Simple linear regression

Johan A. Elkink

University College Dublin

30 September 2013

Outline

Example

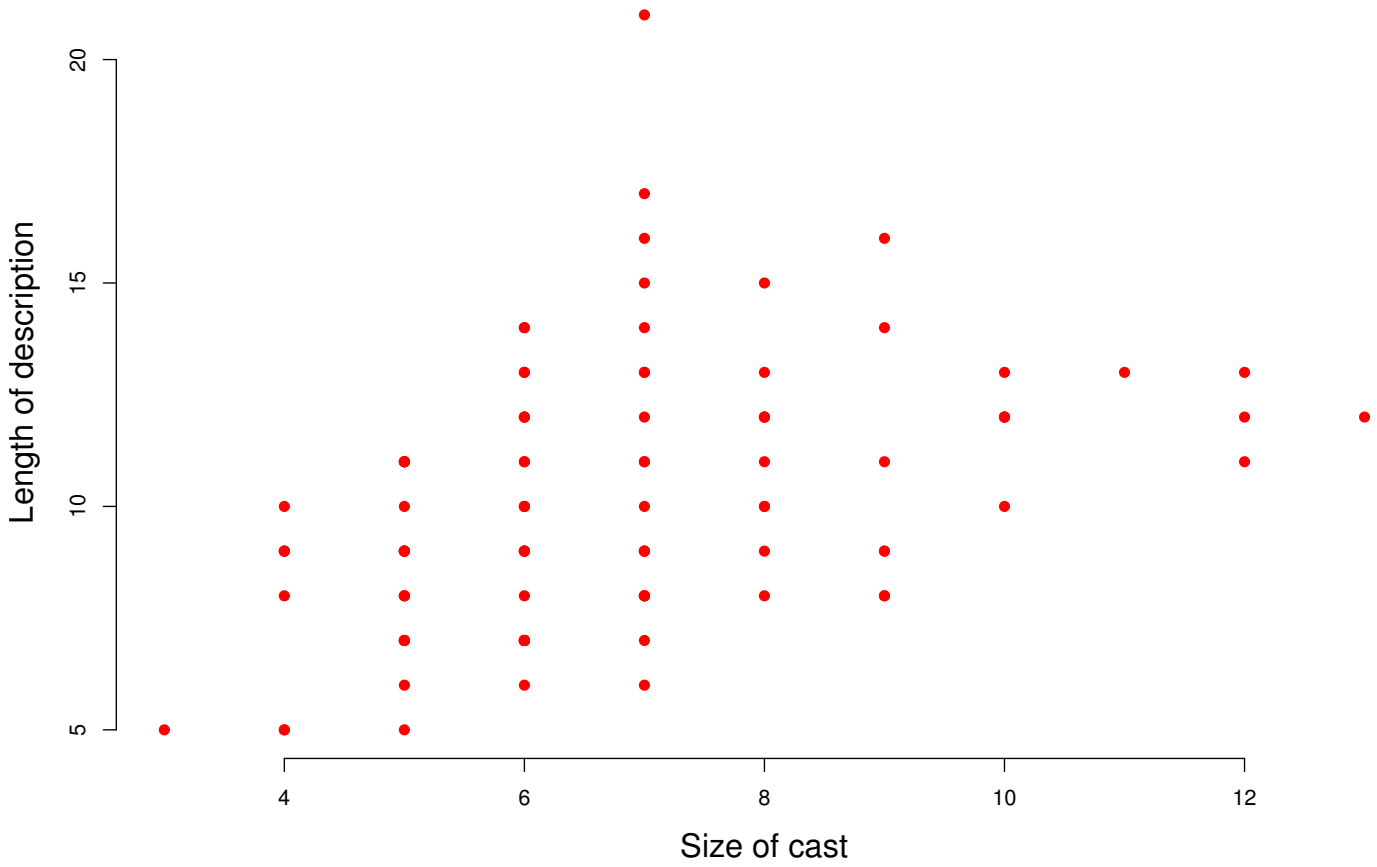
We will make use of the `films.dta` data file again.

Example

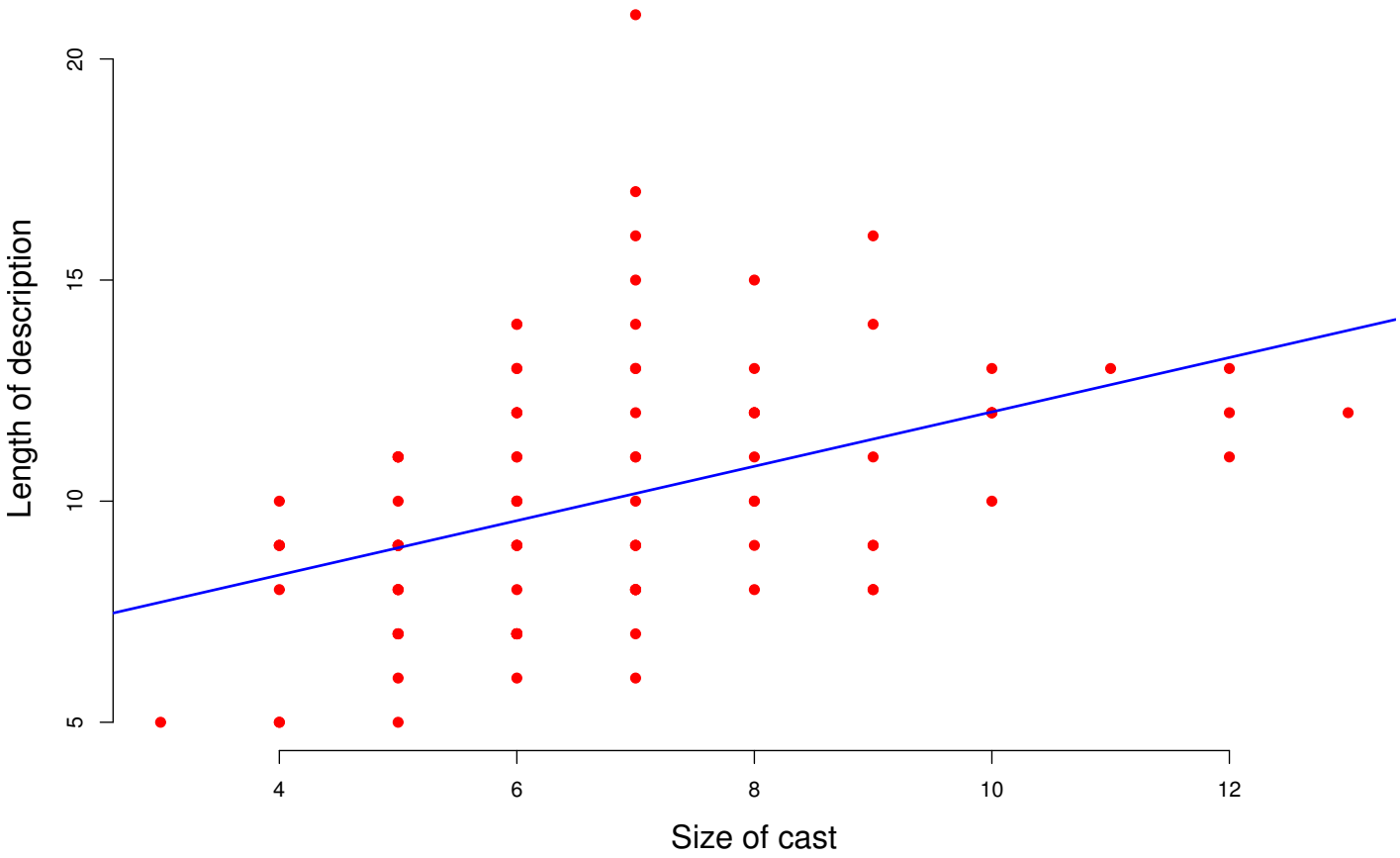
We will make use of the `films.dta` data file again.

We will look at the relation between the size of the cast in a movie and the length of the description of the movie in a 1996 guide. What do you expect to see?

Linear regression



Linear regression



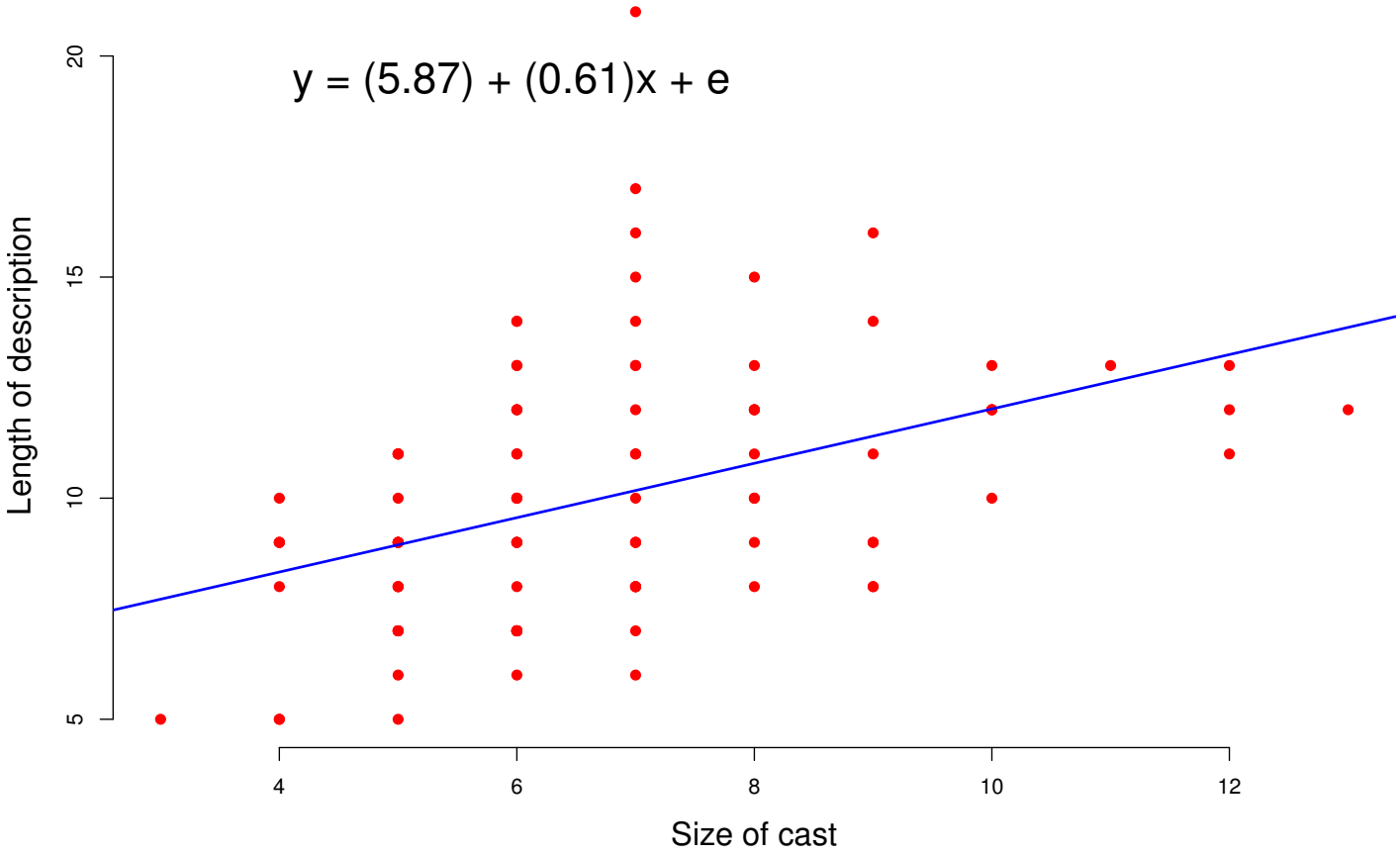
Example

Which is the dependent and which the independent variable in this example?

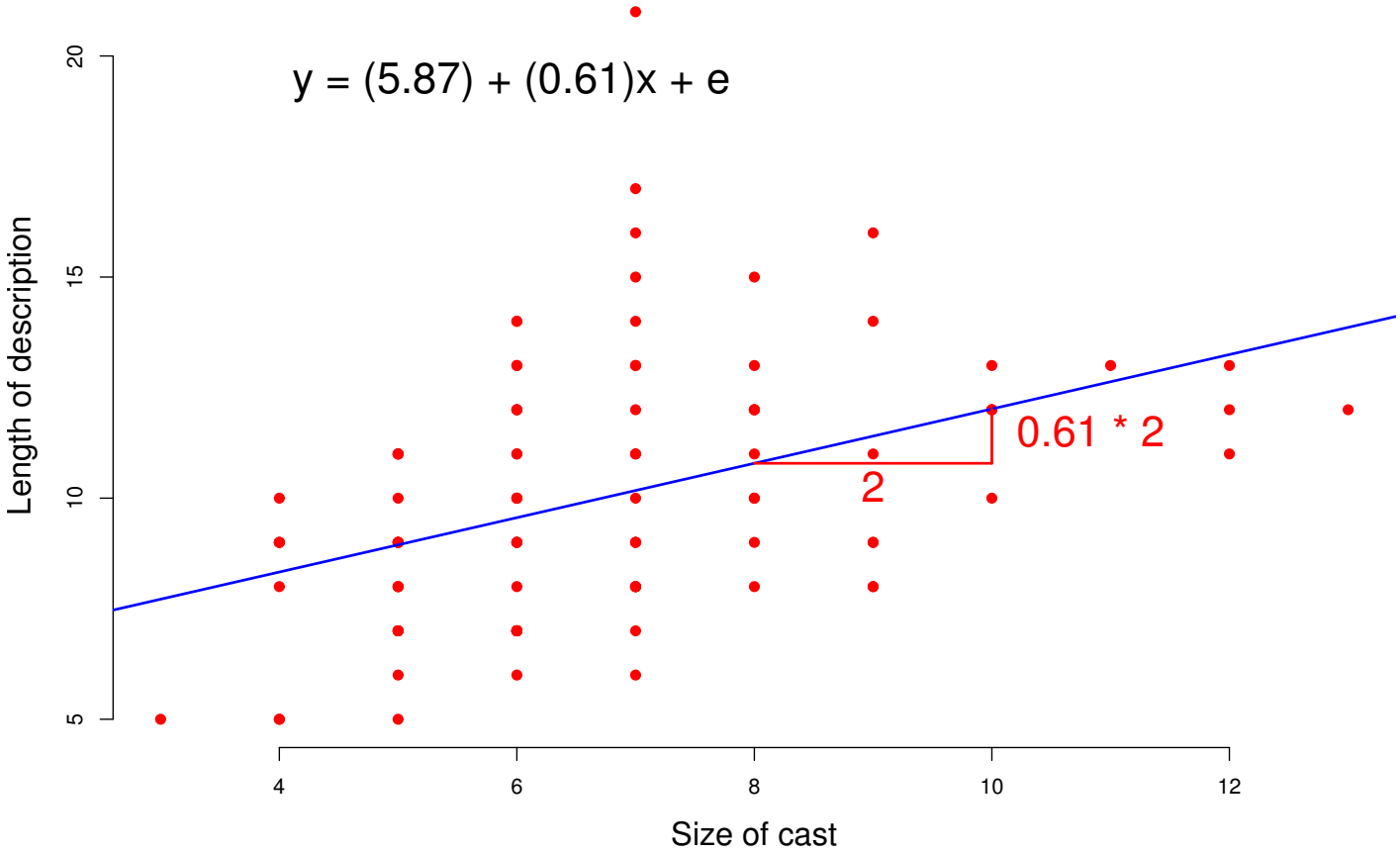
Exercise

- Open `films.dta`
- Create a scatter plot for description length by cast size
- Add a regression line to the plot
- Run a linear regression with those two variables

Linear regression



Linear regression



Example: interpretation

$$y_i = 5.87 + 0.61x_i + \varepsilon_i$$

Example: interpretation

$$y_i = 5.87 + 0.61x_i + \varepsilon_i$$

- There is a positive relation between cast size and description length.

Example: interpretation

$$y_i = 5.87 + 0.61x_i + \varepsilon_i$$

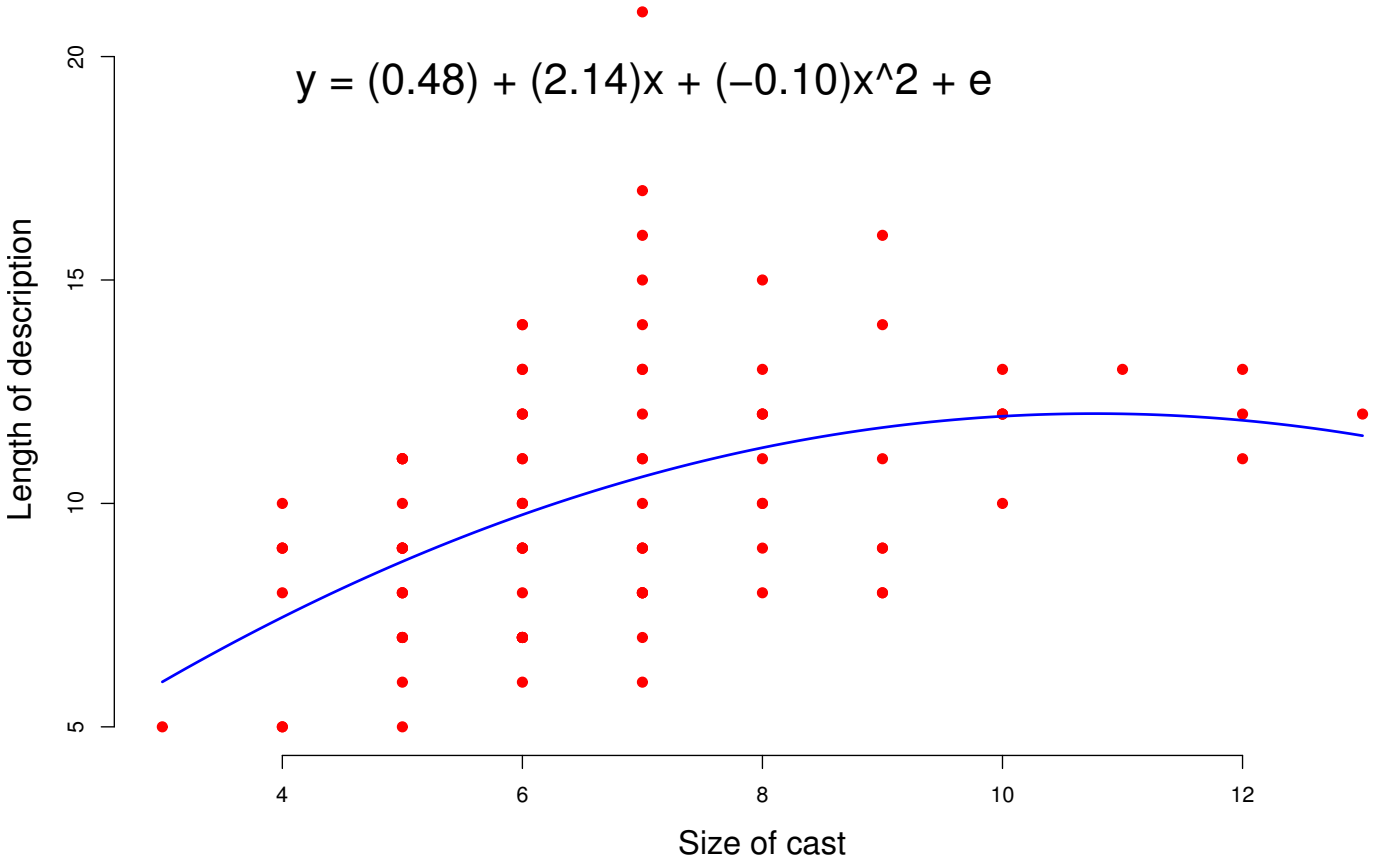
- There is a positive relation between cast size and description length.
- For every increase in cast size by 1, the description length increases by 0.61 lines.

Example: interpretation

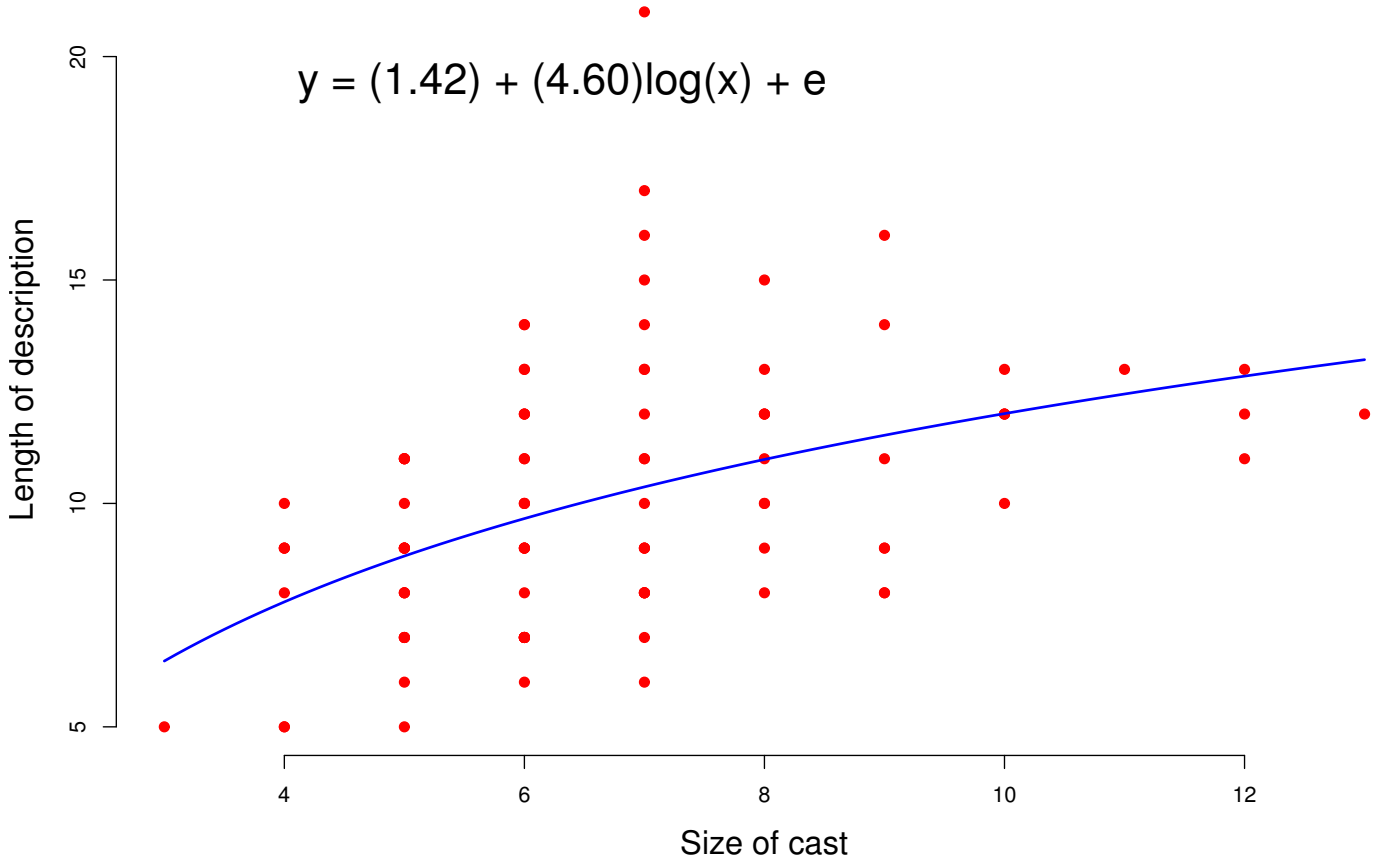
$$y_i = 5.87 + 0.61x_i + \varepsilon_i$$

- There is a positive relation between cast size and description length.
- For every increase in cast size by 1, the description length increases by 0.61 lines.
- For a (hypothetical) film where the cast size is 0, the description would be 5.87 lines.

Linear regression (squared)



Linear regression (log)



Exercise

- Run a linear regression with $log(x)$ instead of x.
- Run a linear regression with x^2 and x instead of just x.

Outline

Linear model

The regression equation here is

$$y_i = b_0 + b_1 x_i + \varepsilon_i,$$

whereby **y** is the dependent variable, **x** the independent variable, i an indicator of the case (film), b_0 and b_1 the model parameters, and ε the error term.

Residuals

$$y_i = b_0 + b_1 x_i + \varepsilon_i$$

The linear prediction given the parameters would be $\hat{y}_i = \hat{b}_0 + \hat{b}_1 x_i$.

Residuals

$$y_i = b_0 + b_1 x_i + \varepsilon_i$$

The linear prediction given the parameters would be $\hat{y}_i = \hat{b}_0 + \hat{b}_1 x_i$.

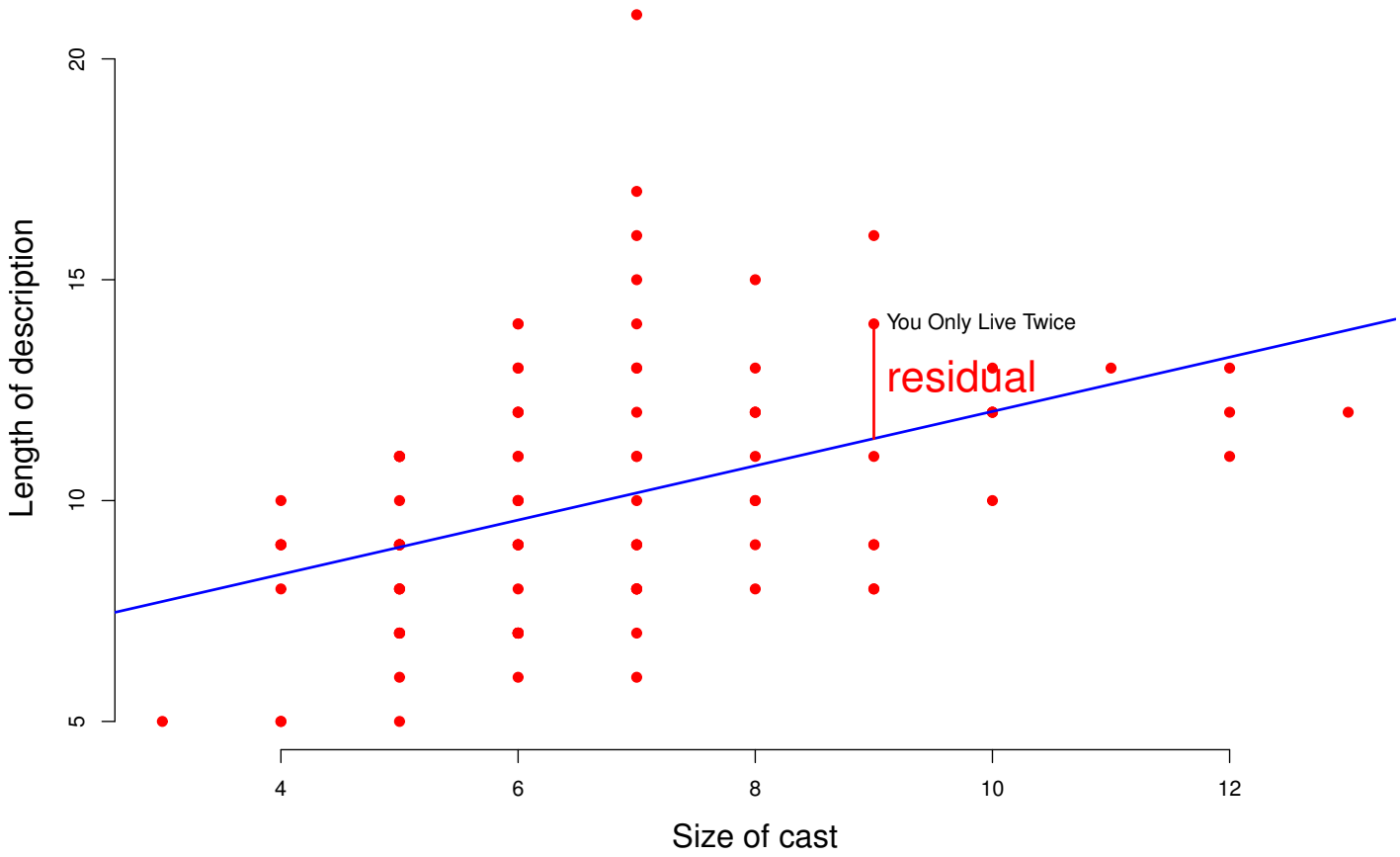
The extend to which the real value differs from the predicted value is:

$$y_i - \hat{y}_i = y_i - \hat{b}_0 - \hat{b}_1 x_i = e_i.$$

Residuals

By this formulation, the residuals ($\mathbf{e}$) are the vertical distance between a point and the regression line (i.e. not the shortest distance between the point and the line).

Linear regression (residuals)



Ordinary Least Squares

To estimate the regression line, we need to estimate the parameters b_0 and b_1.

Ordinary Least Squares

To estimate the regression line, we need to estimate the parameters b_0 and b_1.

Ordinary Least Squares (OLS) is the most common method to do so. With OLS, we estimate the parameters such that the sum of squared residuals are minimized.

Ordinary Least Squares

To estimate the regression line, we need to estimate the parameters b_0 and b_1.

Ordinary Least Squares (OLS) is the most common method to do so. With OLS, we estimate the parameters such that the sum of squared residuals are minimized.

(This is the same as minimizing the variance of the residuals.)

Outline

Model fit

Once we have estimated a line, we might ask how well this line summarizes the relationship between those two variables.

Model fit

Once we have estimated a line, we might ask how well this line summarizes the relationship between those two variables.

A common measure is R^2:

$$R^2 = 1 - \frac{\text{residual sum of squares}}{\text{total sum of squares}} = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}.$$

Model fit

Once we have estimated a line, we might ask how well this line summarizes the relationship between those two variables.

A common measure is R^2:

$$R^2 = 1 - \frac{\text{residual sum of squares}}{\text{total sum of squares}} = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}.$$

This can be interpreted as the proportion of the variation in **y** explained by this model.

Model fit

Once we have estimated a line, we might ask how well this line summarizes the relationship between those two variables.

A common measure is R^2:

$$R^2 = 1 - \frac{\text{residual sum of squares}}{\text{total sum of squares}} = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}.$$

This can be interpreted as the proportion of the variation in **y** explained by this model.

Note the relation with correlation coefficient Peason's r: $r = \sqrt{R^2}$.

Outline

Exercise

Repeat for both the duration of the movie and the year of publication:

1. Plot description length against film length.
2. Regress description length on film length.
3. Interpret the regression results.
4. Evaluate the model fit.

Exercise

Open `demdev.dta` and investigate the relation between income per capita and the Polity IV democracy score in 1980. Repeat for the logged value of the income per capita.