STAT115: Introduction to Biostatistics

University of Otago

Ōtākou Whakaihu Waka

Lecture 21: Checking the Assumptions of the Linear Regression Model

Outline

- Previous:
  - Fitting a statistical model
  - Method of least squares
- Today:
  - Assumptions underlying linear regression
    - What are the assumptions?
    - How do we check the assumptions?

Motivation

- Exploring relationship between total length (mm) and head length (mm) of brushtail possums
- Recall: fitting linear model

```
m_possum = lm(head_l ~ total_l, data = possum) # possum data
```

- Linear regression model allows us to:
  - Estimate the effect of x (total length) on y (head length)
  - Estimate the mean response of y (head length) given x (total length)
    - E.g. estimate mean head length of possums that have total length $x = 820$ mm
- Problem: the model relies on assumptions
  - Interpretations and conclusions may be invalid if assumptions are badly wrong
- We need to test the model assumptions (so far as possible)

Assumptions for Simple Linear Regression

- Recall that the linear regression model is

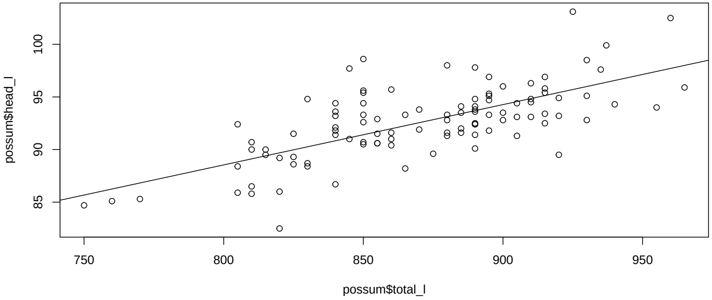
$$y = \underbrace{\beta_0 + \beta_1 x}_{\mu_y} + \varepsilon$$

- The underlying assumptions are:
  - **Linearity:** The mean response μ_y is described by a straight line
  - **Independence:** The errors $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ are independent
  - **Normality:** The error terms ε are normally distributed
  - **Equal variance:** The errors terms all have the same variance, σ_ε^2 ('homoscedastic')
- These are often remembered using the mnemonic **LINE**.

Tools for checking assumptions

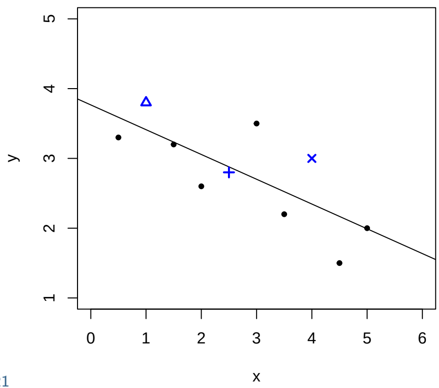
- Fitted line plot: compare the observed data to the fitted model
  - Useful, but not extensively used for checking assumptions
- Show code for plotting data and fitted model

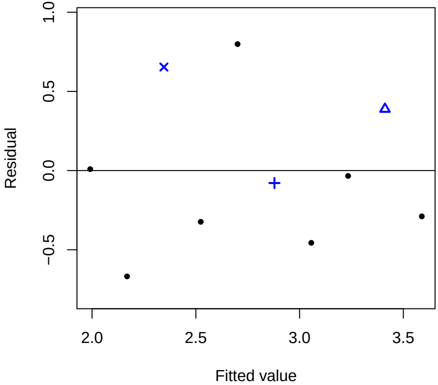
```
plot(possum$total_l, possum$head_l) # plot(x,y): x gives x values, y gives y values
abline(m_possum) # draws the fitted regression line
```



Residual plots

- It is more common to use a residual plot
  - Residuals $\hat{\varepsilon}$ are on the y-axis
    - Recall: $\hat{\varepsilon} = y - \hat{y}$
- Look at a small example



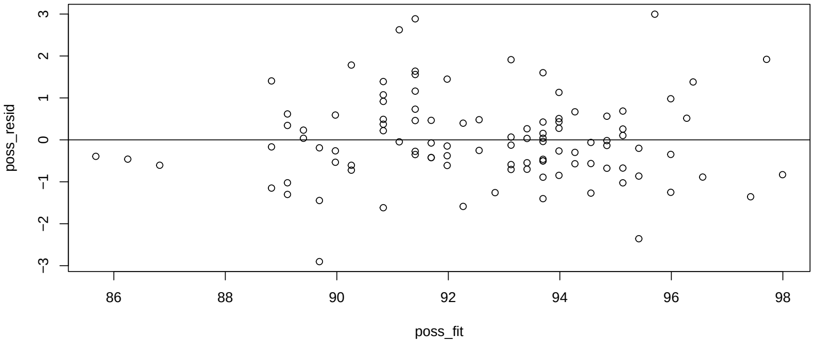


More on residuals: $\hat{\varepsilon} = y - \hat{y}$

- The residual is $\hat{\varepsilon} = y - \hat{\beta}_0 - \hat{\beta}_1 x$
- Residuals are estimates of error terms (ε)
  - Can be used to check assumptions about error terms (ε)
- The residual $\hat{\varepsilon}$ is often called a raw residual
  - Standardised or studentised residuals are often preferred
    - We will use studentised residuals in this course
  - What are studentised (or standardised) residuals?
    - Transformed to have standard deviation ≈ 1
    - (Mathematical) details are beyond the scope of the course
  - Find them in R using function `rstudent`
    - e.g. for model object `m_possum` we find studentised residuals using `rstudent(m_possum)`

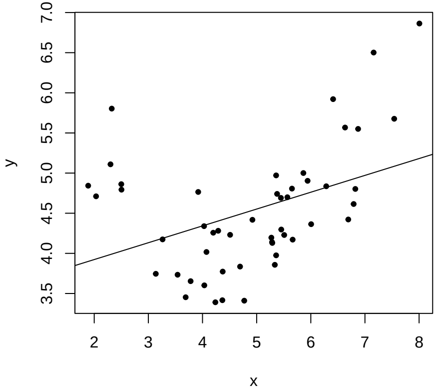
Plotting residuals in R

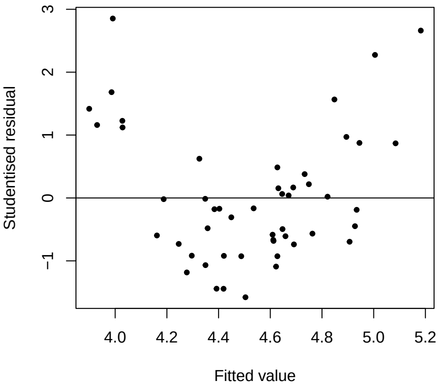
```r
poss_fit = fitted(m_possum) # finds the fitted values of the model m_possum
poss_resid = rstudent(m_possum) # finds the studentized residuals of the model m_possum
plot(poss_fit, poss_resid) # plots residuals against fitted values
abline(h=0) # draws a horizontal line at 0
```



Checking the linearity assumption

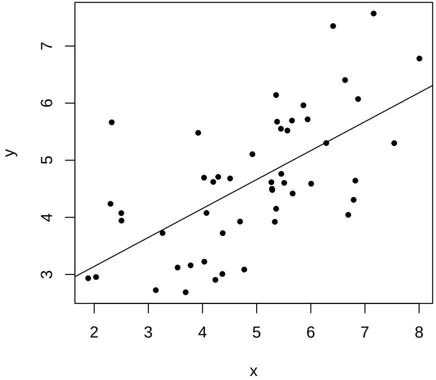
- Looking for clear departure for linearity in trend of data.
  - Look for patterns in plot of residuals against fitted values
- Plots below illustrate failure of linearity assumption (bad)

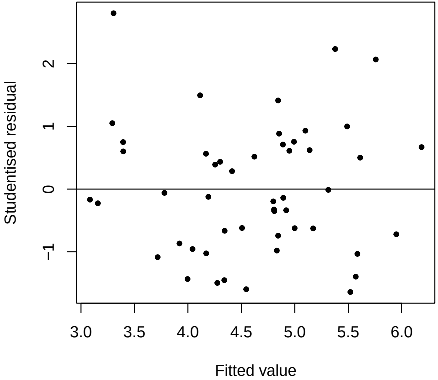




Checking the linearity assumption

- Looking for clear departure for linearity in trend of data.
  - Look for patterns in plot of residuals against fitted values
- Plots below: no evidence of failure of linearity assumption (good)





The independence assumption

- Independence assumption: errors $\varepsilon_1, \ldots, \varepsilon_n$ are independent
- What does it mean that errors ε_1 and ε_2 are independent?
  - Knowing ε_1 tells us nothing about ε_2
    - $\varepsilon_i = y_i - (\beta_0 + \beta_1 x_i)$
- For the possum example, independence means
  - Knowing how much above average one possum's head length is, gives no information about how far above average another possum's head length is.

Checking the independence assumption

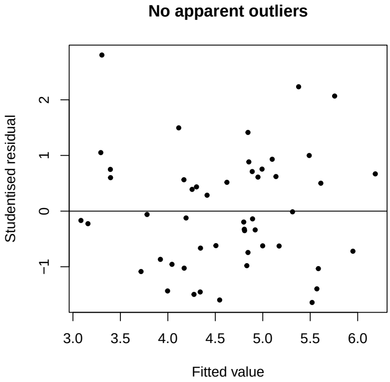
- In general: difficult to assess
  - We are unable to check it by looking at fitted line or residual plots.
- In certain situations, we may be able to check it
  - If the data are collected in time (time series)
    - Expect observations close together in time to be correlated
  - If the data are collected in space (spatial data)
    - Expect observations close together in space to be correlated
  - If there are multiple measurements from each participant (repeated measures)
    - Expect observations from a given participant to be correlated
- We can look at more complex statistical models for each of the cases above
  - Outside the scope of this course

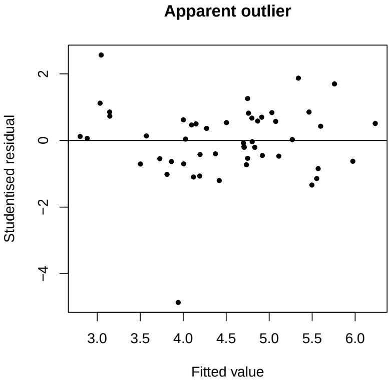
Checking the normality assumption

- Assumption: errors ε are normally distributed
- The importance of the normality assumption depends on sample size
  - Sample size small: important, but hard to check
  - As sample size increases (say $n > 50$) it becomes increasingly less important
    - Looking for large violations of normality
- An example of such a violation are outliers / extreme observations

Checking for outliers

- Studentized residuals should be approximately normal with standard deviation 1:
  - Most (approx 95%) within ± 2
  - Nearly all ($> 99\%$) within ± 3
  - Values exceeding ± 4 are unusual: outliers



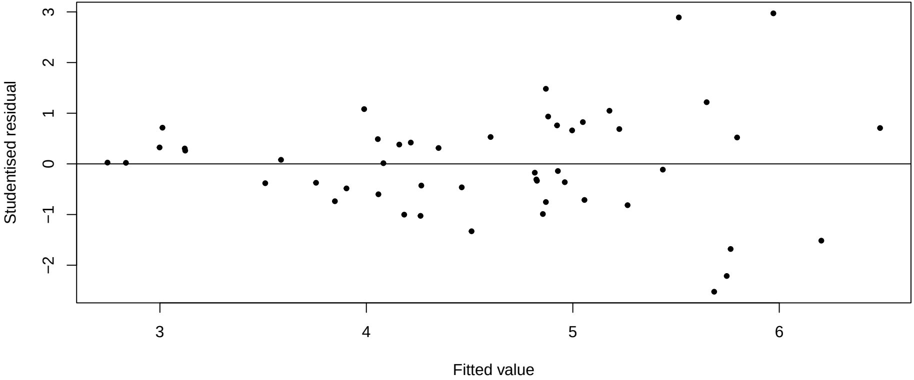


Checking equal variance assumption (homoscedasticity)

- Assumption: error terms $\varepsilon_1, \varepsilon_2, \ldots, \epsilon_n$ have the same variance
  - The magnitude of spread of data about regression line should not change too much with x
- In contrast, if (say) variance of error terms increases with x
  - We would expect to see data more dispersion as x increases.
- Best seen with residual plot against fitted values.

Checking equal variance

- Example where there is evidence of non-constant variance
  - Variance of residuals increases with fitted value



What to do when assumptions fail: linearity

- Failure of the linearity assumption is critical
  - Conclusions drawn from the model will be invalid
- Paths forward include
  - Consider transforming outcome or predictor variables (where appropriate)
  - Explore more sophisticated models
    - Move beyond a simple linear regression model
- Both of these are outside the scope of the course
  - Considered further in STAT 210, 310

What to do when assumptions fail: independence or equal variance

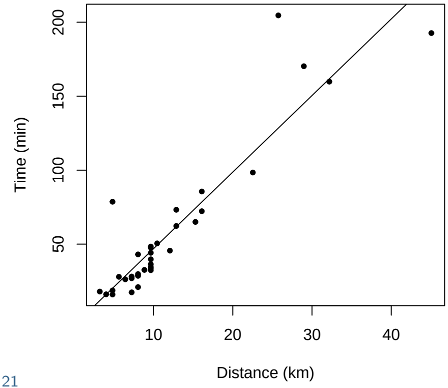
- When independence or equal variance assumptions fail
  - Estimates of parameters remain valid
  - Estimates can be inefficient
    - They can be improved
- Follows that fitted regression line is useable
- Confidence intervals and hypothesis tests will be invalid.
- Failure of assumptions can be rectified by sophisticated modelling techniques.
  - Details beyond this course.

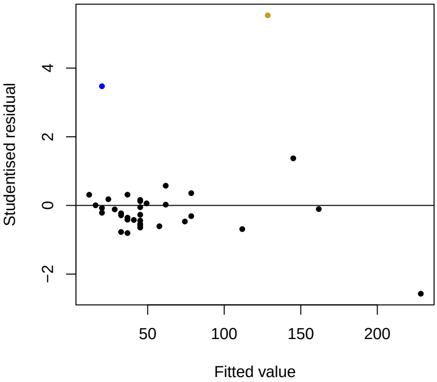
What to do when assumptions fail: normality / outliers

- Outliers can have a dramatic effect on the estimated regression
- If outliers are present: check that the data are correctly recorded
- If outliers remain we may consider removing them, however:
  - Think carefully first
    - Often outliers (or unexpected values in general) are the most interesting
    - They could be revealing something important about what we are studying
  - If we do remove observations, we must be transparent
    - It should clear and obvious that values were removed and why
- Look at an example

Scottish hill racing

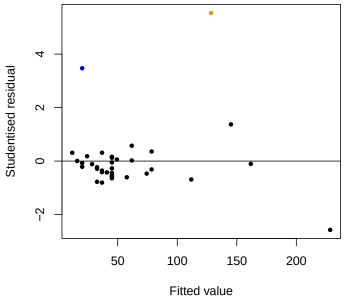
- Data are the record times in 1984 for 35 Scottish hill races (running)
- Interested in the relationship between distance and record time
  - Outcome variable (y): record time (in minutes)
  - Predictor variable (x): distance (in km)





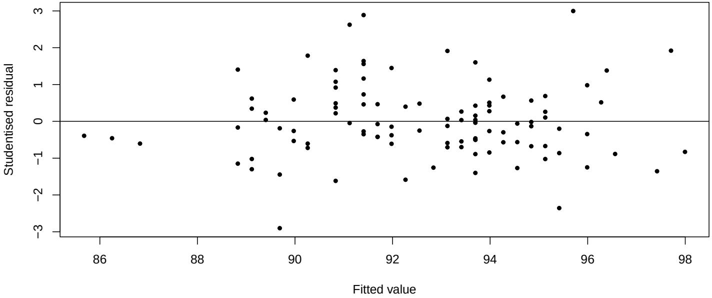
Scottish hill races: Investigate the outliers

- Knock Hill: record incorrectly recorded
  - Recorded as 78 minutes 39 seconds
  - It should have been 18 minutes 39 seconds.
- Bens of Jura: other important information?
  - This race has the largest climb by over 700 m
  - Consider (extended) model that includes climb?



Residuals: possum data

```
plot(fitted(m_poss), rstudent(m_poss), pch = 20, xlab = "Fitted value",
     ylab = "Studentised residual") # xlab (x label), ylab (y label), pch (point)
abline(h = 0)
```



- Linearity: no evidence of a trend
- Outliers: no apparent outliers
- Constant variance: no obvious change in magnitude of spread of residuals

Summary

- Assumptions of linear regression
  - LINE
    - Linearity
    - Independence
    - Normality
    - Equal variance
- Introduced residual plots
  - Can be used to check assumptions of linear regression model