ENGSCI 311 2016 – Data Analysis Assignment
DUE: Wednesday 14 September at 7:00pm on Canvas

This assignment requires you to conduct statistical analyses on three data sets. For each task, we expect to see the following, as done in the case studies discussed in lectures:

- exploratory analysis
- checking modelling assumptions via appropriate plots
- model building / variable selection as appropriate
- appropriate inference, including predictions where required
- reports: models and assumption check and an executive summary

In your submission, you should include all your R code and output from the R console, including all plots produced by R.

Submission

Each task should be prepared as a **separate document** and converted to a **single PDF file**, which should be submitted to the appropriate Canvas dropbox prior to the due date.

Clear and succinct communication is an important part of Engineering, regardless of specialisation. We expect that you will write clear and concise English detailing your understanding of the material. In Executive Summaries, this means desribing analysis in context, not using variable names, using units when known, rounding sensibly and not using technical language (e.g. *p*-value).

Most of the marks in each task are allocated to the Methods and Assumption Checks and Executive Summary. These must be consistent with your R output for credit.

For R code and output, we ask that you use a `fixed-width font` such as Courier New or Consolas.

You may wish to hand-write your Models and Assumption Checks and Executive Summaries. This is permitted as long as you merge your files such that only one PDF file is submitted per task.

There will be penalties for not following instructions!

Late submissions will be penalised per the policy on Canvas.

Rmarkdown

This is NOT compulsory. You do not need to use Rmarkdown to attain full marks.

To get started, a Rmarkdown template has been provided on Canvas. If you have RStudio installed, you can publish your R code, output, plots and comments directly without copying / pasting. You should examine the template carefully, taking note of how commands relate to published output.

Open the template provided in RStudio and click the Knit button. Relevant packages will be installed automatically. You should choose to Knit Word (if you have Microsoft Word installed), or Knit HTML. Knit PDF only works if you have a LaTeX distribution installed.

This is *entirely optional* – it is completely acceptable to produce your assignment by copying and pasting R code and output directly into a word processor of your choice.

Assistance available

Piazza is the best place to receive assistance from your peers and your lecturer.

Kevin is available in person at the following times in Weeks 4 – 6:
- Mon 3:30 – 4:30pm in 439-223, Level 2 (left from elevators) UniServices House, 70 Symonds Street
- Wed 1 – 2pm in the Lech Study Area

Office hours during the mid-semester break will be announced in due course.

Piazza will be **closed to new questions in the 19 hours before the assignment is due.**
DO NOT LEAVE QUESTIONS TO THE LAST MINUTE.

Task 1: Gas Turbines

Failures in aircraft gas turbine engines due to high cycle fatigue is a pervasive problem. A single crystal nickel base superalloy was designed and manufactured to address this problem. We are interested in fitting a model to estimate mean cycles to failure from strain amplitude for turbine blades made of this new material.

The file `gasturbines.txt` contains the following variables:

$$
\begin{array}{ll}
\texttt{Strampl} & \text{strain amplitude} \\
\texttt{Cycfail} & \text{cycles to failure}
\end{array}
$$

Hints:

- we suspect that a multiplicative relationship exists between the two variables. You should check whether this is the case thorough appropriate diagnostic plots.
- remember to check for the effect of influential observations.

Task 2: Hot Dogs

Hot dogs contain large amounts of sodium which is detrimental to health. We are interested in building a model to estimate the effect of the different meat types and calorie count on sodium content. We also want to predict the sodium content of a single hot dog made from poultry containing 425 Calories.

The file `hotdogs.txt` contains the following variables:

`Sodium`	Milligrams of sodium per hot dog
`Type`	Type of hot dog (beef, meat, or poultry)
`Calories`	Calories per hot dog

Hints:

- check whether there is interaction between the explanatory variables.
- you will need to *rotate* the factors to make all possible comparisons between the pair of meat types.

Task 3: Wire Bonds

The pull strength of a wire bond is an important characteristic when mass producing components with integrated circuits. We are interested in building a *simple* model to investigate the relationship between pull strength and other measured characteristics.

The file `wirebonds.txt` contains the following variables:

`pullStr`	pull strength		
`dieH`	die height	`wireL`	wire length
`postH`	post height	`bondWidthDie`	bond width on the die
`loopH`	loop height	`bondWidthPost`	bond width on the post

Hints:

- you should treat all variables as continuous.
- some *variable selection* will be required.