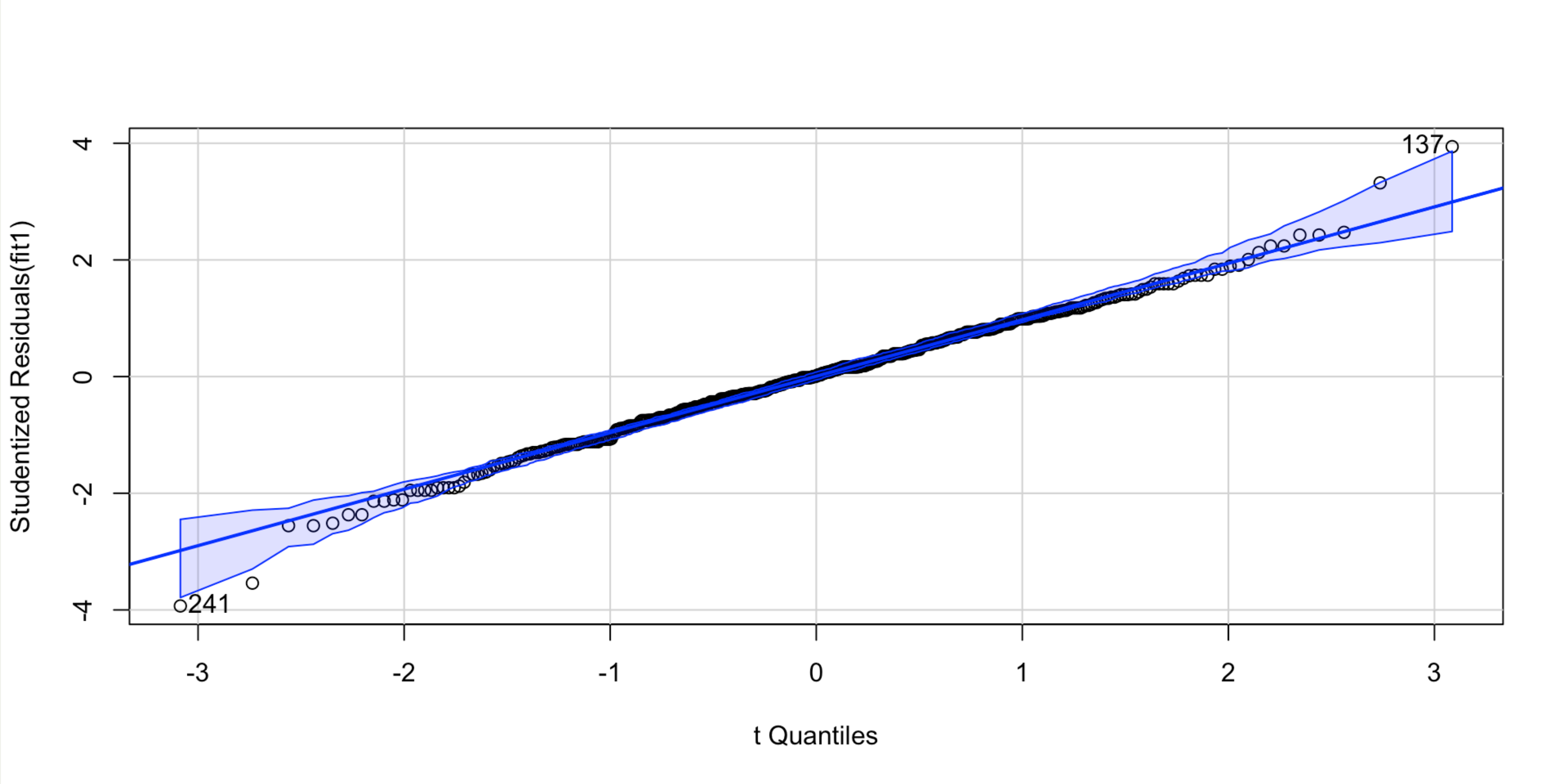
Base R Q-Q Plot



`qqPlot()` Function from the `car` Package

```
car::qqPlot(fit1)
```

```
[1] 137 241
```



F-Test of Model Variance

- F-Test is a test that verifies that the variances of variables are close to equal
- Uses the F Distribution
  - With 2 degrees of freedom as parameters
  - Serves as a test of significance for the model as a whole
- Shown in the `summary()` function output for the `lm()` function

F-Test for the Son-Father Heights Model

```
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
```

Summary of the Sum of Squares

- Total Sum of Squares

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

- Error Sum of Squares

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

- Regression Sum of Squares

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

R^2 – Coefficient of Determination

- Measure of how much the regression line explains the variance in Y
- Ratio of SSR to SST

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

- Calculated by `lm()`
- Appears in `summary(lm)`
- Varies between 0 and 1
- $\sqrt{R^2} = r$ (correlation coefficient)

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
```

Importance of R^2

- If 100% of the variance in Y can be explained by the regression

– $SSR = SST$

- $\therefore\ R^2 = SSR/SST = 1$
- Variance completely explained by the regression
- Means there is no error
- In general, the degree to which the regression explains the model variance