ST303/ST633 Linear Models
Assignment Sheet 2

Due: Fri 11th Nov, 11:59am.

- Only one, randomly chosen question will be marked.

- Answer Questions 1,3 and 4 using R. Submit questions 2-4.

- If you are familiar with RMarkdown, you may wish to use it to knit your results to a .pdf file (but this is not strictly necessary).

- If so, place your name and student number under author in the YAML header, e.g.

  ```
  ---
  title: "Assignment 2"
  output: pdf_document
  author: John Doe 87654321
  ---
  ```

- Either way, your handwritten and/or typed work should be submitted in a single, combined .pdf along with relevant output.

- Make sure you attend the tutorial scheduled in the week ahead of the assignment submission. The tutor will work through Qu 1, answer questions and help students getting started with R (finding/reading in data, knitting an Rmd file etc).

1. An experiment was conducted to assess the relationship between storage temperature of milk and the bacterial count in the milk after a period of time. The results are given in bacteria.csv.

   (a) Fit a simple linear regression model to these data and provide appropriate graphics to assess the fit of the model. Identify the issues with the model fit.

   (b) Try appropriate transformations to the response and / or predictor to find a model that is better fit to the data than the model in (a). Briefly describe each of the models you fit, discuss how well each model fits and indicate which one you deem most appropriate to model the data.

   The code you may need is below:

   ```
   bacteria <- read.csv("Bacteria.csv")
   fit <- lm(count ~ temp, data = bacteria)
   summary(fit)
   plot(bacteria$temp, bacteria$count)
   abline(fit, col = 2)
   plot(fit, which = 1)
   plot(fit, which = 2)
   fit2 <- lm(sqrt(count)~temp, data = bacteria)
   fit3 <- lm(log(count)~temp, data = bacteria)
   ```

2. A simple linear regression model was fitted to four different data sets and the resulting plots of standardised residual versus fitted value are shown. In each case, discuss whether the simple linear regression model is appropriate. In any case where it is not, suggest the most appropriate way to remodel the data giving reason(s) for your answer.
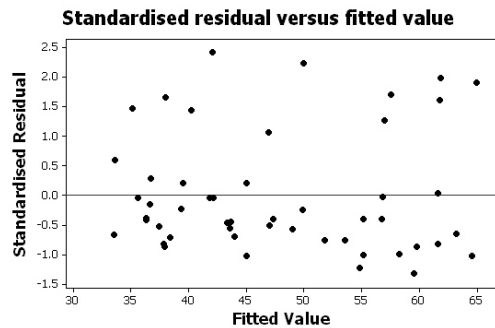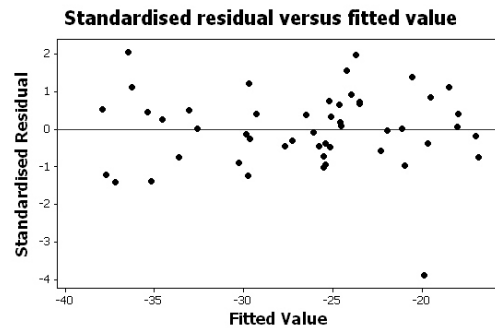
Data set 1

Standardised residual versus fitted value

Data set 2

Standardised residual versus fitted value

Data set 3

Standardised residual versus fitted value

Data set 4

Standardised residual versus fitted value
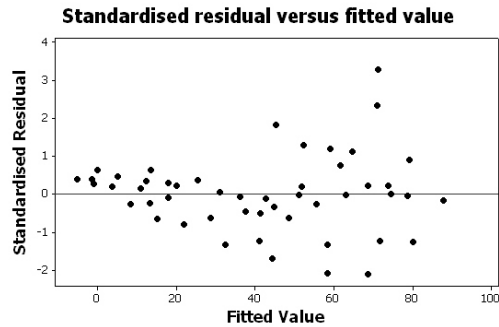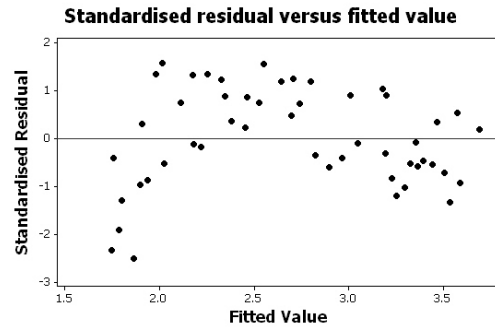
3. Construct some data as follows:
For x generate a column of values $1, 2, ..., 30$.
For y construct a column with values $10 + 2x + \epsilon$, where ϵ contains 30 values randomly generated from the Normal distribution with a mean of 0 and a standard deviation of 4.

   (a) Plot y versus x for the following generated data. Fit a line

   $$y = a + bx$$

   by least squares and draw it on the above plot.

   (b) Fit a line

   $$x = c + dy$$

   by least squares and draw it on the above plot.(Hint: look at ?abline to add a line to base R plot).

   (c) Are the lines in parts (a) and (b) the same? Explain.

4. The data set pollen.txt gives the proportions of pollen removed and visit duration by bumblebee queens and honeybee workers. For the bumblebee queens only (code=1):

    (a) Plot pollen removed versus time spent on flower. Fit the regression of pollen removed on time spent on flower. Plot the residuals versus the fitted values. Does the linear regression model seem appropriate? What problems are evident in the response versus predictor plot? What problems are evident in the residuals versus fitted values plot?

    (b) Do log transformations of Y and / or X help resolve the problems in (a)? (You do not need to include plots, just give a brief text explanation.)

    (c) Try fitting the regression only for those times less than 31 seconds (i.e. excluding the two longest times). Does this fit better?

The code you may need is below:

```
library(tidyverse)
pollen <- read.table("pollen.txt", header =  TRUE)
head(pollen)
pollen <- pollen %>% filter(code==1)
pollen_c <- pollen  %>%
  filter(duration < 31)
```