Topic 10 – Linear functions

ENVX1002 Introduction to Statistical Methods

Liana Pozza
The University of Sydney

Dec 2024

THE UNIVERSITY OF SYDNEY

Recap

Last week...

- Correlation r: a measure of the strength and direction of the linear relationship between two variables
- Is there a causal relationship between two variables?
  - ➟ **No**: use correlation analysis
  - ➟ **Yes**: use *regression analysis*

Simple linear regression modelling

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

> Basically, a deterministic straight line equation $y = c + mx$, with added random variation that is normally distributed

$$Y = c + mx + \epsilon$$

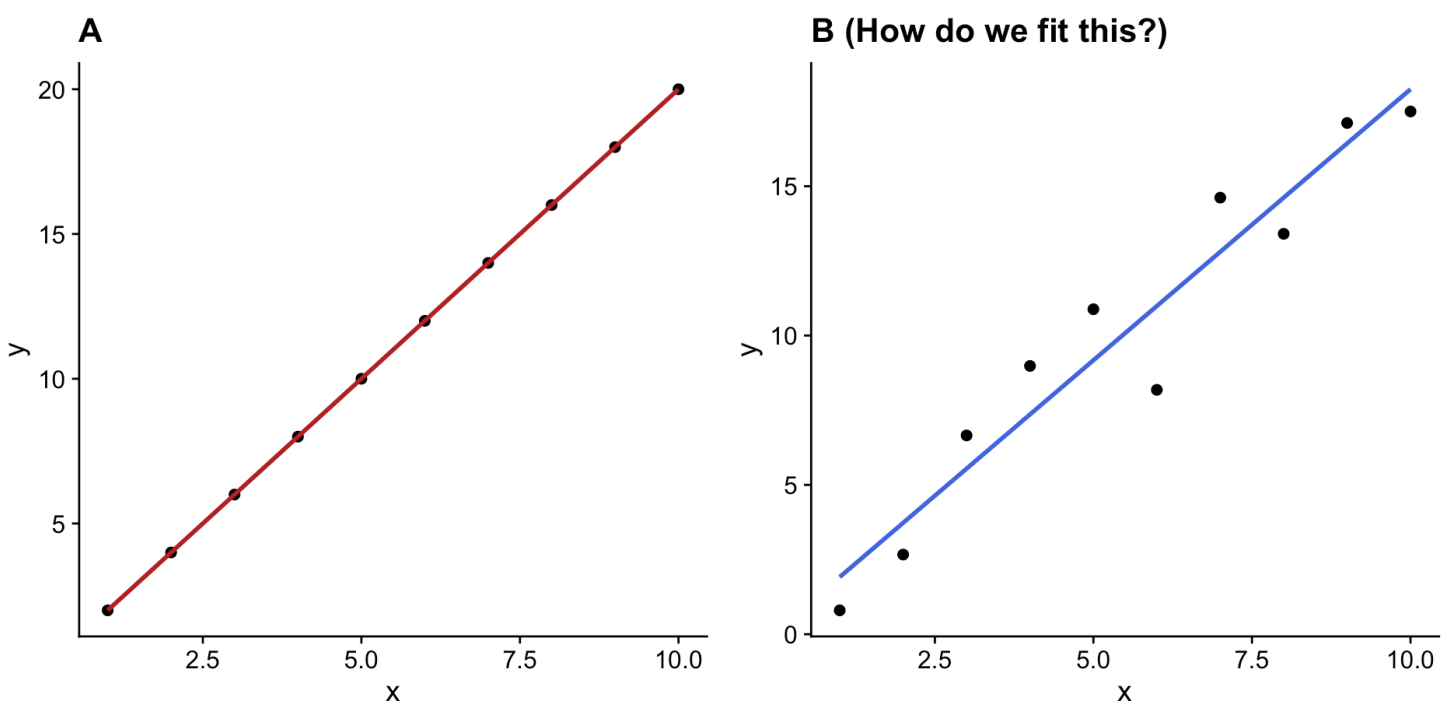
Fitting the line

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

$$Y = c + mx + \epsilon$$

How do we fit a line to data if data are “noisy”?

▶ Code



Least squares

> The method of least squares is the **automobile of modern statistical analysis**: despite its limitations, ocassional accidents and incidental pollution, it and its numerous variations, extensions, and related conveyances **carry the bulk of statistical analyses**, and are known and valued by nearly all.

– Stigler, 1981 (emphasis added)

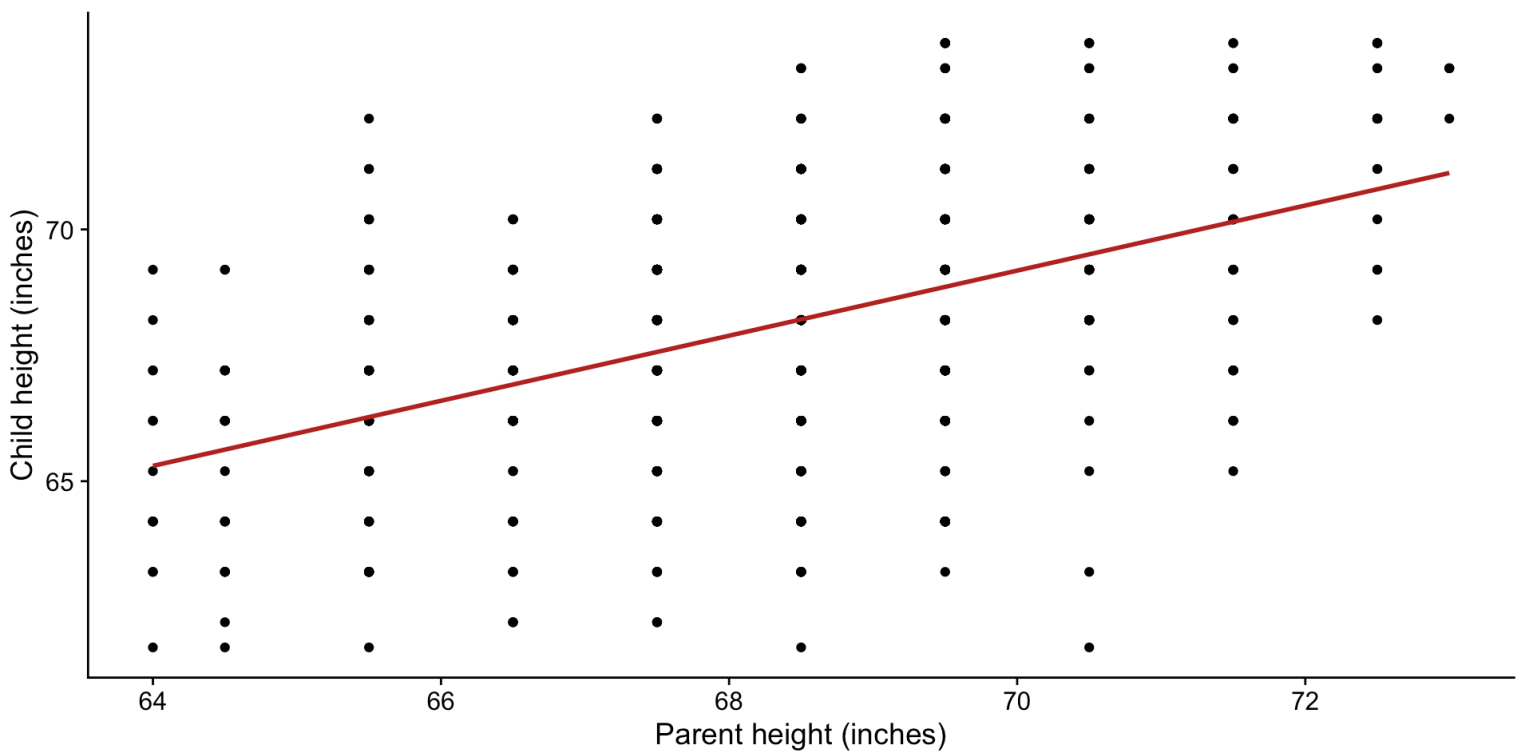
Usage

- **Student's t-test**
- **linear regression**
- ANOVA
- logistic regression
- nonlinear regression
- ridge regression
- lasso regression
- principle component analysis
- generalised linear model
- etc...

Galton's data revisited

- Galton's data on the heights of parents and their children.
- Is there a relationship between the heights of parents and their children?

▶ Code



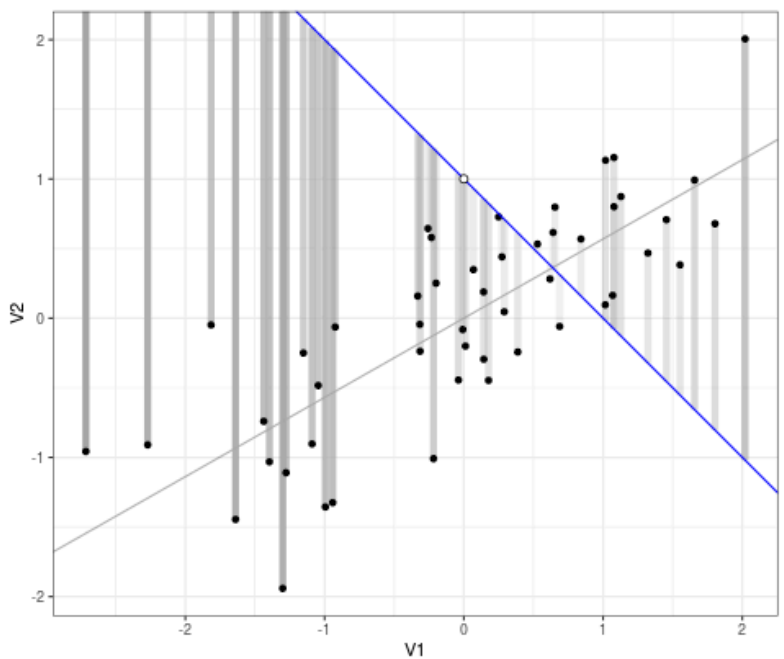
How did we end up with the line in the plot above?

Fitting the model

How do we fit a line?

- *Minimise* the sum of the squared residuals:

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

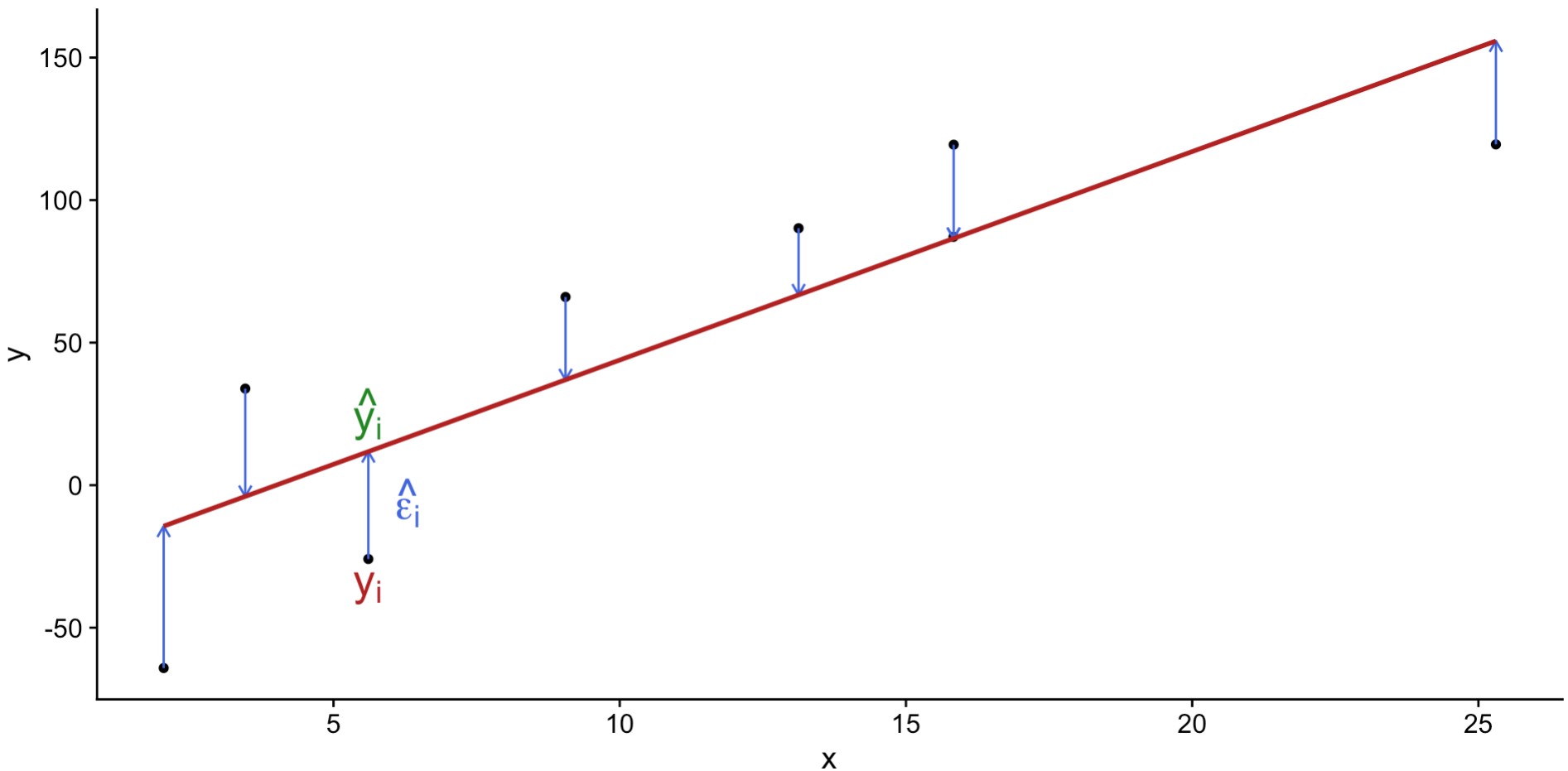


Source

Residuals, $\hat{\epsilon}$

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

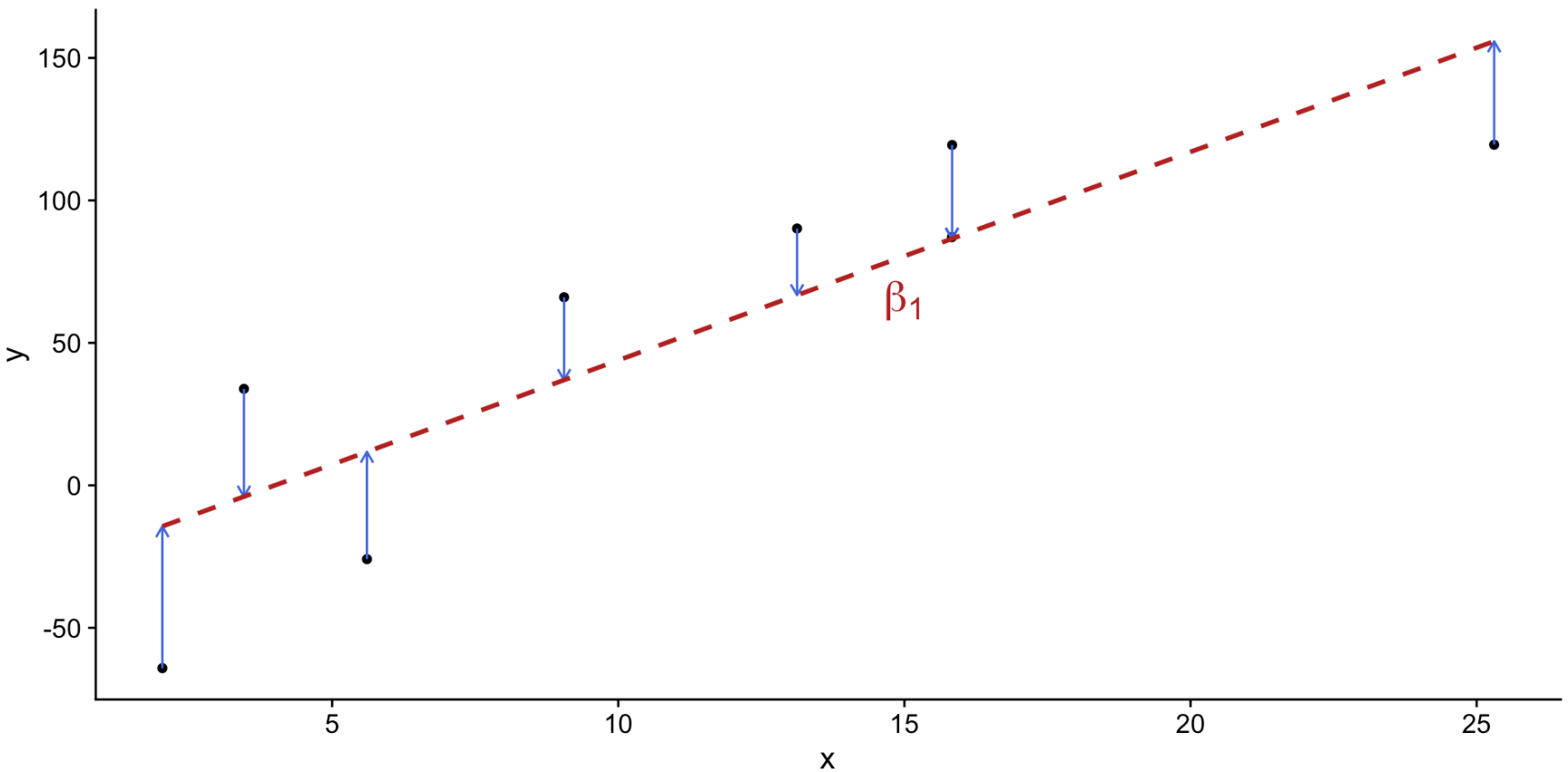
▶ Code



Slope, β_1

$$\beta_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{Cov(x,y)}{Var(x)} = \frac{SS_{xy}}{SS_{xx}}$$

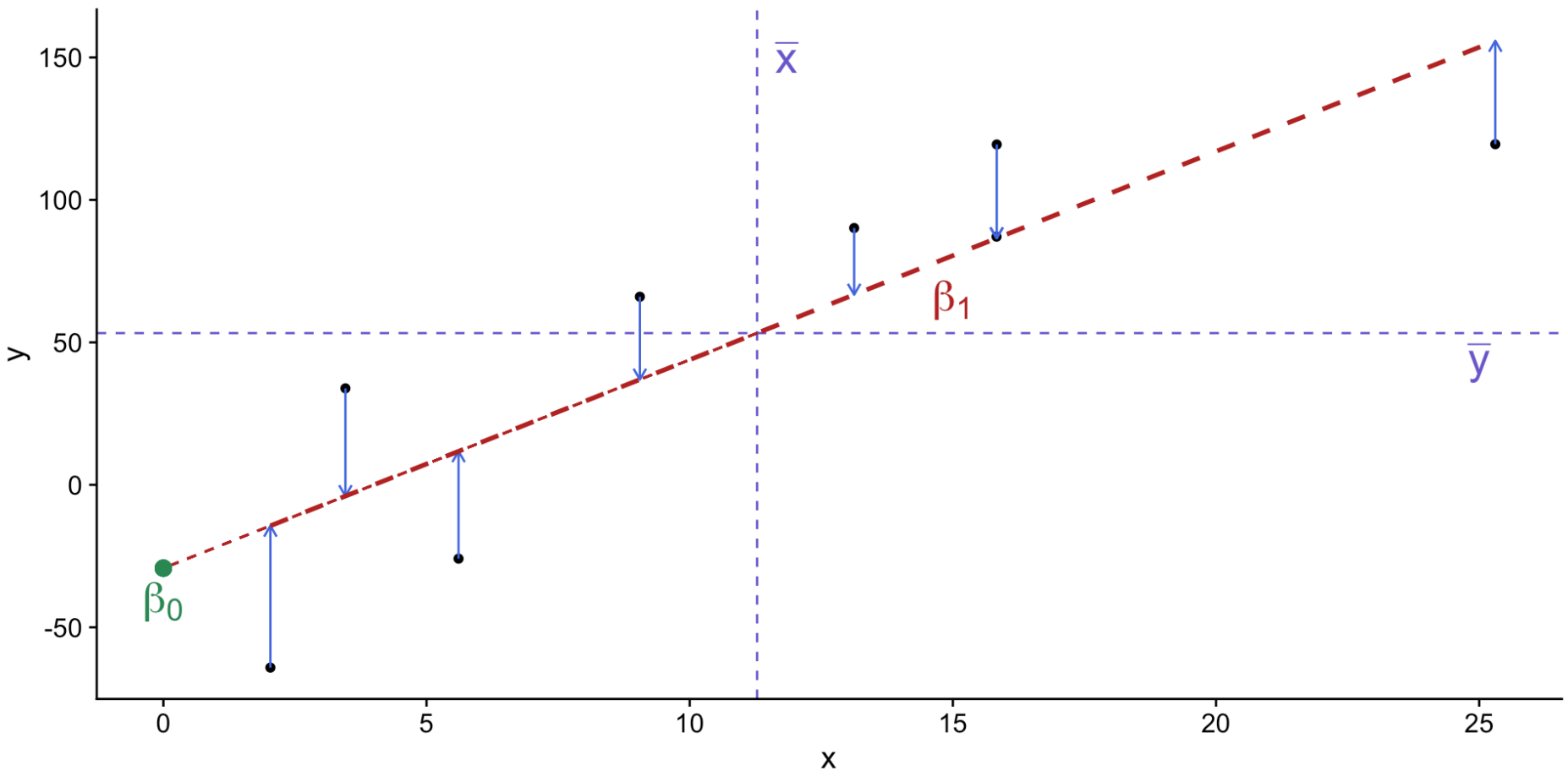
▶ Code



Intercept

$$\beta_0 = \bar{y} - \beta_1 \bar{x}$$

▶ Code



It’s a lot easier in R…

Fitting a linear model in R

Is there a relationship between the heights of parents and their children?

```r
fit <- lm(child ~ parent, data = Galton)
fit
```

```
Call:
lm(formula = child ~ parent, data = Galton)

Coefficients:
(Intercept)       parent
    23.9415       0.6463
```

$$\widehat{child} = 23.9 + 0.646 \cdot parent$$

But is the model any good?

Assessing model fit

Assumptions

The data **must** meet certain criteria, which we often call *assumptions*. They can be remembered using **LINE**:

- **L**inearity. The relationship between y and x is linear.
- **I**ndependence. The errors ϵ are independent.
- **N**ormal. The errors ϵ are normally distributed.
- **E**qual Variance. At each value of x, the variance of y is the same i.e. homoskedasticity, or constant variance.

> **Tip**
>
> All but the independence assumption can be assessed using diagnostic plots.

Assumptions: Why do we care?

- If the assumptions are met, then we can be confident that the model is a good representation of the data.
- If they are *not* met, the results are still presented, but our interpretation of the model is likely to be flawed.

> ⚠ **Warning**
>
> R will not warn you if the assumptions are not met. It is up to you to check them!

How do we check the assumptions?

Recall that the linear model is a **deterministic straight line equation** $y = c + mx$ plus some **random noise** ϵ:

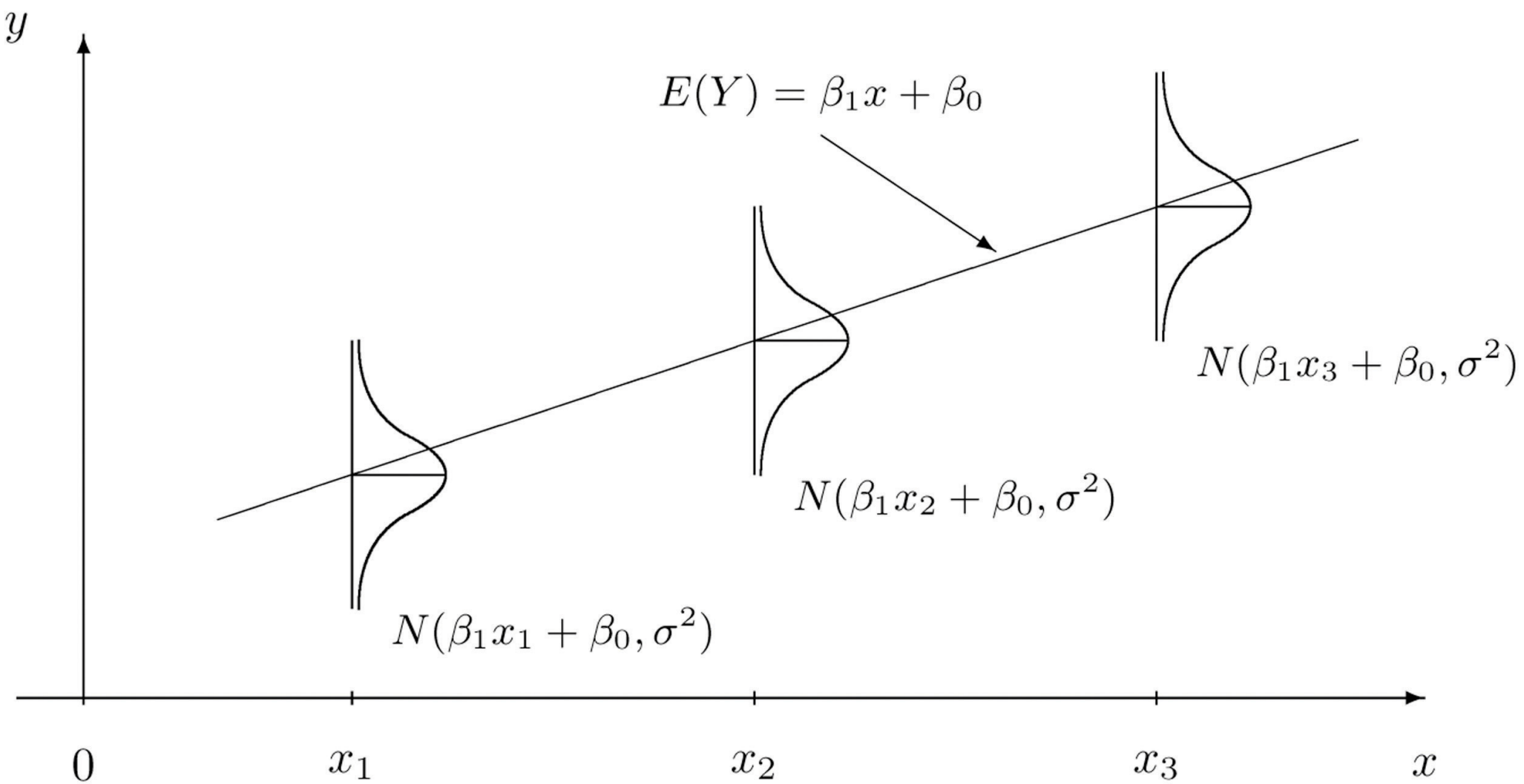
$$Y_i = \beta_0 + \beta_1 x + \epsilon$$

- **If the only source of variation in y is ϵ, then we can check our assumptions by just looking at the residuals $\hat{\epsilon}$.**

How do we get the residuals?

- Fit the model...
- Residuals need to be calculated from the model, not from the raw data.
- In R, these values are stored automatically.

Another way to look at residuals



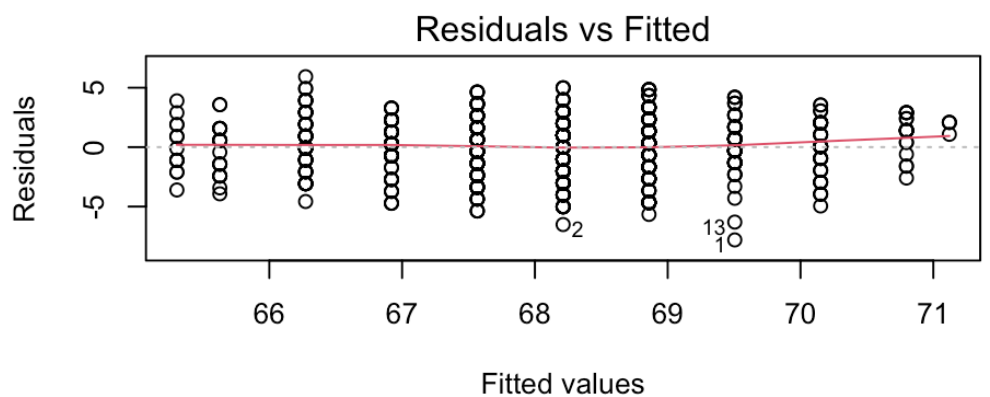
Once you have fitted the line, it does not change. The residuals are the vertical distances between the points (not shown) and the line.

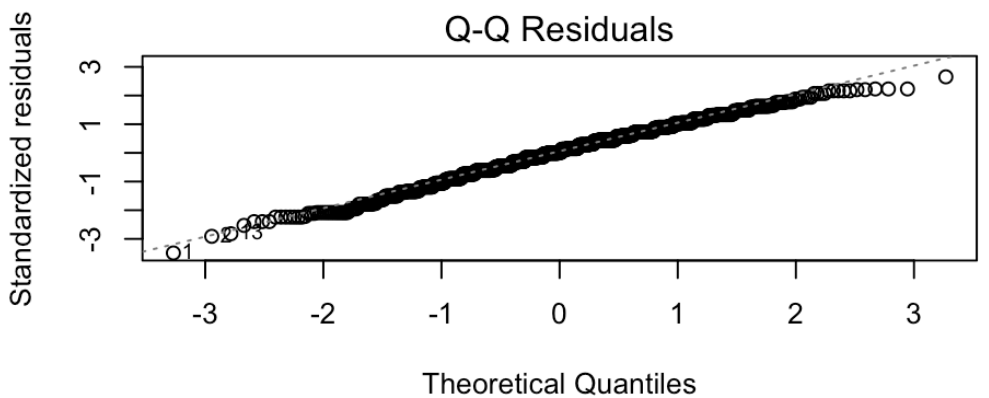
Checking assumptions

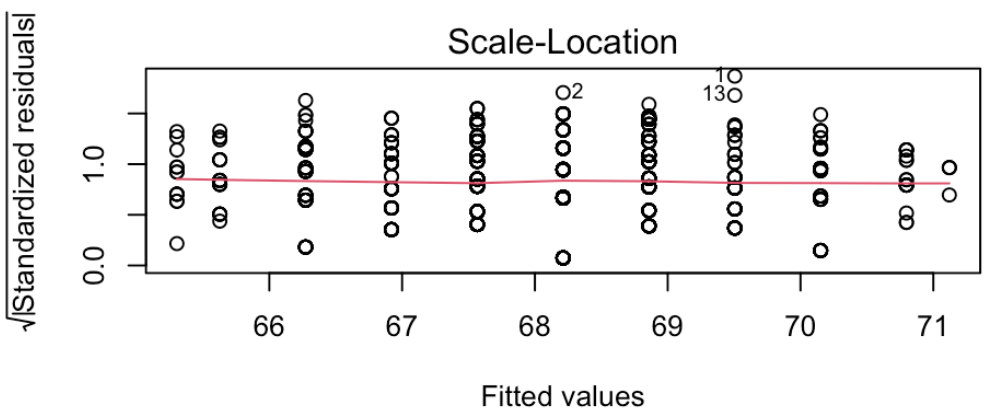
linearty | normality | equal variance | outliers

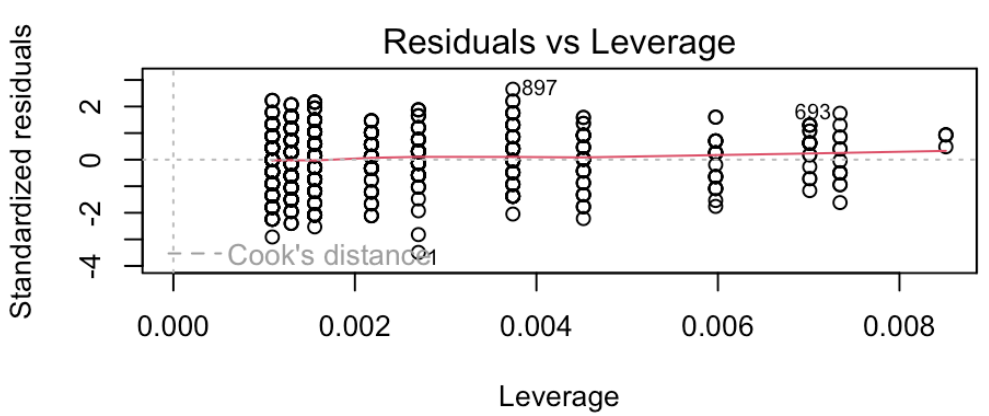
1-step

```r
par(mfrow = c(2, 2)) # need to do this to get 4 plots on one page
plot(fit)
```

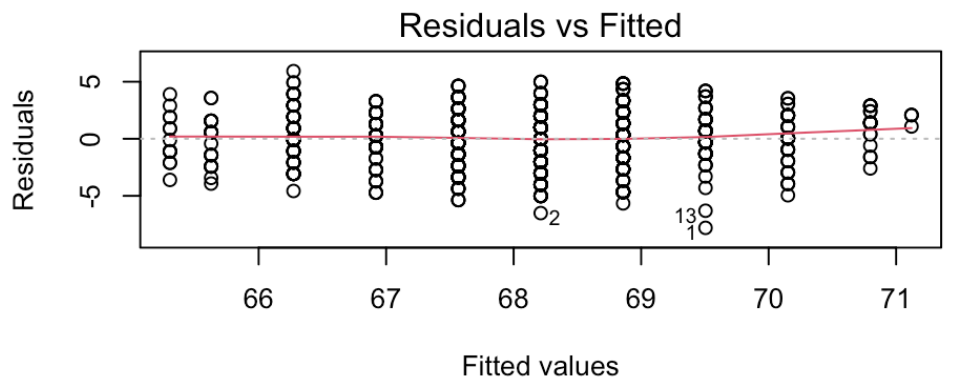


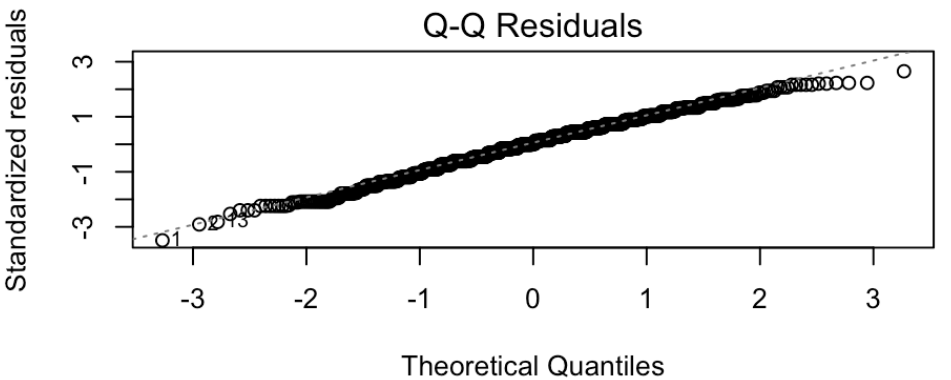


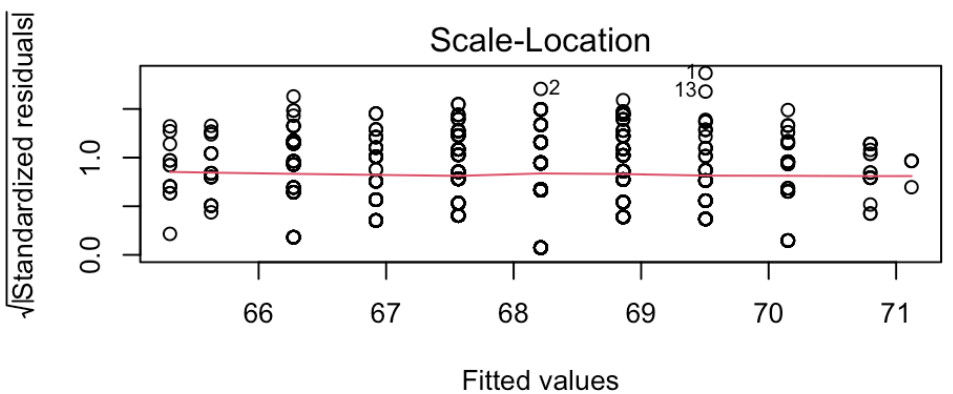


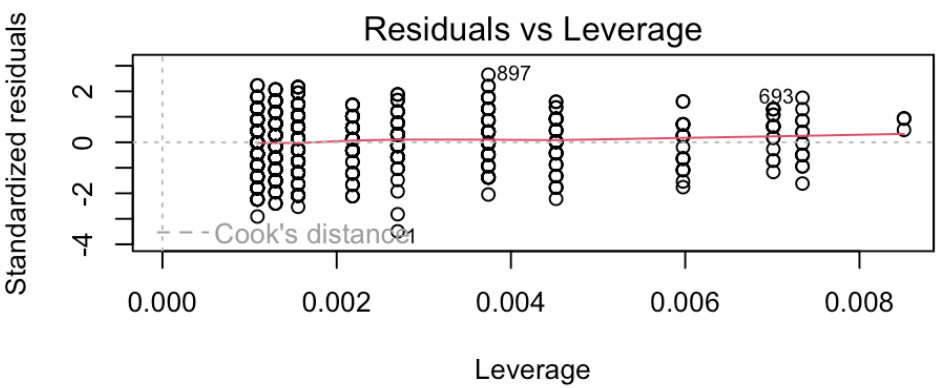


- **Residuals vs. Fitted**: check for linearity, equal variance.
- **Q-Q Residuals**: check for normality.
- **Scale-Location**: check for equal variance (standardised).
- **Residuals vs. Leverage**: check for outliers (influential points).





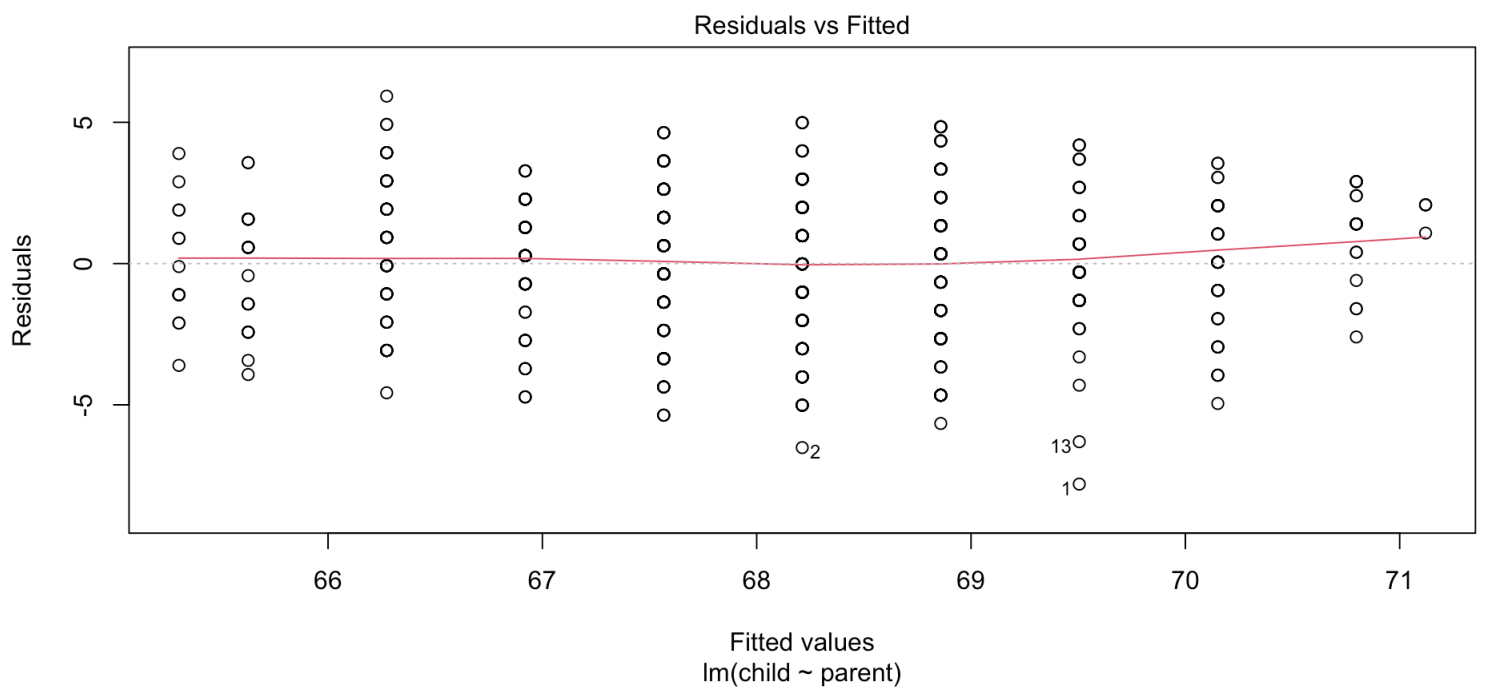




Assumption: Linearity

- Residuals vs. fitted plot looks at the relationship between the residuals and the fitted values.
- If the relationship is linear:
  - Residuals should be randomly scattered around the horizontal axis.
  - The red line should be reasonably straight.

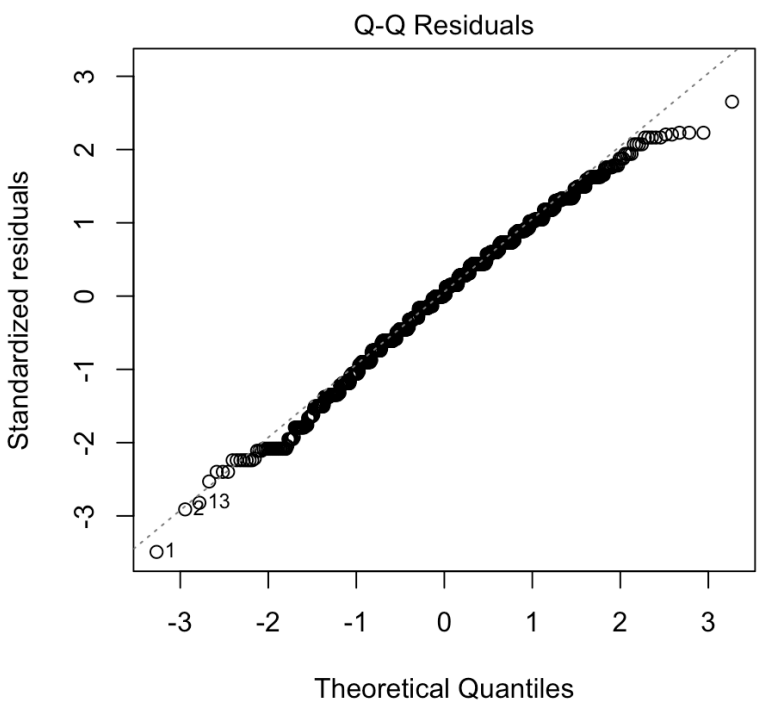
```
1  plot(fit, which = 1)
```

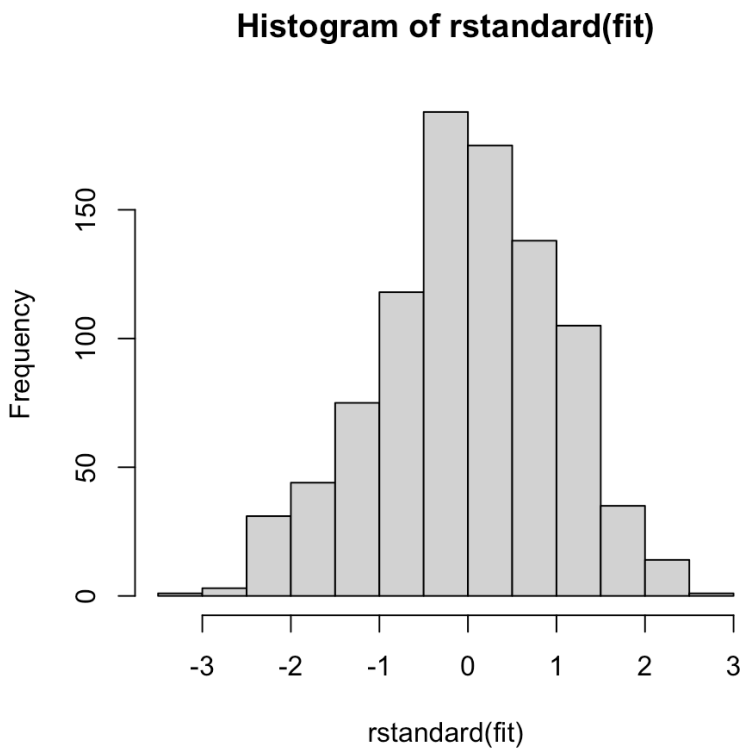


Assumption: Normality

- Q-Q plot looks at the distribution of the residuals against a normal distribution function (the dotted line).
- Sometimes, a histogram is still useful to see the shape of the distribution.

```
par(mfrow = c(1, 2))
plot(fit, which = 2)
hist(rstandard(fit))
```





Assumption: Normality

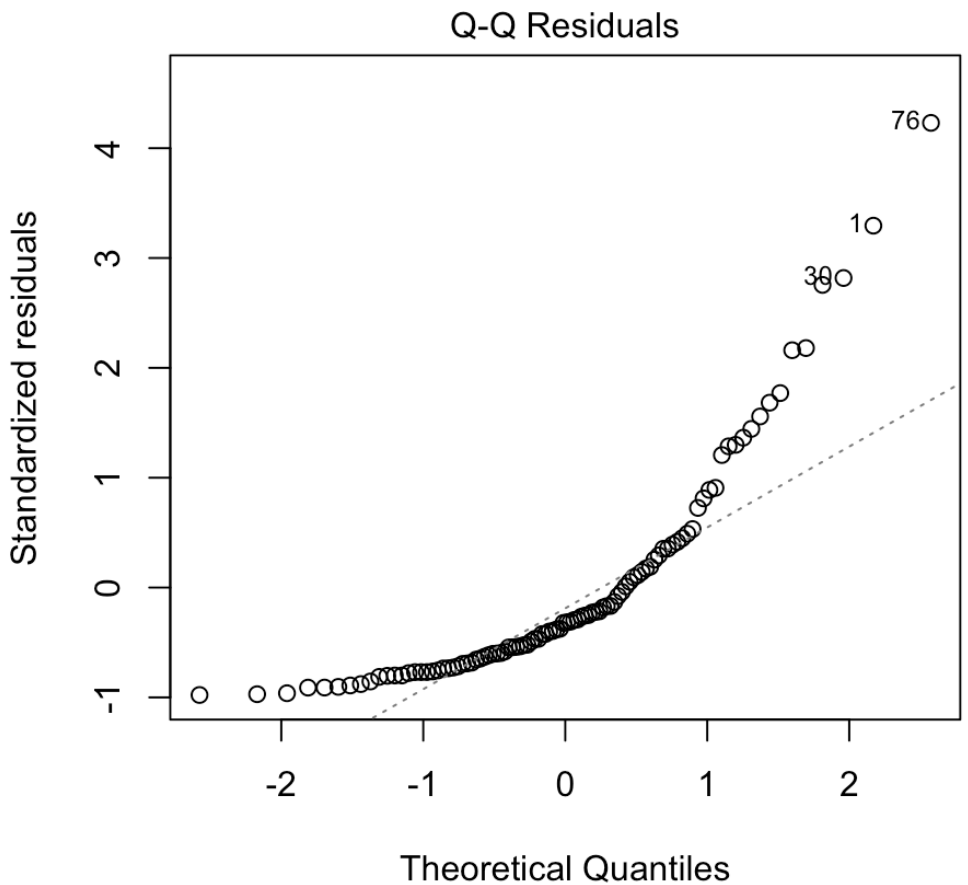
- If **normally distributed**, the points should follow the red line.
- Deviation from the red line is common in the tails (i.e. the ends), but not in the middle.

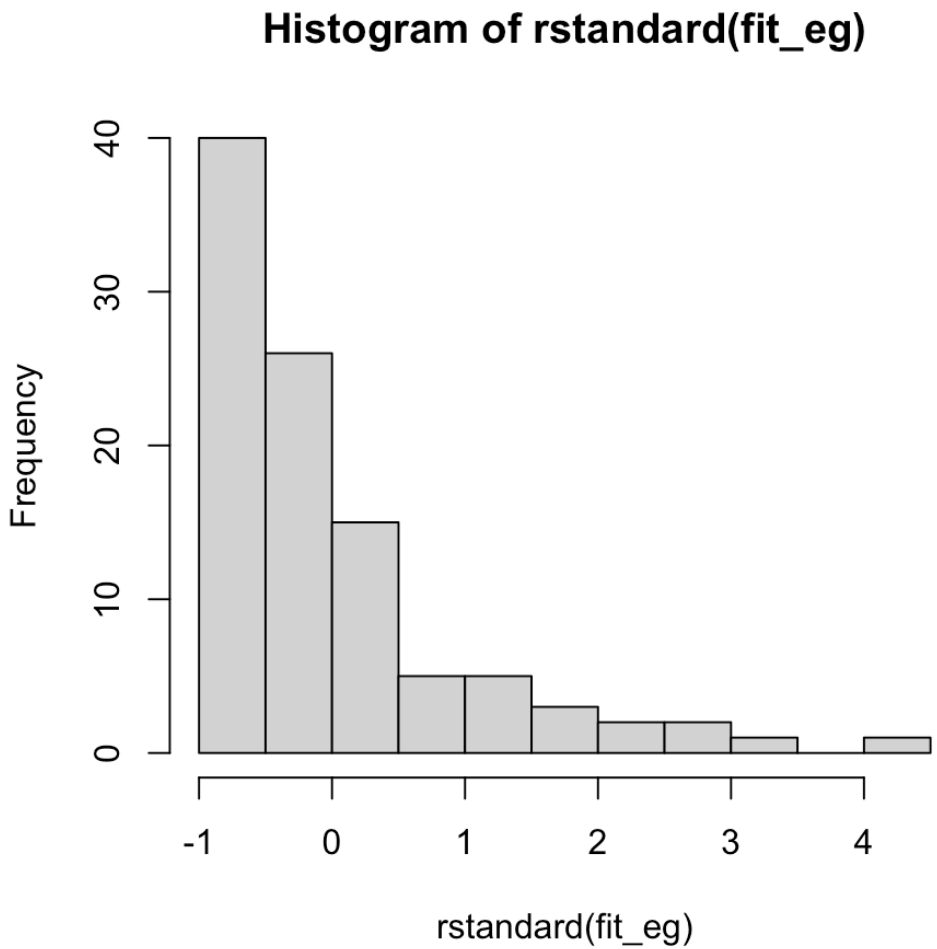
Tips

- **Light-tailed**: small variance in residuals, resulting in a narrow distribution.
- **Heavy-tailed**: many extreme positive and negative residuals, resulting in a wide distribution.
- **Left-skewed** (n shape): more data falls to the left of the mean.
- **Right-skewed** (u shape): more data falls to the right of the mean.

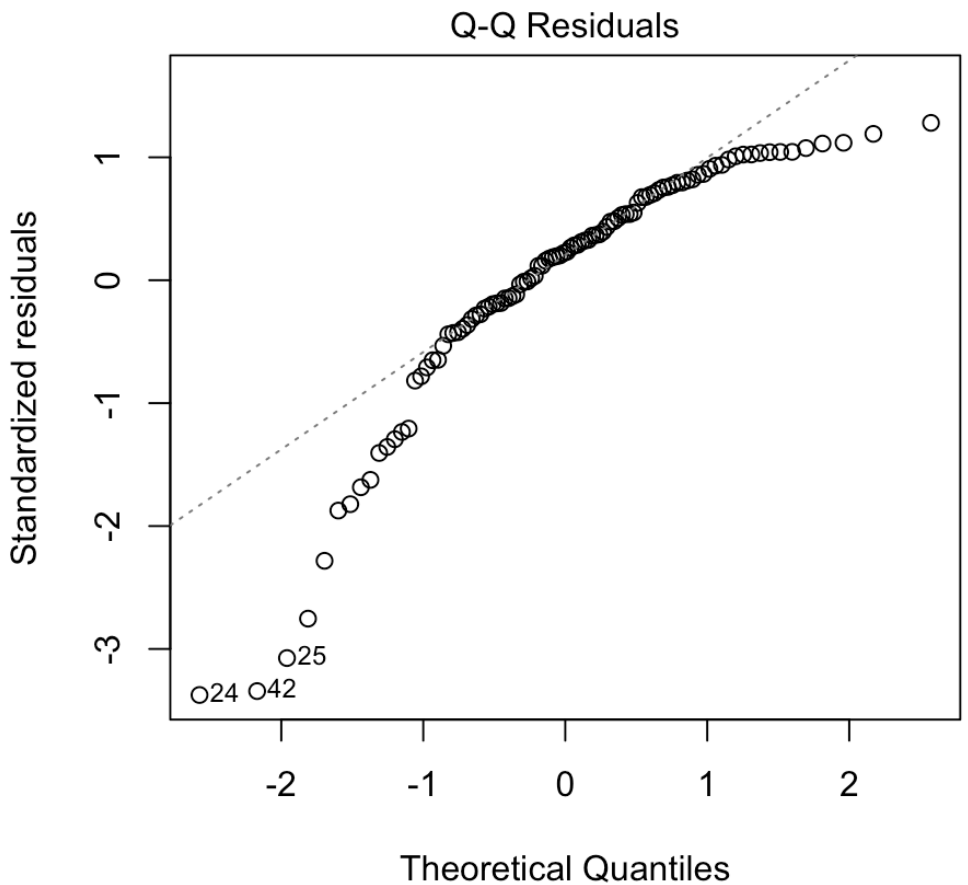
Practice

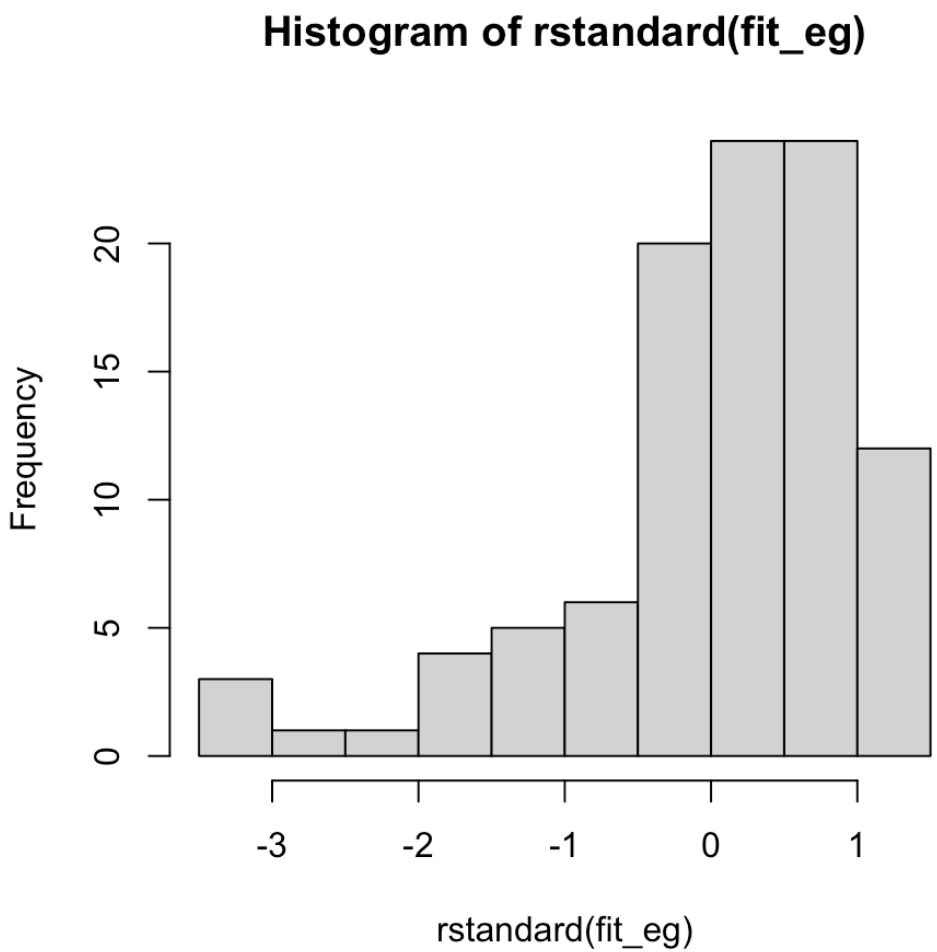
▶ Code


Q-Q Residuals
Standardized residuals
Theoretical Quantiles
76
1
30
-2
-1
0
1
2
-1
0
1
2
3
4

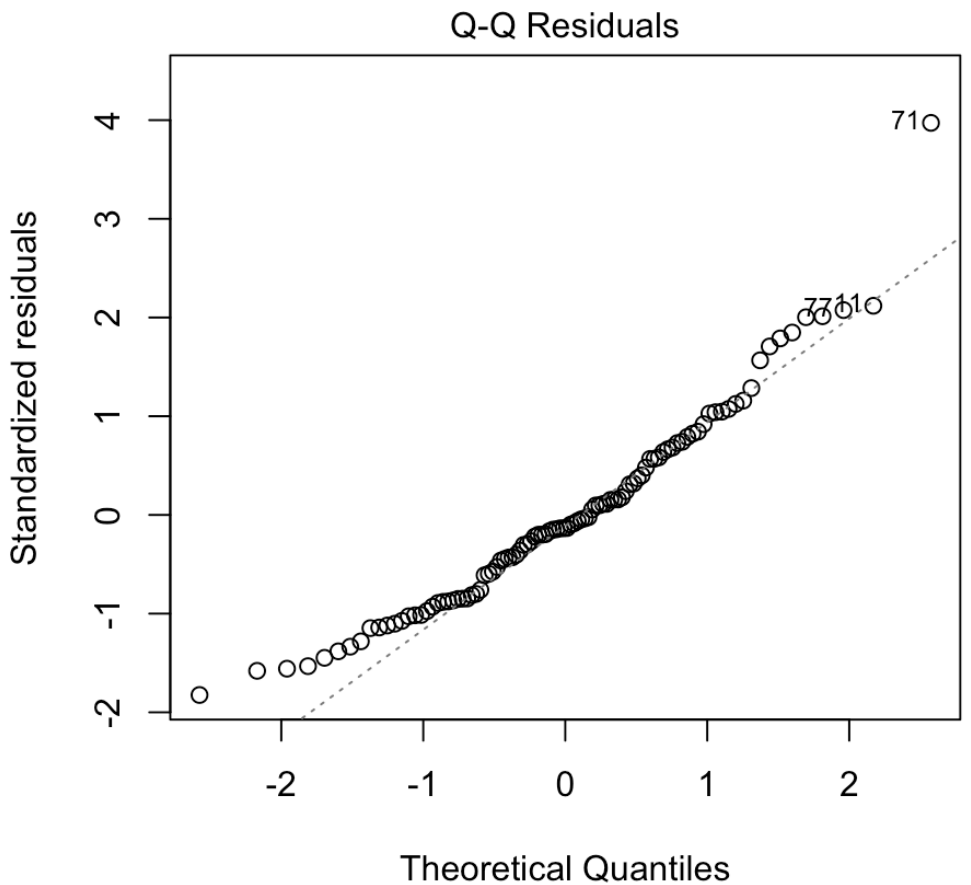

Histogram of rstandard(fit_eg)
Frequency
rstandard(fit_eg)
0
10
20
30
40
-1
0
1
2
3
4

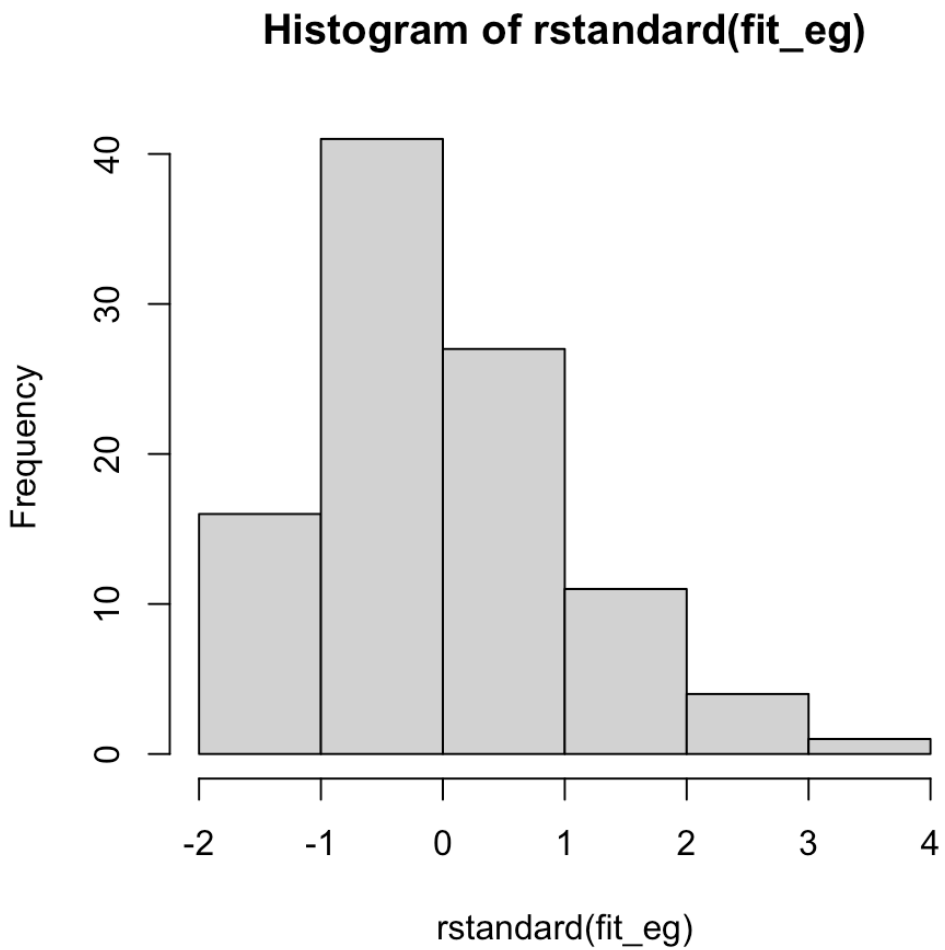
▶ Code


Q-Q Residuals
Standardized residuals
Theoretical Quantiles
-3
-2
-1
0
1
2
24
42
25


Histogram of rstandard(fit_eg)
Frequency
rstandard(fit_eg)
0
5
10
15
20
-3
-2
-1
0
1

▶ Code


Q-Q Residuals
Standardized residuals
Theoretical Quantiles
71
77
-2
-1
0
1
2
3
4

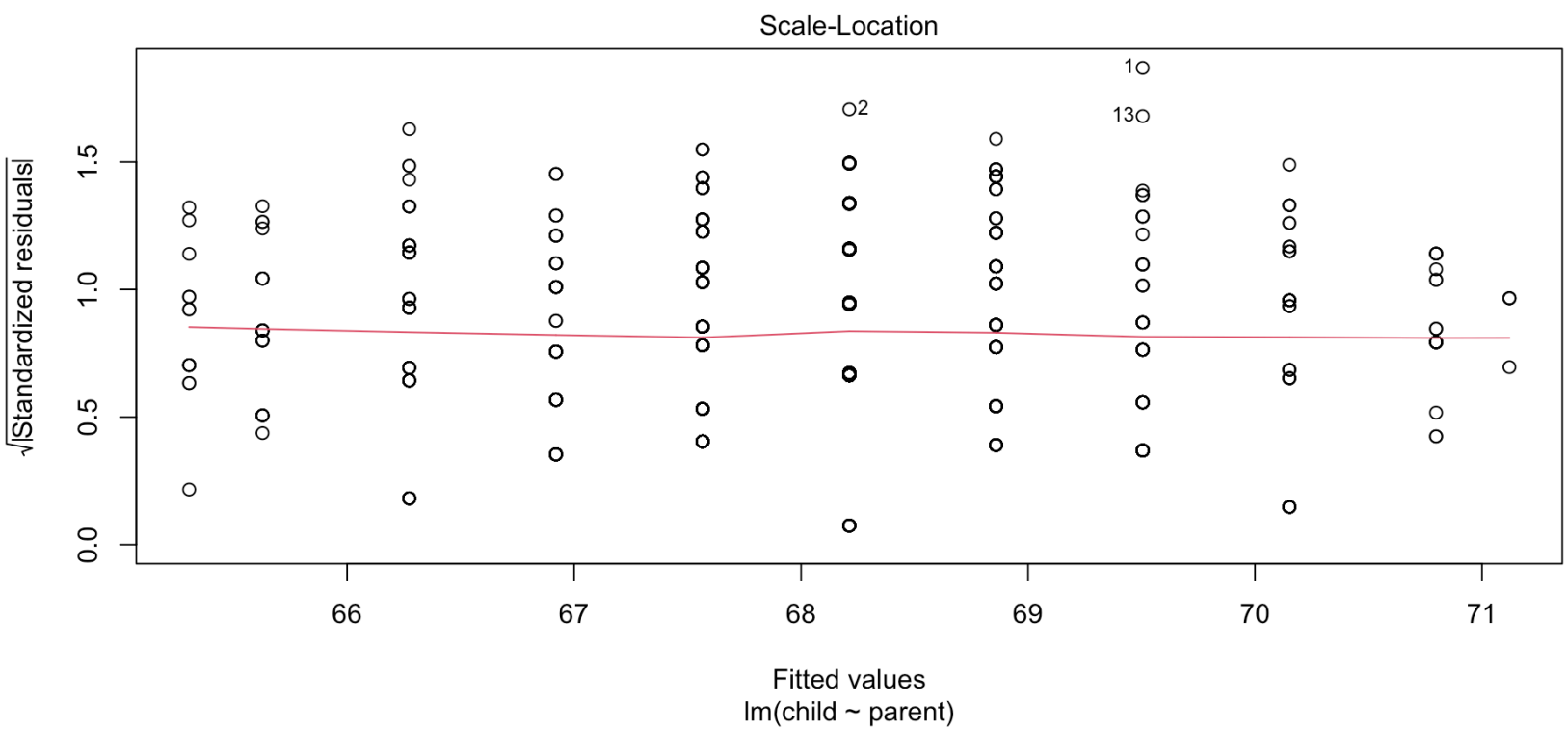

Histogram of rstandard(fit_eg)
Frequency
rstandard(fit_eg)
0
10
20
30
40
-2
-1
0
1
2
3
4

Assumption: Equal variances

Equal variances

- Look at the **scale-location plot**.
- If variances are equal, the points should be randomly scattered around the horizontal axis.
- The red line should be more or less horizontal.

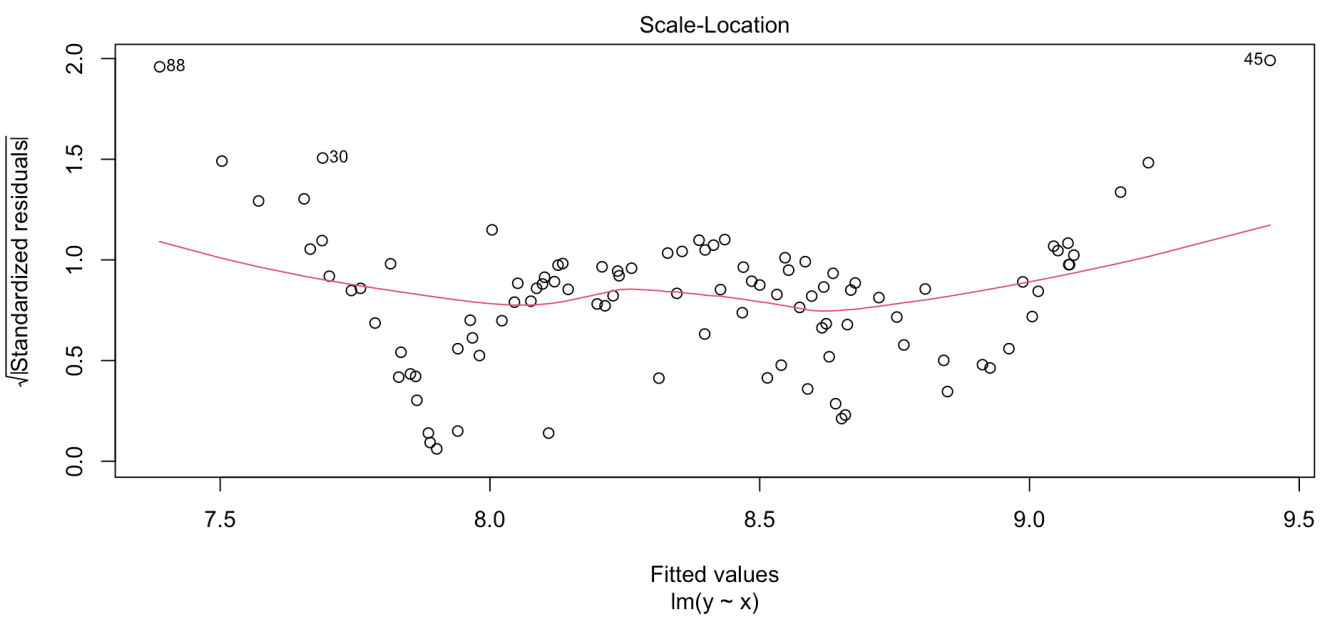
```
1  plot(fit, which = 3)
```



Equal variances

- If variances are not equal we *may* see:
  - A funnel shape, where the points are more spread out at the ends than in the middle. Sometimes also called “fanning”.
  - Patterns in the scale-location plot, such as a curve or a wave, indicating that the variance is changing.
- Look at the red line for a general trend, **but don’t depend on it too much**.

▶ Code

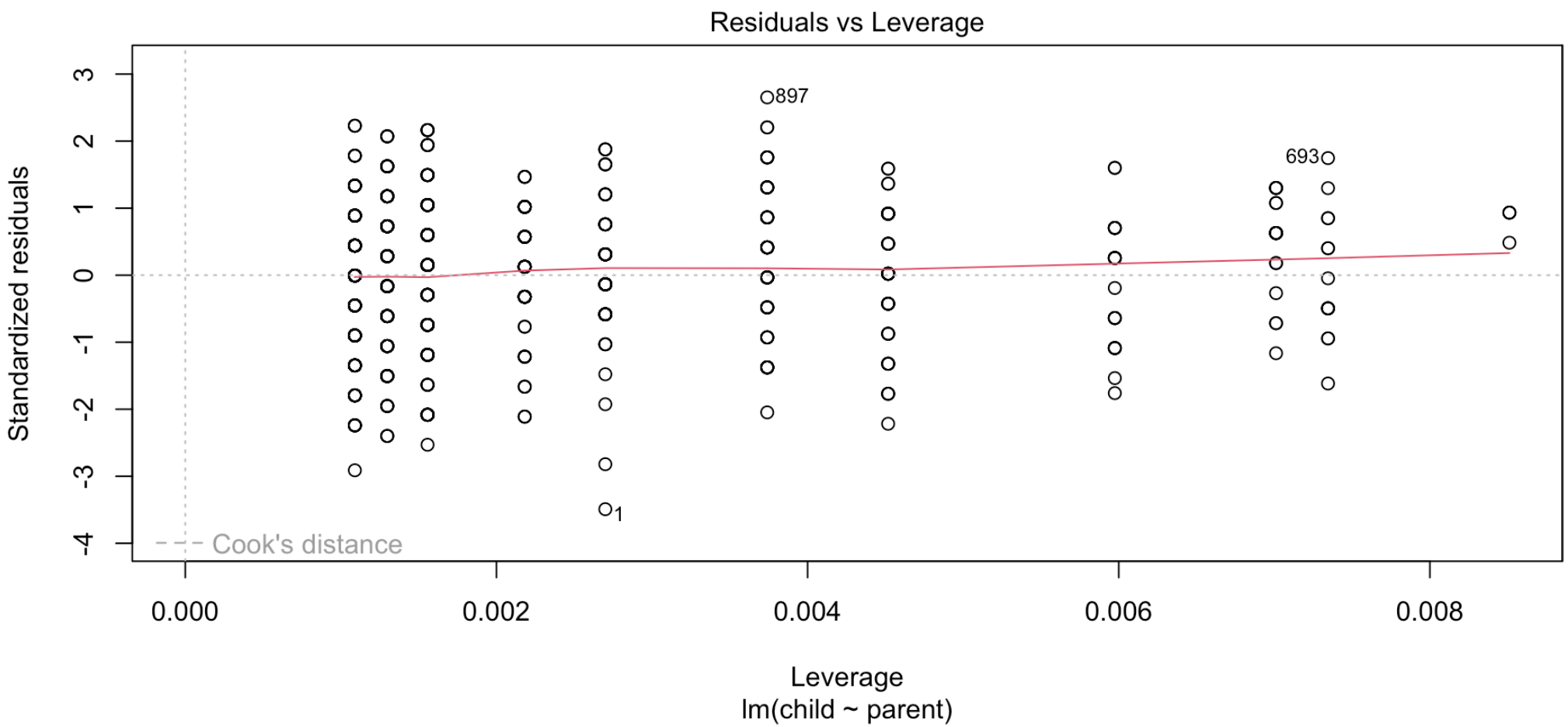


Outliers

- **Leverage** is a measure of how far away the predictor variable is from the mean of the predictor variable.
- The Residuals vs Leverage plot shows the relationship between the residuals and the leverage of each point.
- **Cook's distance** is a measure of how much the model would change if a point was removed.

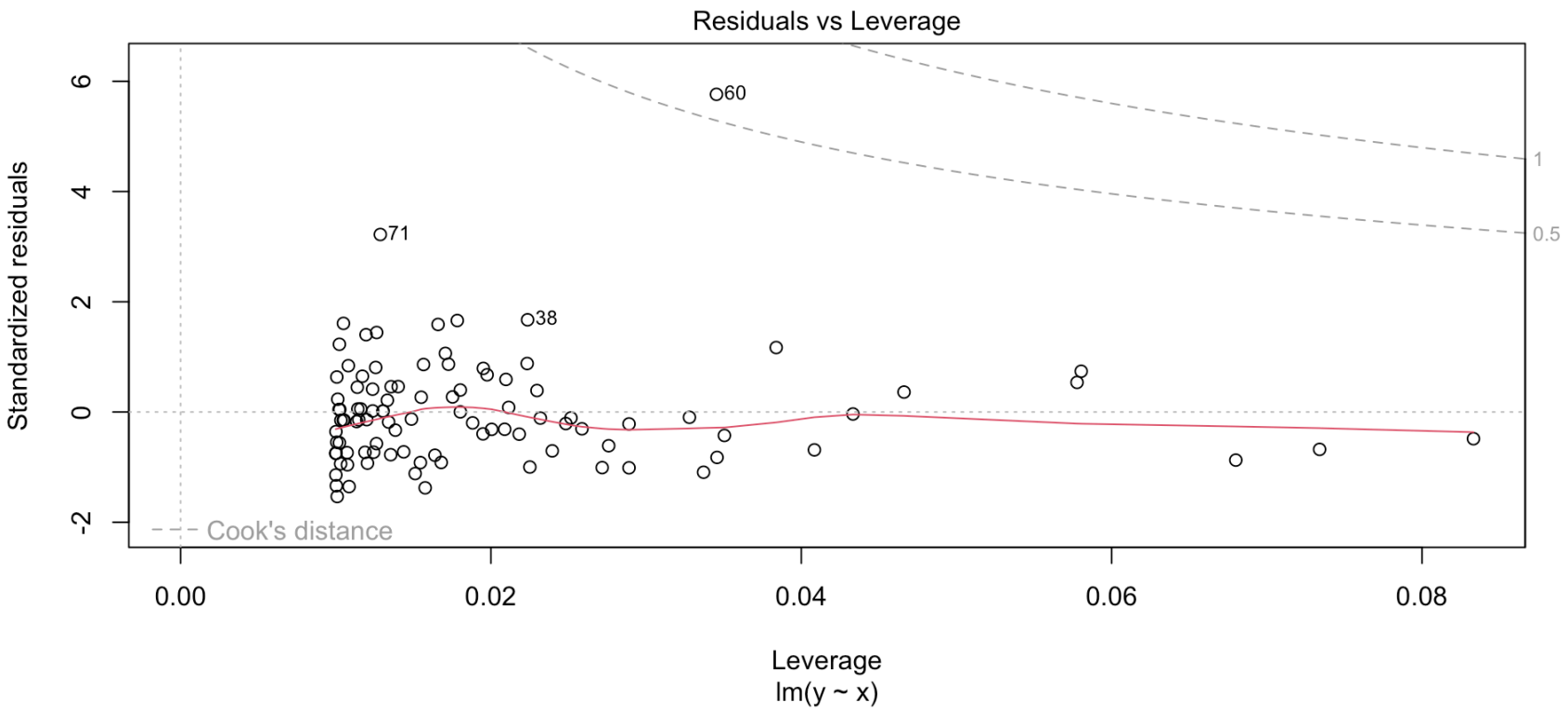
In general, points with **high leverage** and **high Cook's distance** are considered outliers.

```
1  plot(fit, which = 5)
```



Example of an influential outlier

▶ Code



We don’t want points to exceed the dashed line (which appears once they approach the Cook’s distance), because that means they are likely to influence the model greatly.

What can we do if the assumptions aren’t met?

It depends...

- Depends which assumption is not met and the type of data i.e. circumstances.
  - If data is **non-linear**, try a transformation of the response variable y, from light to extreme:
    - root: $\sqrt{y}$ or $\sqrt{y+1}$ if y contains zeros
    - log: $\log(y)$ or $\log(y+1)$ if y contains zeros
    - inverse: $\frac{1}{y}$ or $\frac{1}{y+1}$ if y contains zeros
- If data is **not normally distributed**, try a transformation of the response variable y first, otherwise transform the predictor variable x. Both can be done at the same time.
- If **equal variances** assumption is not met, same as above.
- If **outliers** are present, try removing them, or transforming the response variable y.

What if that doesn’t work?

- If the assumptions are still not met after trying the above, you can try:
  - Using a different model e.g. generalized linear model.
  - Using a different type of regression e.g. logistic regression.
  - Using a non-parametric test.

Model assumptions are met. Now what?

```r
summary(fit)
```

```
Call:
lm(formula = child ~ parent, data = Galton)

Residuals:
    Min      1Q  Median      3Q     Max
-7.8050 -1.3661  0.0487  1.6339  5.9264

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 23.94153    2.81088   8.517   <2e-16 ***
parent       0.64629    0.04114  15.711   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.239 on 926 degrees of freedom
Multiple R-squared:  0.2105,    Adjusted R-squared:  0.2096
F-statistic: 246.8 on 1 and 926 DF,  p-value: < 2.2e-16
```

Inference

What can we say about the model based on our data?

> What can we understand about the relationship between `child` and `parent`?

The model so far

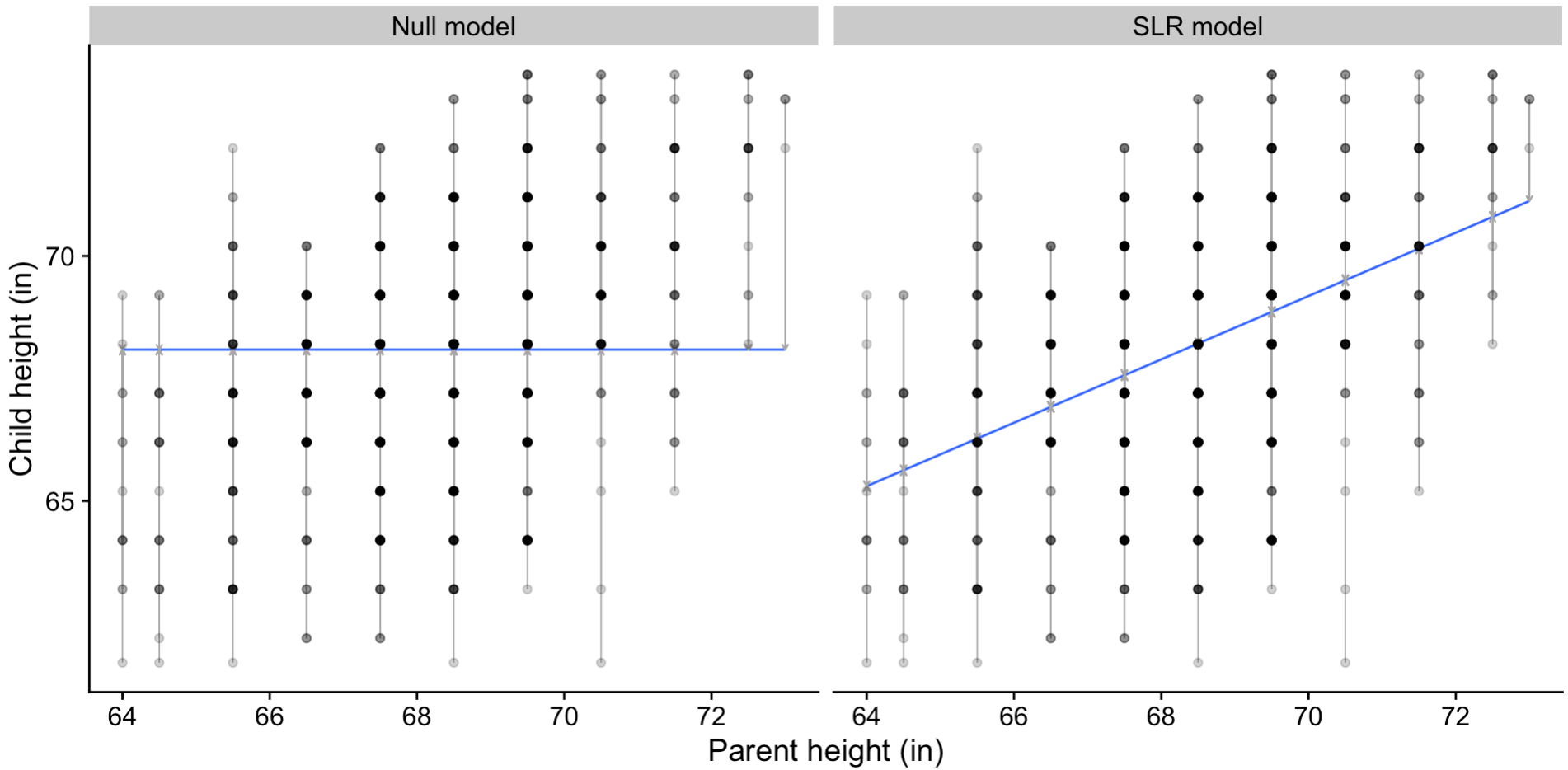
```r
library(HistData)
data(Galton)
fit <- lm(child ~ parent, data = Galton)
summary(fit)
```

```
Call:
lm(formula = child ~ parent, data = Galton)

Residuals:
    Min      1Q  Median      3Q     Max
-7.8050 -1.3661  0.0487  1.6339  5.9264

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 23.94153    2.81088   8.517   <2e-16 ***
parent       0.64629    0.04114  15.711   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.239 on 926 degrees of freedom
Multiple R-squared:  0.2105,    Adjusted R-squared:  0.2096
F-statistic: 246.8 on 1 and 926 DF,  p-value: < 2.2e-16
```

Hypothesis testing

How does our null ($H_0 : \beta_1 = 0$) model compare to the linear ($H_0 : \beta_1 \neq 0$) model?

▶ Code



What are we testing?

- The null model is a model with no predictors, i.e. $y = \beta_0 + \epsilon$
- The linear model is a model with one predictor, i.e. $y = \beta_0 + \beta_1 x + \epsilon$
- We use the t-test to compare the two models:

$$t = \frac{estimate - 0}{Standard\ error} = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)}$$

where $SE(\hat{\beta}_1)$ is the standard error of the slope estimate:

$$SE(\hat{\beta}_1) = \frac{\hat{\sigma}}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

Assesing the model

Interpreting the output

```
Call:
lm(formula = child ~ parent, data = Galton)

Residuals:
    Min      1Q  Median      3Q     Max
-7.8050 -1.3661  0.0487  1.6339  5.9264

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 23.94153    2.81088   8.517   <2e-16 ***
parent       0.64629    0.04114  15.711   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.239 on 926 degrees of freedom
Multiple R-squared:  0.2105,    Adjusted R-squared:  0.2096
F-statistic: 246.8 on 1 and 926 DF,  p-value: < 2.2e-16
```

- `Call`: the model formula

Interpreting the output

```
Call:
lm(formula = child ~ parent, data = Galton)

Residuals:
    Min      1Q  Median      3Q     Max
-7.8050 -1.3661  0.0487  1.6339  5.9264

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 23.94153    2.81088   8.517   <2e-16 ***
parent       0.64629    0.04114  15.711   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.239 on 926 degrees of freedom
Multiple R-squared:  0.2105,    Adjusted R-squared:  0.2096
F-statistic: 246.8 on 1 and 926 DF,  p-value: < 2.2e-16
```

- `Residuals`: distribution of the residuals

Interpreting the output

```
Call:
lm(formula = child ~ parent, data = Galton)

Residuals:
    Min      1Q  Median      3Q     Max
-7.8050 -1.3661  0.0487  1.6339  5.9264

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 23.94153    2.81088   8.517   <2e-16 ***
parent       0.64629    0.04114  15.711   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.239 on 926 degrees of freedom
Multiple R-squared:  0.2105,    Adjusted R-squared:  0.2096
F-statistic: 246.8 on 1 and 926 DF,  p-value: < 2.2e-16
```

- `Coefficients`: a summary table of the coefficients, their standard errors, t-values, and p-values addressing the hypothesis that the coefficient is 0

Interpreting the output

```
Call:
lm(formula = child ~ parent, data = Galton)

Residuals:
    Min      1Q  Median      3Q     Max
-7.8050 -1.3661  0.0487  1.6339  5.9264

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 23.94153    2.81088   8.517   <2e-16 ***
parent       0.64629    0.04114  15.711   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.239 on 926 degrees of freedom
Multiple R-squared:  0.2105,    Adjusted R-squared:  0.2096
F-statistic: 246.8 on 1 and 926 DF,  p-value: < 2.2e-16
```

Interpreting the output

```
Call:
lm(formula = child ~ parent, data = Galton)

Residuals:
    Min      1Q  Median      3Q     Max
-7.8050 -1.3661  0.0487  1.6339  5.9264

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 23.94153    2.81088   8.517   <2e-16 ***
parent       0.64629    0.04114  15.711   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.239 on 926 degrees of freedom
Multiple R-squared:  0.2105,    Adjusted R-squared:  0.2096
F-statistic: 246.8 on 1 and 926 DF,  p-value: < 2.2e-16
```

- We can also use the `Estimate` values to write the equation of the regression line:

$$\widehat{child} = 23.94153 + 0.64629 \cdot parent$$

- For every one-inch increase in the parent height, the child height is predicted to increase by 0.64629 inches.

Interpreting the output

```
Call:
lm(formula = child ~ parent, data = Galton)

Residuals:
    Min      1Q  Median      3Q     Max
-7.8050 -1.3661  0.0487  1.6339  5.9264

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 23.94153    2.81088   8.517   <2e-16 ***
parent       0.64629    0.04114  15.711   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.239 on 926 degrees of freedom
Multiple R-squared:  0.2105,    Adjusted R-squared:  0.2096
F-statistic: 246.8 on 1 and 926 DF,  p-value: < 2.2e-16
```

- `Residual standard error`: the standard deviation of the residuals.
  - ➡ Interpretation: the average amount that the response will *deviate* from the true regression line.
- `degrees of freedom`: the number of observations minus the number of parameters being estimated. Used in hypothesis testing and calculating the standard error of the regression coefficients.
  - ➡ Can estimate sample size from this number.

Interpreting the output

```
Call:
lm(formula = child ~ parent, data = Galton)

Residuals:
    Min      1Q  Median      3Q     Max
-7.8050 -1.3661  0.0487  1.6339  5.9264

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 23.94153    2.81088   8.517   <2e-16 ***
parent       0.64629    0.04114  15.711   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.239 on 926 degrees of freedom
Multiple R-squared:  0.2105,    Adjusted R-squared:  0.2096
F-statistic: 246.8 on 1 and 926 DF,  p-value: < 2.2e-16
```

Interpreting the output

```
Call:
lm(formula = child ~ parent, data = Galton)

Residuals:
    Min      1Q  Median      3Q     Max
-7.8050 -1.3661  0.0487  1.6339  5.9264

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 23.94153    2.81088   8.517   <2e-16 ***
parent       0.64629    0.04114  15.711   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.239 on 926 degrees of freedom
Multiple R-squared:  0.2105,    Adjusted R-squared:  0.2096
F-statistic: 246.8 on 1 and 926 DF,  p-value: < 2.2e-16
```

- `F-statistic`: the ratio of the variance of the regression model to the variance of the residuals.
  - ➦ Also known as the partial F-test between the full model and the intercept-only (null) model.
- `p-value`: for the linear model, the p-value is the probability that the F-statistic is greater than the observed value under the null hypothesis.
  - ➦ A significant p-value indicates that the linear model is a better fit than the intercept-only model.

Reporting

Two methods

Using ANOVA

`anova(fit)`

```r
fit <- lm(formula = child ~ parent, data = Galton)
anova(fit)
```

```
Analysis of Variance Table

Response: child
           Df Sum Sq Mean Sq F value    Pr(>F)
parent      1 1236.9 1236.93  246.84 < 2.2e-16 ***
Residuals 926 4640.3    5.01
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.'
0.1 ' ' 1
```

Using Regression

`summary(fit)`

```r
summary(fit)
```

```
Call:
lm(formula = child ~ parent, data = Galton)

Residuals:
    Min      1Q  Median      3Q     Max
-7.8050 -1.3661  0.0487  1.6339  5.9264

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 23.94153    2.81088   8.517   <2e-16 ***
parent       0.64629    0.04114  15.711   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.'
0.1 ' ' 1

Residual standard error: 2.239 on 926 degrees of
freedom
Multiple R-squared:  0.2105,    Adjusted R-squared:
0.2096
F-statistic: 246.8 on 1 and 926 DF,  p-value: < 2.2e-
16
```

Two methods

Using ANOVA

> The ANOVA suggests that the main effect of parent is statistically significant and large (F(1, 926) = 246.84, p < .001)

Using Regression

> We fitted a linear model (estimated using OLS) to predict child with parent (formula: child ~ parent). The model explains a statistically significant and moderate proportion of variance (R^2 = 0.21, F(1, 926) = 246.84, p < .001, adj. R^2 = 0.21). Within this model, the effect of parent is statistically significant and positive (β = 0.65, t(926) = 15.71, p < .001).

Tip

For **simple linear models**, `summary()` provides more information than `anova()`, but the results are the same.

Let’s practice

Can we predict the weight of an alligator from its length? Download data ⬇

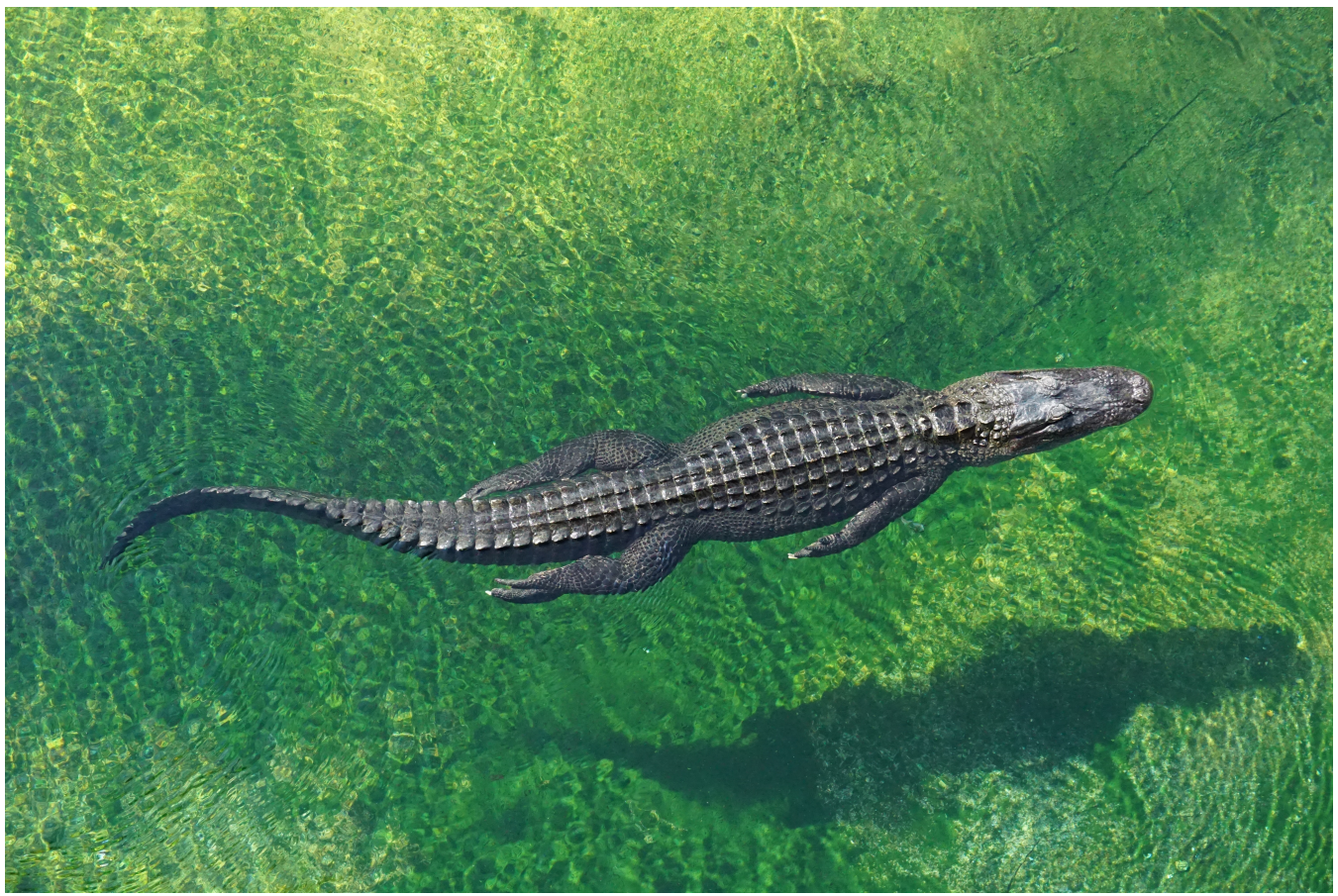


Photo by Shelly Collins

Explore

Read the data:

```
library(readxl) # load the readxl package

alligator <- read_excel(path = "data/ENVX1002_Lecture_wk10_data.xlsx",
  sheet = "Alligator") # read in the data
```

What does the data look like?

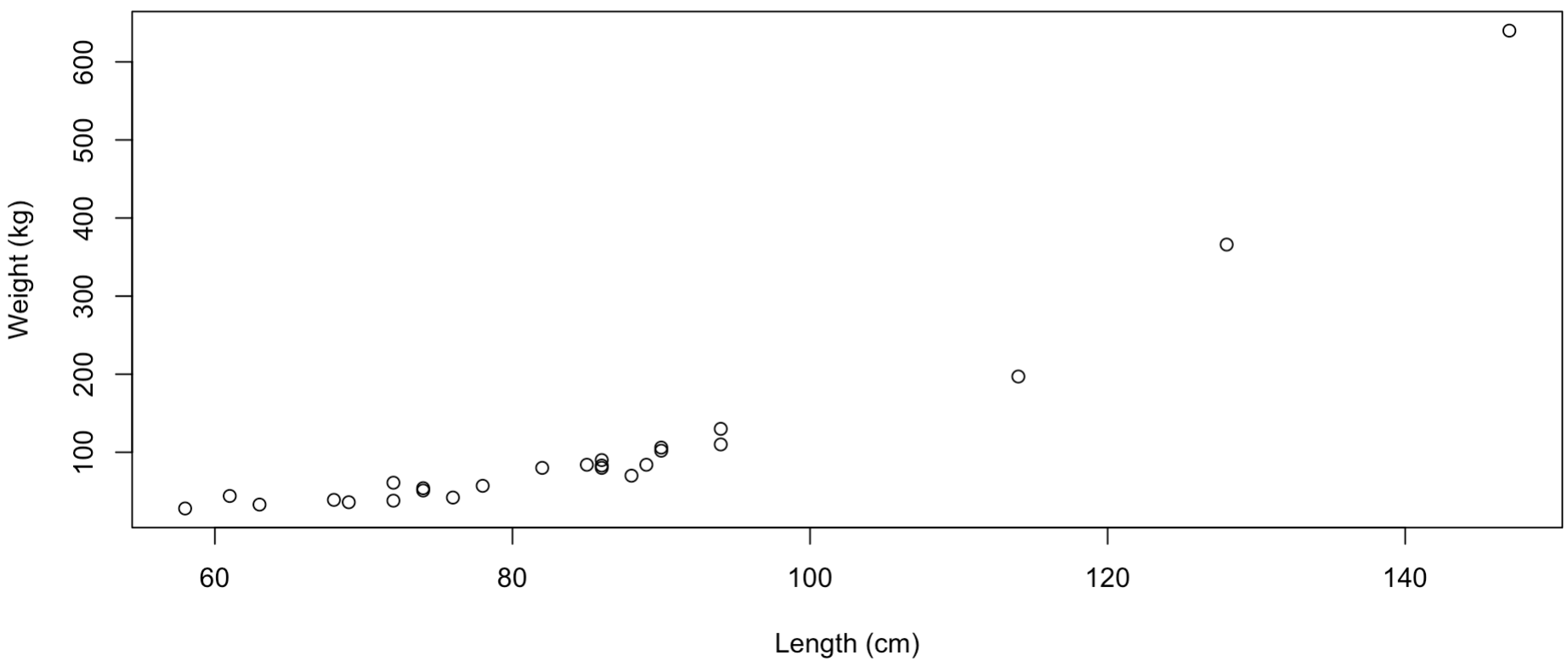
```
str(alligator)
```

```
tibble [25 × 2] (S3: tbl_df/tbl/data.frame)
 $ Length: num [1:25] 58 61 63 68 69 72 72 74 74 76 ...
 $ Weight: num [1:25] 28 44 33 39 36 38 61 54 51 42 ...
```

Plot

Using base R Using `ggplot2`

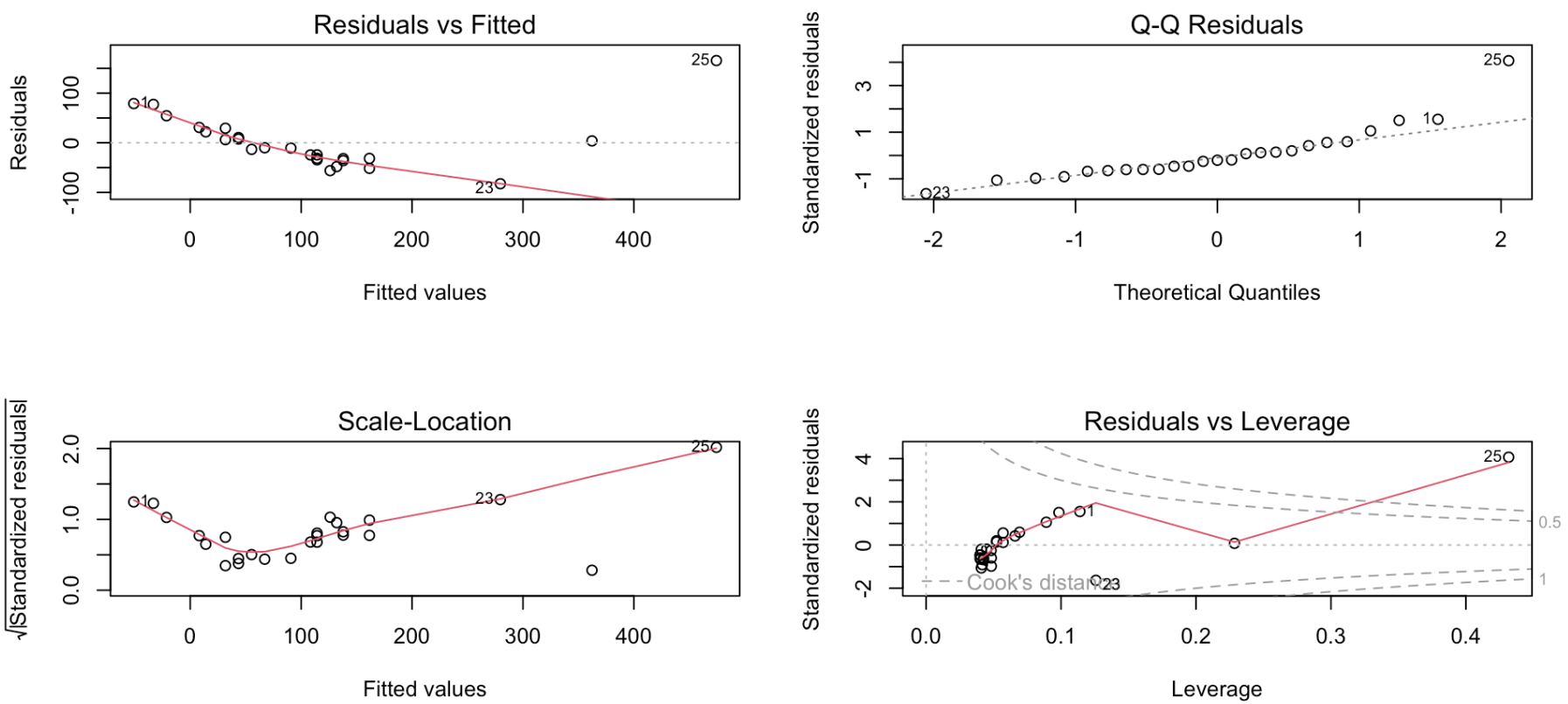
```
plot(x = alligator$Length, y = alligator$Weight,
  xlab = "Length (cm)", ylab = "Weight (kg)")
```



Plot residual diagnostics

To check assumptions, we need to fit the model first, then plot the model.

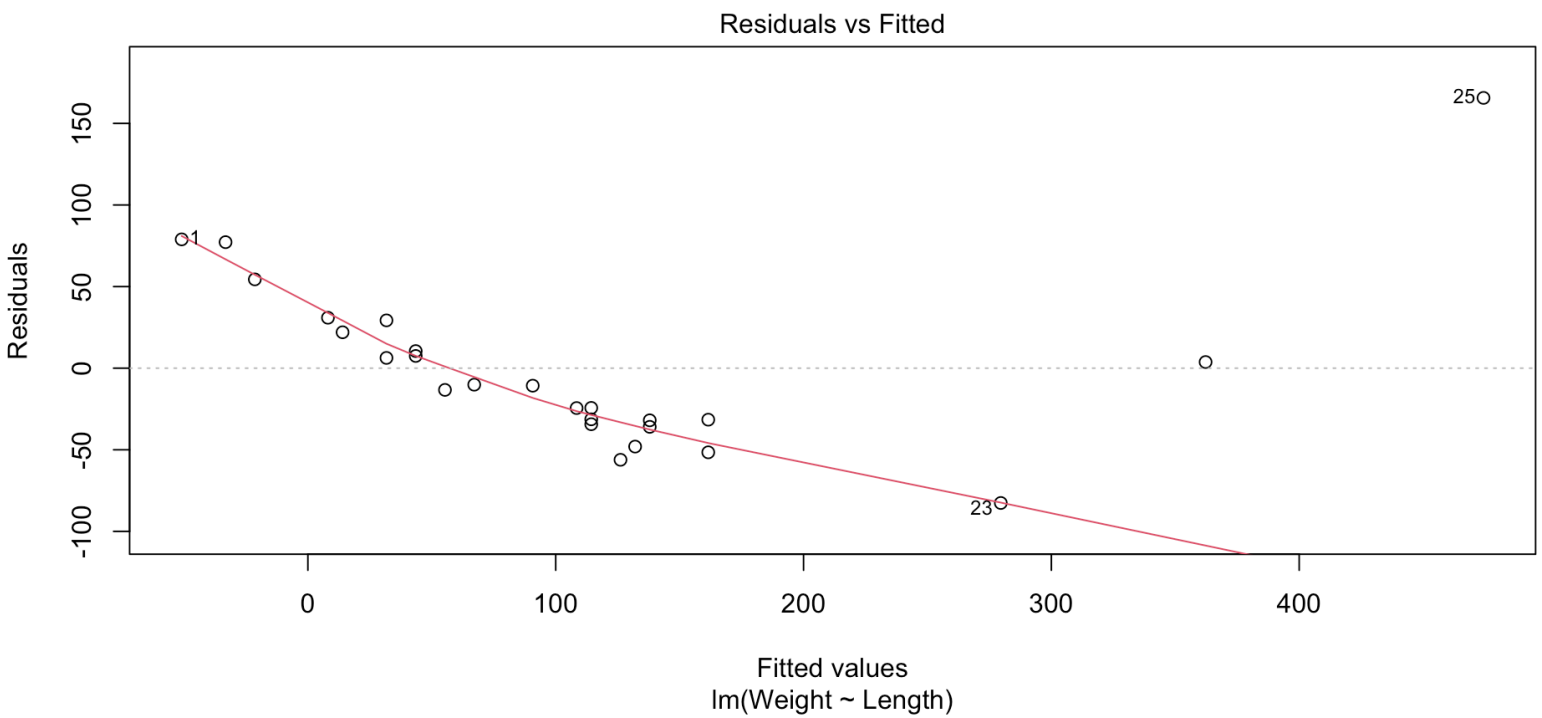
```
fit <- lm(formula = Weight ~ Length, data = alligator)
par(mfrow = c(2, 2)) # set up a 2 x 2 grid for plots
plot(fit)
```



Check assumptions

Is the relationship linear?

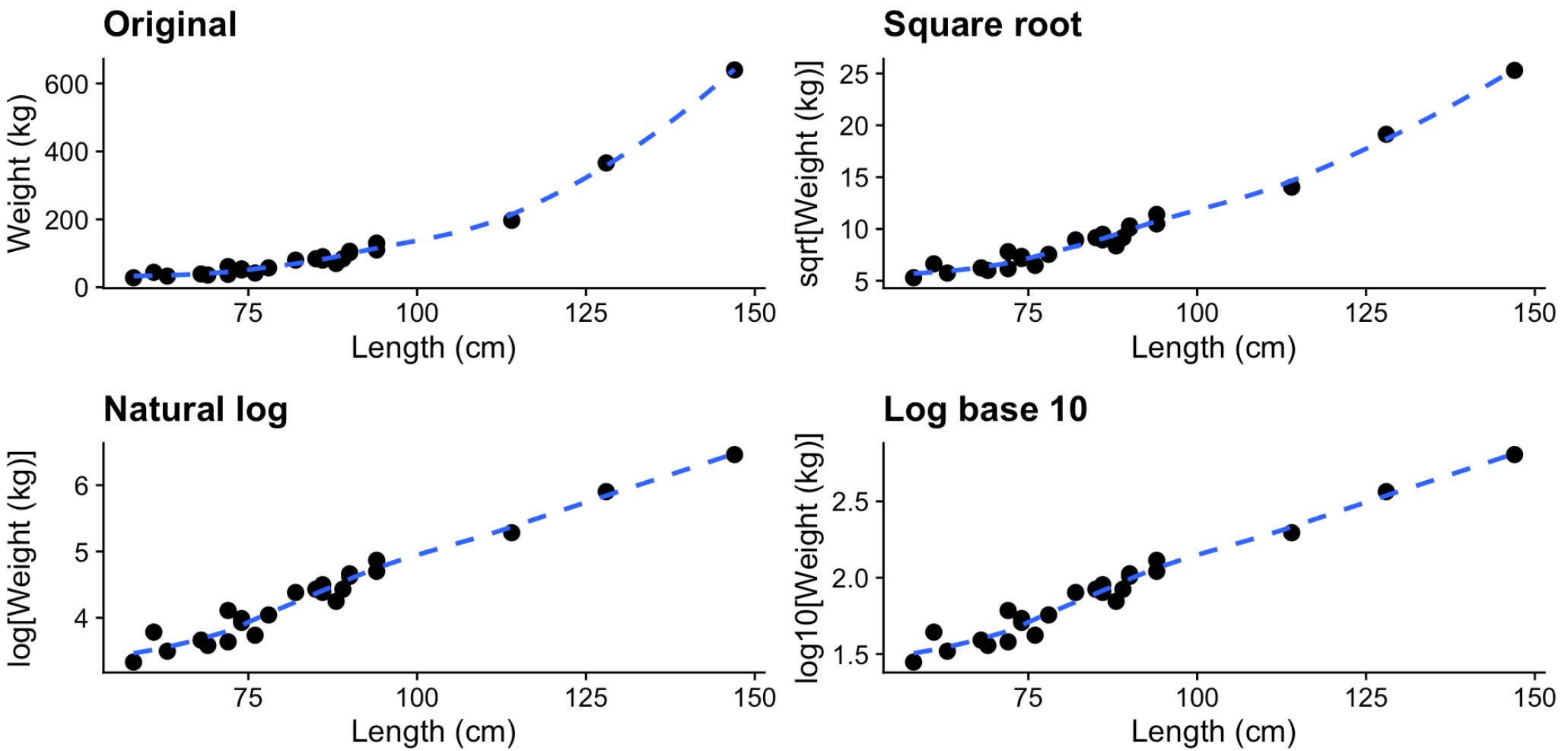
```
1  plot(fit, which = 1)
```



If the linearity assumption is not met, there is no reason to validate the model since it is no longer suitable for the data.

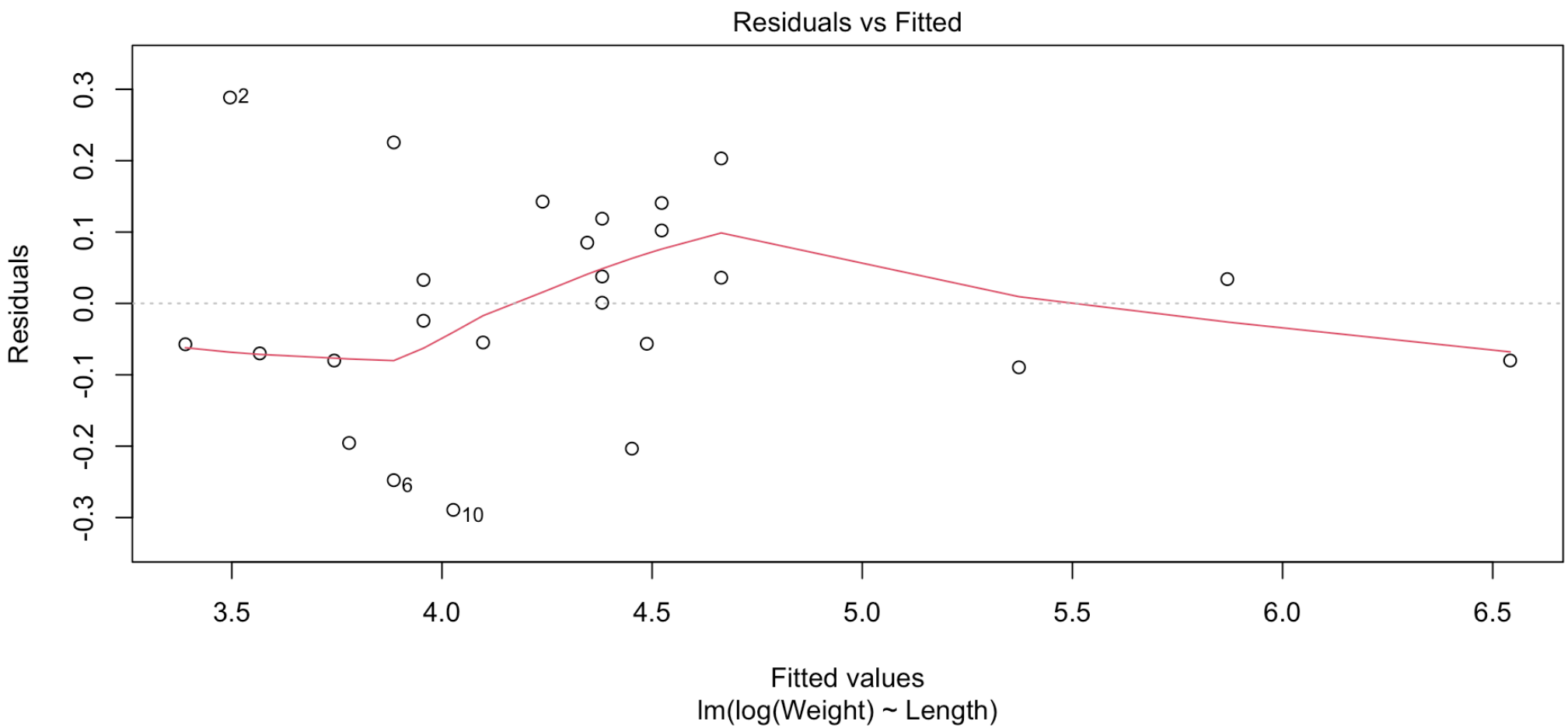
Dealing with non-linearity: transform the data

▶ Code



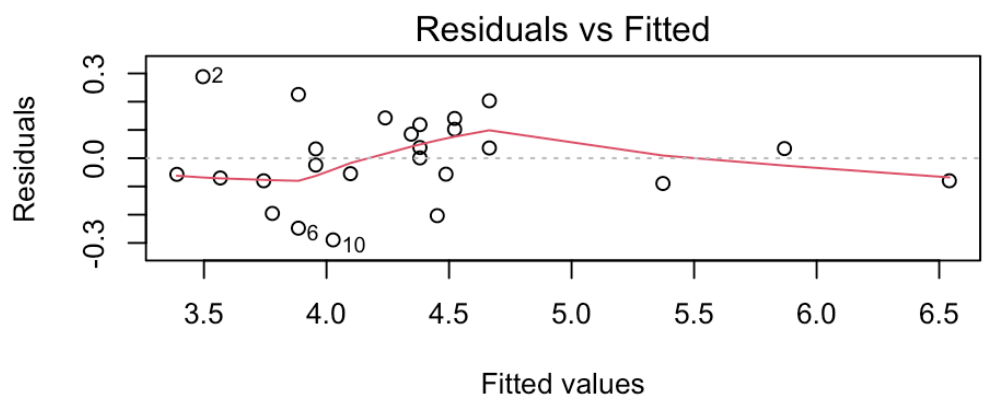
Natural log transformation

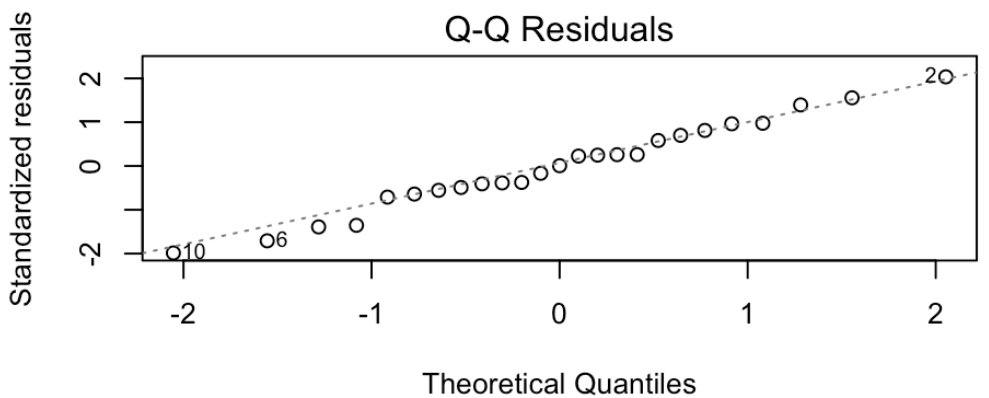
```
1 fit <- lm(formula = log(Weight) ~ Length, data = alligator)
2 plot(fit, which = 1)
```

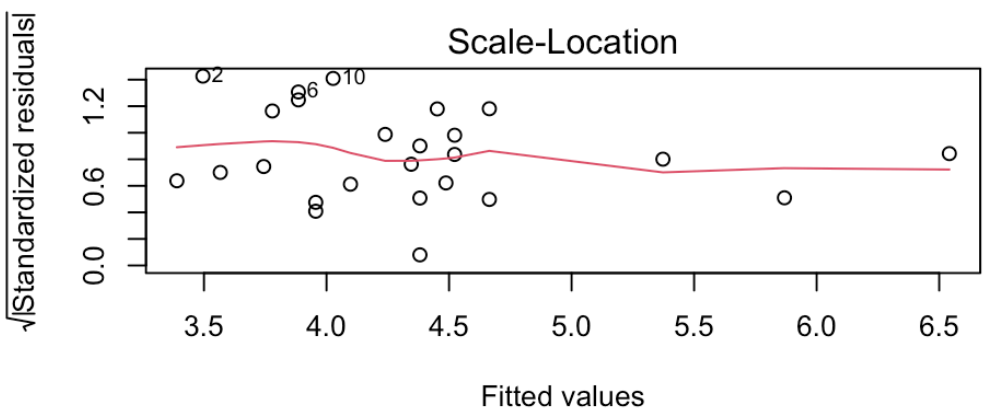


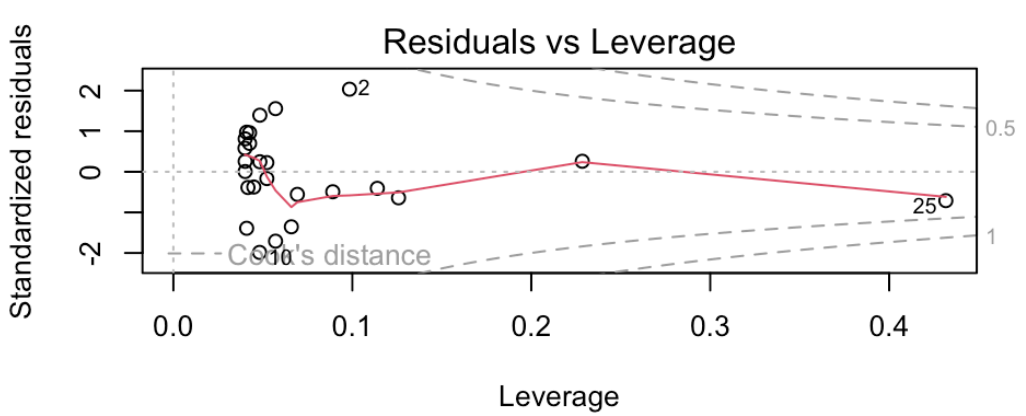
Natural log transformation – Check assumptions again

```r
par(mfrow = c(2, 2)) # set up a 2 x 2 grid for plots
plot(fit)
```









Interpretation

```
1  summary(fit)
```

```
Call:
lm(formula = log(Weight) ~ Length, data = alligator)

Residuals:
      Min        1Q    Median        3Q       Max
-0.289266 -0.079989  0.000933  0.102216  0.288491

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.335335   0.131394   10.16 5.63e-10 ***
Length      0.035416   0.001506   23.52  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1493 on 23 degrees of freedom
Multiple R-squared:  0.9601,    Adjusted R-squared:  0.9583
F-statistic:   553 on 1 and 23 DF,  p-value: < 2.2e-16
```

- `Length` is a statistically significant predictor of `log(Weight)` ($p < .001$).
- The model explains a statistically significant and large proportion (96%) of variance ($R^2 = 0.96$, $F(1, 23) = 553$, $p < .001$)
- For every 1 cm increase in `Length`, `log(Weight)` increases by 0.0354.
    - ➠ *Or, for every 1 cm increase in `Length`, percent increase in `Weight` is 3.54% (only works when transforming using natural log).*

Summary

You should know the workflow by now

1. Explore
2. Plot
3. Fit model and plot residual diagnostics
4. Check assumptions, transform data if necessary. Go back to step 3.
5. Interpret

Thanks!