STAT115: Introduction to Biostatistics

University of Otago

Ōtākou Whakaihu Waka

Lecture 20: Fitting Linear Regression Models

Outline

- Previous
  - Model for linear regression
  - $y = \beta_0 + \beta_1 x + \varepsilon$
- Today:
  - Fitting the model
    - Estimating β_0 and β_1
    - Fitted model
    - Residuals

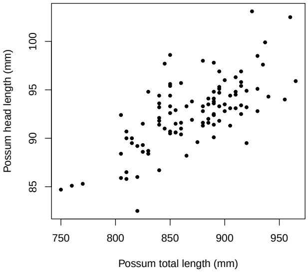
Recall: motivating data

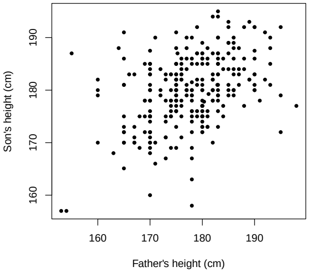
- The size of brushtail possums
  - Exploring relationship between total length (mm) and head length (mm)
- Height of STAT110 students
  - Compare father's height (cm) and son's height (cm)
- Squat weight of international power lifters
  - Look at the relationship between body weight (kg) and max squat weight (kg)

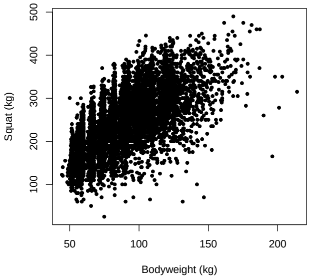
Recall: importing data into R

- Import the data into R

```
possum = read.csv('possum.csv')
height = read.csv('height.csv')
powerlift = read.csv('powerlift.csv')
```







Fitting a regression model

- The (simple) linear regression model is

$$y = \underbrace{\beta_0 + \beta_1 x}_{\text{mean response}} + \varepsilon$$

- β_0 and β_1 are parameters
  - Estimate parameters (population) with statistics (sample)
  - What statistics could we use to estimate β_0 and β_1?

Fitting a regression model

- The (simple) linear regression model is

$$y = \underbrace{\beta_0 + \beta_1 x}_{\text{mean response}} + \varepsilon$$

- β_0 and β_1 are parameters
  - Estimate parameters (population) with statistics (sample)
  - What statistics could we use to estimate β_0 and β_1?
    - We could guess by eye: use paper, pencil and ruler (or electronic equivalents)
    - Later in the lecture: find general approach for estimating β_0 and β_1
- For now: assume we have some way to find estimates $\hat{\beta}_0$ and $\hat{\beta}_1$
- Work through using the possum data to illustrate concepts

Fitted model

- The (simple) linear regression model is

$$y = \underbrace{\beta_0 + \beta_1 x}_{\text{mean response}} + \varepsilon$$

- Once we have estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ we can write the fitted model

$$\hat{\mu}_y = \hat{\beta}_0 + \hat{\beta}_1 x$$

- The fitted model is commonly written as

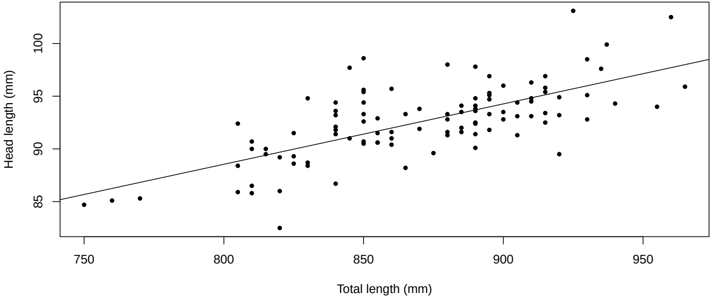
$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

- The fitted model gives the estimate of the mean at a given x value

Fitted model: possum data

- Use estimates $\hat{\beta}_0 = 42.7$ and $\hat{\beta}_1 = 0.0573$
- Fitted model is

$$\hat{y} = 42.7 + 0.0573x$$



Residuals

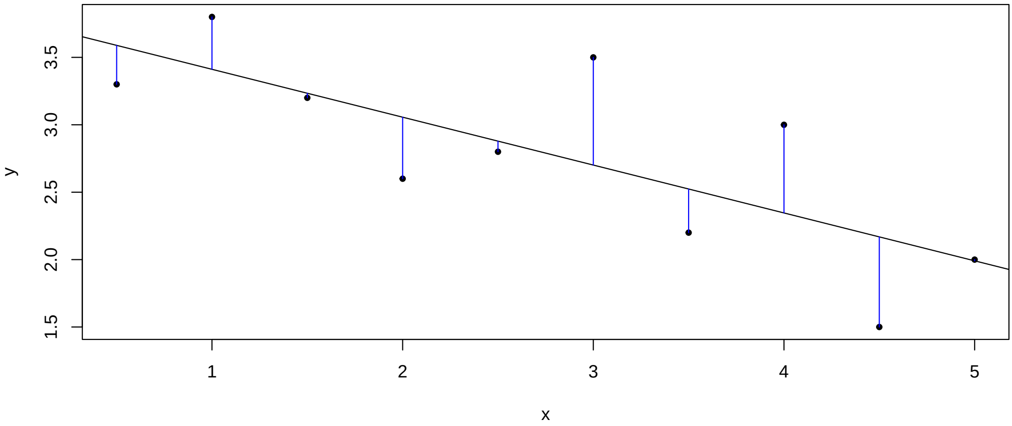
- The statistical model can be expressed as
  - observation = mean response + error
- After fitting the model, we have
  - observation = fitted model + residual
- The residual $\hat{\varepsilon}$ is our best guess (estimate) of the error ε
  - It is the difference between the observation (y) and the mean response $(\hat{y})$

$$\hat{\varepsilon} = y - \hat{y}$$

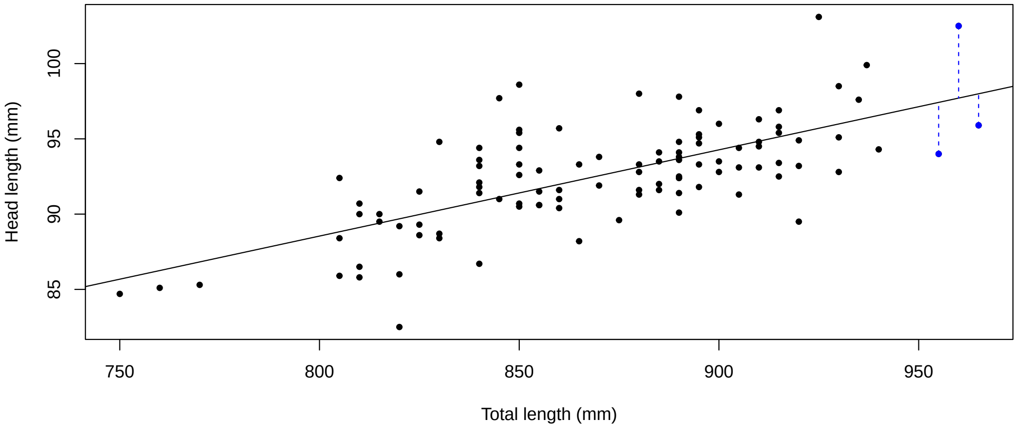
- We often index by i: for the ith observation (x_i, y_i) the residual is

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

Residuals: blue lines



Residuals: possum data (three points in blue)



How do we fit the model?

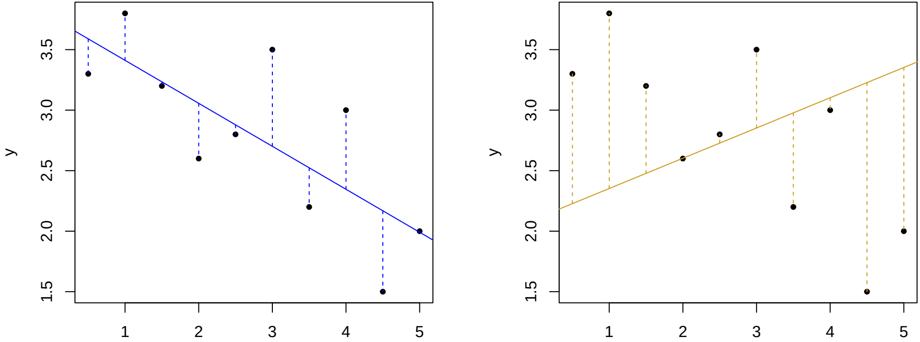
- The (simple) linear regression model is

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

- Estimate parameters β_0 and β_1
  - Find β_0 and β_1 that give the 'best' description of relationship between x and y
- Suppose we had a choice between two possible fitted models
  1. One of them has many large residuals (large positive and large negative residuals)
  2. The other one has mostly small residuals (small positive and small negative residuals)
- Which is better?
  - Look graphically

Graphical representation

- Same data, two possible fitted models
  - One with larger residuals (magnitude): gold
  - One with smaller residuals (magnitude): blue
- Which describes the relationship between x and y better?

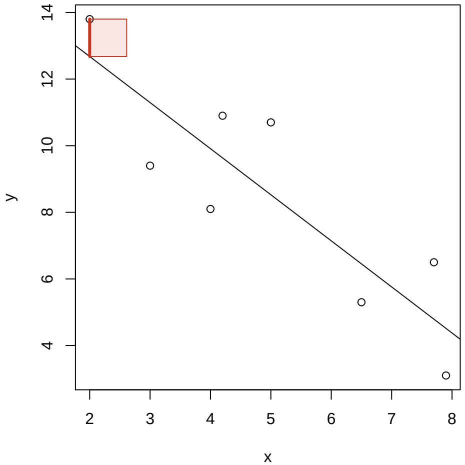


Least squares

- We want the (magnitude of the) residuals to be as small as possible
- We will find estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ using the method of least squares
  - Find the values $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize the sum of squared residuals
- Explain the process graphically

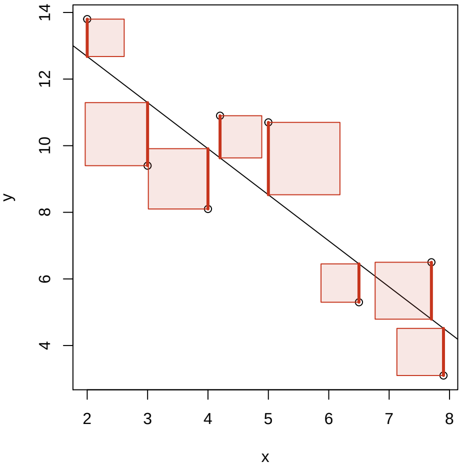
Least squares

- We can visualise the squared residual by drawing a square!
  - Squared residual is the area of red square



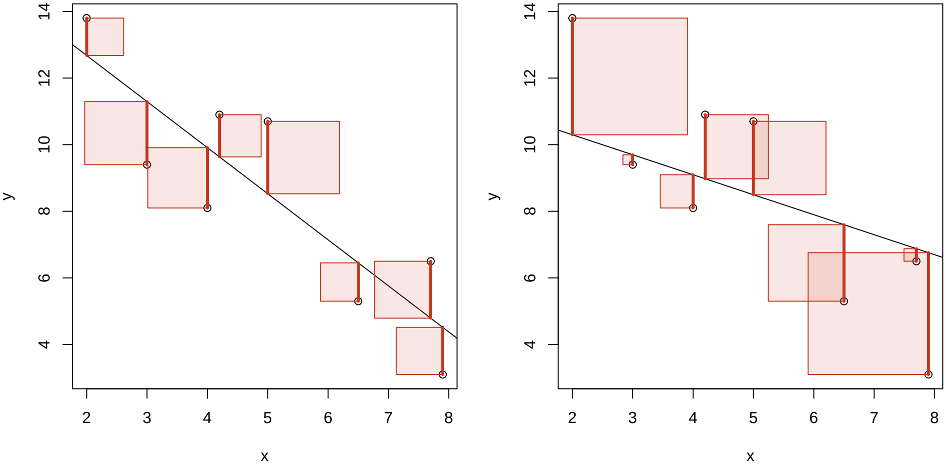
Least squares

- The sum of squared residuals
  - Combined area of the red squares



Least squares

- Minimise the sum of squared residuals (minimise combined area)
  - Left plot: better fit (to the same data)



Least squares

- The sum of squared residuals:

$$\sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$
$$= \sum_{i=1}^{n} (y_i - [\hat{\beta}_0 + \hat{\beta}_1 x_i])^2$$

  - Note: $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$

- Find $\hat{\beta}_0$ and $\hat{\beta}_1$ that make $\sum \hat{\varepsilon}_i^2$ as small as possible

Parameter estimates

- We can use calculus to find estimates
  - $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimise sum of square residuals

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{s_y}{s_x} r$$

  - s_y: sample standard deviation of outcome y
  - s_x: sample standard deviation of predictor x
  - r: sample correlation between x and y

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

- Details of how to find these: outside the scope of the course

In R

- We can find the least squares estimates using R
- The R code is

```r
lm(y ~ x)
```

- Look at each piece in turn:
  - `lm`: function for fitting a **l**inear **m**odel
  - `y`: outcome variable
  - `x`: predictor variable
  - `~`: thought of as 'is modelled by'
  - `lm(y ~ x)`: is saying that we are fitting a linear model where the outcome variable y is modelled in terms of the predictor variable x

Fitting the possum data

```
m_possum = lm(possum$head_l ~ possum$total_l) # assigned the output to object m_possum
summary(m_possum) # shows a summary of the results
```

```
##
## Call:
## lm(formula = possum$head_l ~ possum$total_l)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -7.188 -1.534 -0.334  1.279  7.397
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    42.70979    5.17281    8.26  5.7e-13 ***
## possum$total_l  0.05729    0.00593    9.66  4.7e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.6 on 102 degrees of freedom
## Multiple R-squared:  0.478,Adjusted R-squared:  0.472
## F-statistic: 93.3 on 1 and 102 DF,  p-value: 4.68e-16
```

Estimates in R

- The estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ are given in column headed `Estimate`
  - $\hat{\beta}_0 = 42.7098$
  - $\hat{\beta}_1 = 0.0573$
- R labels the estimates in terms of the variable names
  - `(Intercept)`
  - `possum$total_l`

Detour: `data` option in `lm`

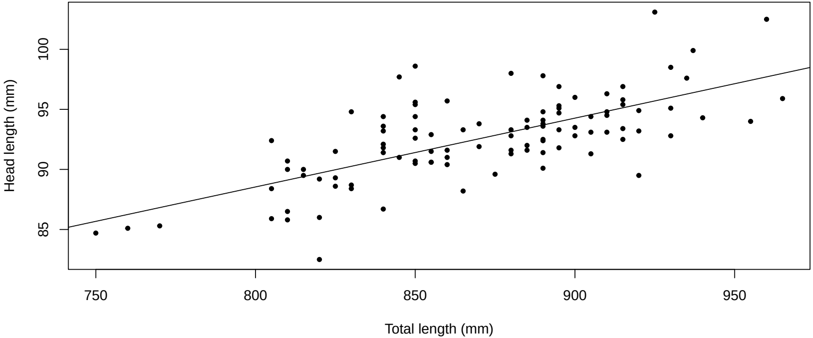
- The `lm` function includes a `data` option that can make specification easier
- Separate the variable (e.g. `head_l`) from the data frame object (`possum`)
- The code is

```
m_possum2 = lm(head_l ~ total_l, data = possum)
```

- This is fitting the same model as in the slide above

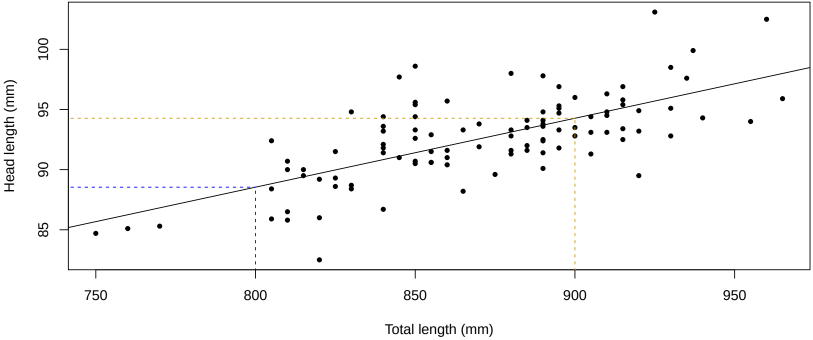
Fitted model: possum data

- The fitted model is $\hat{y} = 42.7098 + 0.0573x$
  - Recall: y is head length, x is total length
  - We could also write: $\widehat{\text{head}} = 42.7098 + 0.0573\,\text{total}$



Fitted model: possum data

- Fitted model is $\hat{y} = 42.7 + 0.0573x$
  - For $x = 800$ we have $\hat{y} = 42.7 + 0.0573 \times 800 = 88.5$
  - For $x = 900$ we have $\hat{y} = 42.7 + 0.0573 \times 900 = 94.3$



Interpretation

- Fitted model $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$
  - For the possum data: $\hat{y} = 42.7 + 0.0573x$
- Our interest is $\hat{\beta}_1$:
  - We estimate that the average head length of a possum will increase by 0.0573 mm for a 1 mm increase in total length.
- This is a comparison of two subpopulations
  - If we compare possums whose total length is x mm to possums whose total length is $x + 1$ mm, the estimated increase in their expected (or mean) head length is 0.0573 mm.
- $\hat{\beta}_0$ is the estimated mean head length of possums with total length 0 mm
  - Makes no biological sense
  - Do not interpret in this case

Summary

- Fitting a linear regression model
  - Fitted values: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$
  - Residuals: $\hat{\epsilon} = y - \hat{y}$
- Method of least squares
  - Minimise the sum of squared residuals
  - Fit the model using `lm` in R: `lm(y ~ x)`