STU22005 Laboratory Session 4

Professor Caroline Brophy

Week beginning 8th March 2021

Instructions

- In this lab sheet, there will be a series of notes for you to work through. For each section, you should run the R code yourself on your own computer and verify the output provided. Ask your lab demonstrator any questions you have as you work through the material.
- In the last section of the sheet, there are several questions where you will write your own code. Your lab demonstrator is there to help if you get stuck in this section.
- Create a script for all the code that you write in this lab session. Save the script into the folder you have for the module. Save your code regularly throughout the class. (Do not write your code into the "Console" window as this will not save in an R script.)

Introduction

Today we will analyse the data from the in-class experiment carried out in the first lecture of term. This is the lab session you have all been waiting for I'm sure!!

Experiment

First, a recap on the experiment.

On the first day of term, you were all asked to answer one question, assigned to you at random from a group of six. Here were the questions:

- Put the following list of words in alphabetical order, either:
  - bouncing handle kitchen tracksuit university
  - **w**ashing **w**eird **w**hich **w**isdom **w**onderful
  - **mo**bile **mo**del **mo**ment **mo**untain **mo**vie
  - **sta**ck **sta**ke **sta**ndard **sta**pler **sta**tistics
  - **star**board **star**dom **star**fish **star**ry **star**tle
  - **cross**bar **cross**fire **cross**ing **cross**road **cross**word

(Of course, the bolding was not included in the actual experiment.)

The data

Three variables were recorded, Question (1-6), Seconds (duration in seconds it took to complete the question), and Correct (0 wrong or 1 correct).

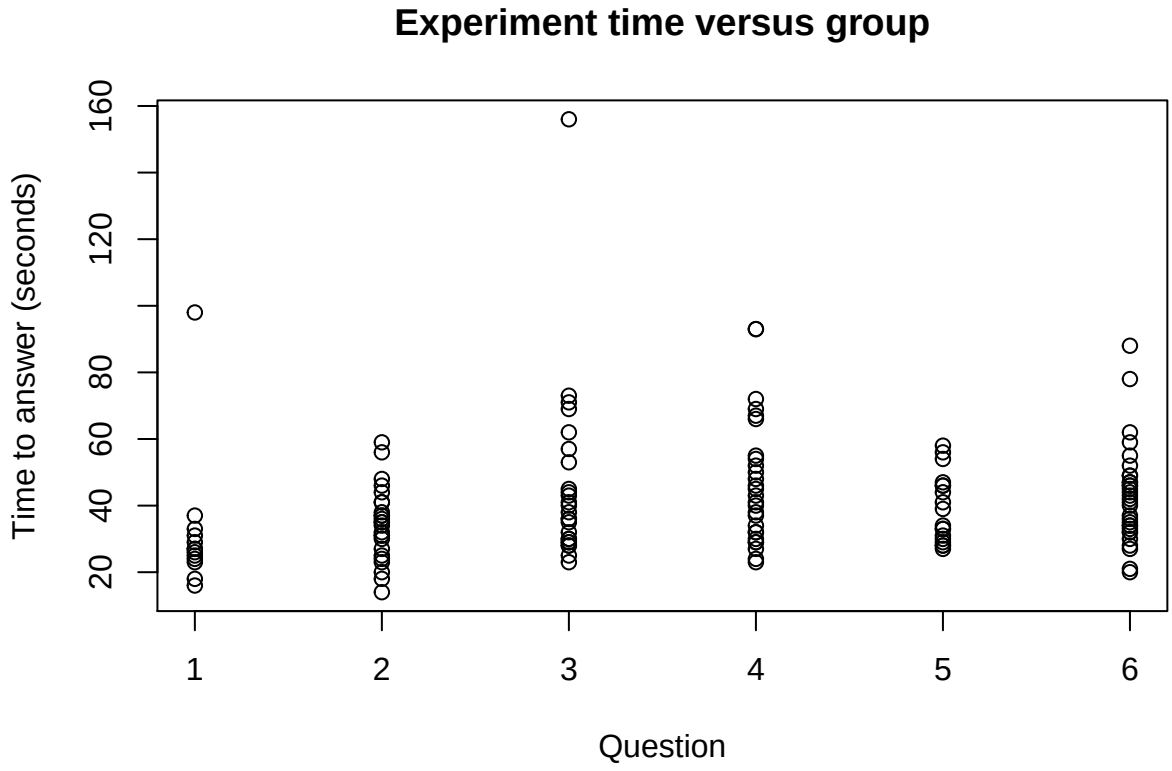
- Read the data into R, view the first few lines and visualise the data.

```
# Read the data into R
exp <- read.csv("Lab4_ExperimentData.csv")

# View the first few lines of the data
head(exp)
```

```
##   Question Seconds Correct
## 1        1      16       1
## 2        1      18       1
## 3        1      23       1
## 4        1      24       1
## 5        1      25       1
## 6        1      26       1
```

```
# Scatterplot
plot(exp$Question, exp$Seconds, main = "Experiment time versus group",
     xlab = "Question", ylab = "Time to answer (seconds)")
```



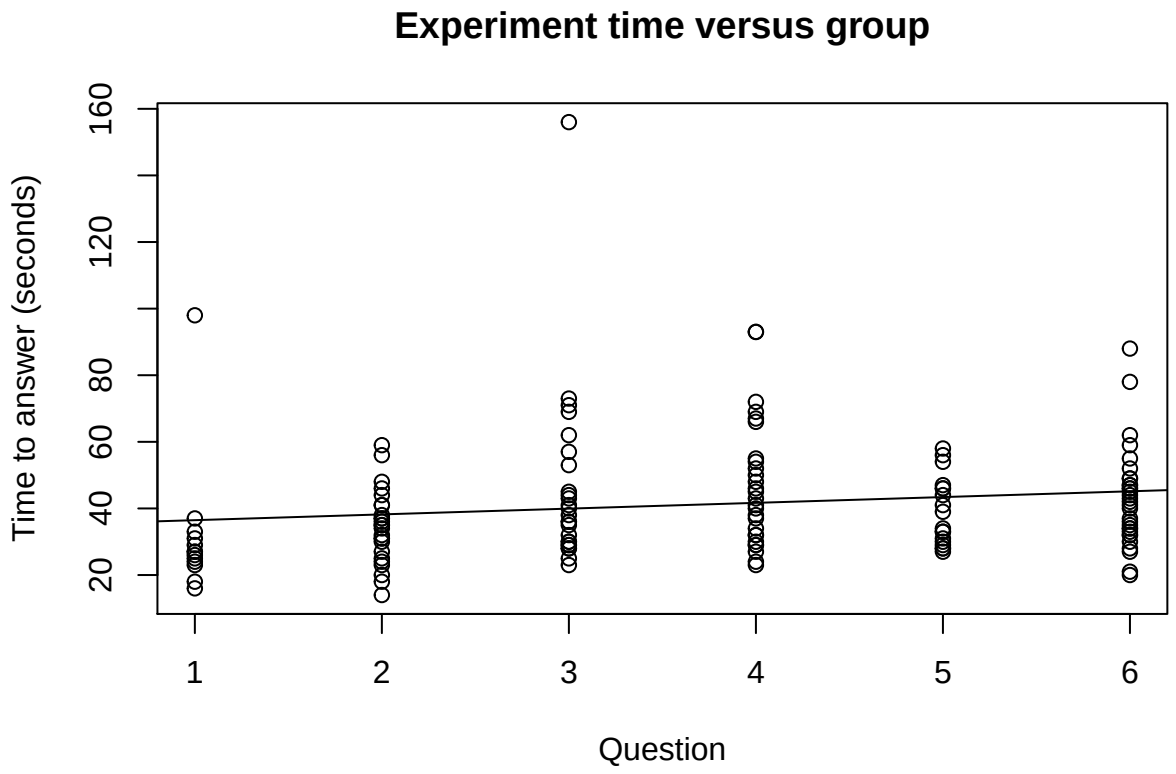
Fitting a simple linear regression (SLR) model

- Fit a simple linear regression model to the data.

```
# Fit and summarise the slr model fit to the data
lm1 <- lm(Seconds ~ Question, data = exp)
summary(lm1)
```

```
##
## Call:
## lm(formula = Seconds ~ Question, data = exp)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -25.145 -11.340  -3.676   4.191 116.058
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  34.7387     3.9001   8.907 2.85e-15 ***
## Question      1.7343     0.9514   1.823   0.0705 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.44 on 137 degrees of freedom
## Multiple R-squared:  0.02368,    Adjusted R-squared:  0.01656
## F-statistic: 3.323 on 1 and 137 DF,  p-value: 0.07048
```

```
# Scatterplot with lm fit
plot(exp$Question, exp$Seconds, main = "Experiment time versus group",
     xlab = "Question", ylab = "Time to answer (seconds)")
abline(lm1)
```



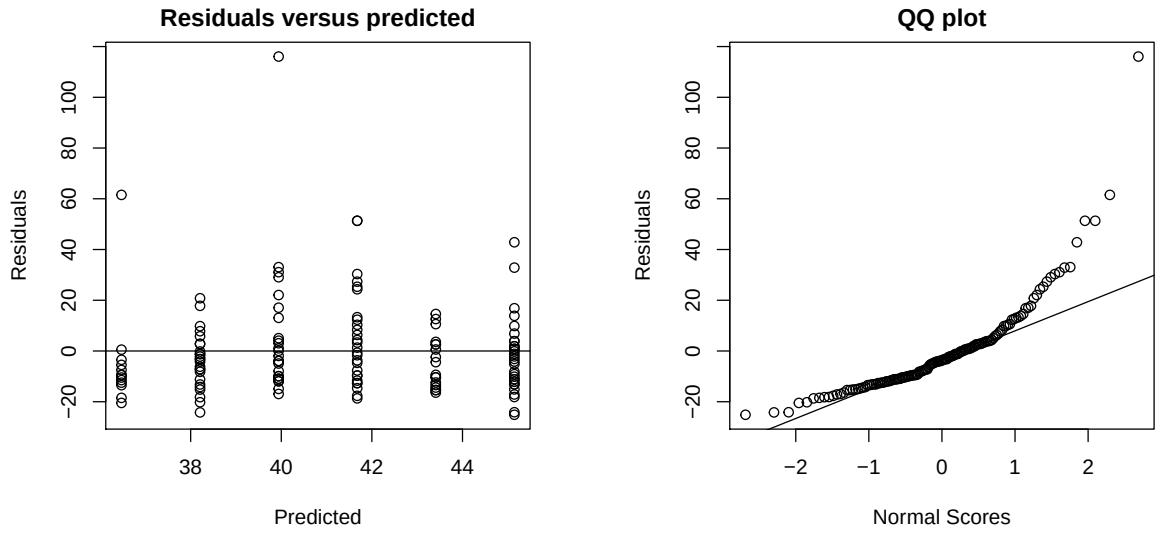
What are the assumptions of a simple linear regression model and are they satisfied here?

- Take a look at the residuals versus fitted values plot, and the QQ plot.

```r
# Store the residuals, standardised residuals and predicted values
resids <- resid(lm1)
sresids = rstandard(lm1)
preds <- predict(lm1)

# Plot the residuals
par(mfrow = c(1, 2), mar = c(5,6,2,2))
plot(preds, resids, xlab = "Predicted", ylab = "Residuals",
     main = "Residuals versus predicted")
abline(h = 0)

# QQ probability plot
qqnorm(resids, ylab="Residuals", xlab="Normal Scores",
       main="QQ plot")
qqline(resids)
```




```r
par(mfrow = c(1,1), mar = c(5,4,4,2))
```

The assumptions of the model, expressed in terms of the error term (ϵ_i) are:

- $E[\epsilon_i] = 0$.
- $\text{Var}(\epsilon_i) = \sigma^2$ (and does not depend on i).
- ϵ_i are independent.
- $\epsilon_i \sim N(0, \sigma^2)$.

Are the assumptions reasonable for this model?

What might we do next?

Questions

1. Restrict the full dataset to only groups 1 to 4, and omitting two outliers. Use the `which()` function in R as follows.

```r
# Create a subset of the original dataset
exp_subset <- exp[which(exp$Question <5 & exp$Seconds < 98),]
```

```
summary(exp_subset)
```

2. Fit a simple linear regression model to the subset of data, generate a scatter plot with the fitted model line included, and test the model assumptions. This is similar to what we did in class for this dataset.
3. There are two values in the restricted dataset that had `Correct = 0`. Is it appropriate to include these observations in the analysis? Repeat step 2 omitting these two values. Start by creating `exp_subset2` that excludes these two values using the `which()` function again. Do the results change much?
4. What next for analysing this dataset?
5. Now we are going to work with a small dataset to reinforce the ideas behind SLR, least squares estimation and how to find residuals, and predicted values.

- Carry out the following tasks:
  - Read a vector of x values into R: 1, 2, 3, 4, 5, 6, 7, 8. Call it `x`.
  - Read a vector of y values into R: 15, 10, 21, 16, 19, 25, 21, 29. Call it `y`.
  - Create a scatter plot of y versus x.
  - Fit a simple linear regression model using the `lm()` function.
  - Verify the estimates of the least squares estimates of the intercept and slope using a series of manual functions in R (such as `xbar = mean(x)`, and `sumxsq = sum(x^2)`). (Go back to the lecture notes for the formulae for OLS estimation of the SLR model.)
  - Store the residual and predicted values from the `lm` fitted model and combine them into a data frame with the original data (use the `data.frame()` function)
  - Print out the dataset, and verify **by hand** the values for the residuals and predicted values in rows 1 and 2.