ENGSCI 314, First Semester, 2018
Second Statistics Assignment, Due: 1pm Tuesday, 1ST MAY

Instructions concerning this assignment:

I am providing you an R Markdown document called **`ENG314A2.rmd`** (available on Canvas) which will have some answers already filled in. You will need to fill in and complete the rest of the document. The data files you will be using for the assignment are described in the questions and are available from Canvas. Make sure you put these datasets in the same place you put the R markdown document because it is going to look for them there. The first change you need to make to the markdown document is put your name and ID number at the top.

Instructions concerning this assignment:

- The assignment will be worth 4% towards your final grade.
- When asked to use ***R*** for this assignment, include the ***R***-code along with any output.
- **Show your working for any calculations!**
- **Hand the assignment in at the START of class.**
- The total marks for this assignment will be **45** (this includes 5 marks for presentation and communication). **Most of the marks for assignments will tend to be for interpretation**.
- There are **5 Presentation and Communication marks** for this assignment as follows:
  - **Name and ID number** at top of R Markdown document.
  - **Space saving and printing assignment 2-up.** Not printing out unnecessary output (listing ENTIRE data sets or showing erroneous R output). Assignment work printed out in "2-up" layout. 2-up layout prints 2 pages side-by-side reduced to one page.
  - **Readability.** This is for your general communication ability in the assignment. This includes sentences clearly conveying the correct idea; sentences making sense; comments not being excessively long or short; conclusions following logically from previous statements.
  - **Use of Natural Language in Executive Summaries.** In executive summaries, this is for discussing the analysis in context, not using variable names, using units when known and rounding sensibly.
  - **Keeping to the Point in Executive Summaries.** In executive summaries this is for not going into far more detail than required.

Notes: Questions **1 and 3** are open questions. The approach to answering them is:

- Comment on the questions of interest or the goal of the analysis.
- Look at data (plot it and maybe get summary statistics) and comment on it.
- Fit a model to the data
  - Check the model assumptions.
  - Change model and repeat checks as needed. You may have to do this more than once.
- Generate inference output required from final model.
- Write a Method and Assumption Checks section.
  - This will detail the steps you took and why you took them in building the model.
  - It will include brief descriptions of the model assumption checks.
  - It will include a mathematical statement for the final model you fitted.
- Write an Executive Summary.

Question 1. [17 Marks]

Can the performance of professional golf players in the first round of a professional tournament be used to predict how well they will do in the final round? A sports journalist believes it can - the better golf players do on the first round, the better they do overall and furthermore the first round score can be used to give a decent forecast of the final score. In professional golf tournaments there are four rounds. Each round, the players get a score which is the difference between the number of strokes they took and the par* score for the course, so a score of -5 means they took 5 fewer shots than par while a score of 8 means they took 8 more shots than par. Lower values are better. Data was gather for a random sample of 100 players from the US master over the last 30 years. The US masters is one of the one of the four major championships in professional golf and is always played at the same golf course (Augusta National Golf Club).

The data collected is stored in the text file "Golf" which contains the variables:

`Round1`	the players score for the first round of the tournament.
`Final`	the players final score for the tournament (i.e., the total score for the four rounds).

Is there a relationship between the first round score and the total score? How useful is this relationship for prediction? In particular, what final score would we predict for a player who scored par (i.e. 0) on round 1.

In addition to the usual analysis, recreate the original plot of the data showing the final fitted model.

* In golf, "par" is the number of strokes an expert golfer is expected to need to complete an individual hole or to complete all the holes on a golf course.

Question 2. [5 Marks]

Revisit the data from question 1. The journalist has posed the following situation: I believe that the winner of the tournament will have a final score that is at least 8 under par (so -8 or lower). Based on the first round scores, who can we confidently rule out from winning?
Use prediction intervals to answer this question and show this information on a plot.

Question 3. [18 Marks]

A film archivist was interested in being able to predict the length of films using how much film was on the reel. Rather than unwinding and counting the actual length, it is much quicker to just measure the diameter of film on the film reel and use this to predict the length of the film. A random sample of 16 mm films (all on the same style of reel) were selected and their diameter on the reel, as well as the actual length of the film was measured. The data collected is stored in the text file "Film1" which contains the variables:

`Length`	the actual length of film on the reel (in feet).
`Diameter`	the diameter of the film on the reel (in inches).

Can we build a model to estimate the actual length of film on a reel using the diameter of film? What is the estimated equation of this model? How useful is this model for prediction? In particular, predict the length of a film which has an 8.4 inch diameter on the reel.

In addition to the usual analysis, recreate the original plot of the data showing the final fitted model.