ST303/ST633 Linear Models
Assignment Sheet 1

Due: Friday 21th Oct 11:59am

- Only one, randomly chosen question will be marked.
- Use R to answer questions 4 and 5.
- 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 1"
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.

1. Prove the the following properties of ordinary least squares residuals and fits for the regular simple linear regression model.

   (a) $\sum_{i=1}^{n} e_i = \sum_{i=1}^{n} (y_i - \hat{y}_i) = 0$

   (b) $\sum_{i=1}^{n} x_i e_i = \sum_{i=1}^{n} x_i (y_i - \hat{y}_i) = 0$

   (c) $\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{y}_i$

   (d) $\sum_{i=1}^{n} \hat{y}_i e_i = 0$

2. (a) Consider the model

$$y_i = \beta_0 + \epsilon_i, \qquad i = 1, ..., n$$

   where ϵ_i are assumed i.i.d. N(0, σ^2).

   i. Find the ordinary least squares estimator of β_0.

   ii. Find an estimator $\hat{\sigma^2}$ of σ^2.

   (b) Consider the model

$$y_i = \beta_1 x_i + \epsilon_i, \qquad i = 1, ..., n$$

   where ϵ_i are assumed i.i.d. N(0, σ^2), is called the no-intercept simple linear regression model.

   i. Find the ordinary least squares estimator of β_1.

   ii. Find an estimator $\hat{\sigma^2}$ of σ^2.

3. A simple linear regression model may be written as either:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \qquad\qquad i = 1, 2, ..., n$$

or

$$y_i = \alpha_0 + \alpha_1 (x_i - \bar{x}) + \epsilon_i \qquad\qquad i = 1, 2, ..., n$$

   (a) Interpret the parameters $\beta_0, \beta_1, \alpha_0$ and α_1 and find the relationship between the α parameters and the β parameters.

   (b) Using the method of least squares derive expressions for the estimators of α_0 and α_1.

   (c) Find the expected values and variances of $\hat{\alpha}_0$ and $\hat{\alpha}_1$.

   (d) Show that $\text{Cov}(\hat{\alpha}_0, \hat{\alpha}_1) = 0$.

   (e) Give a reason why mean-centering the predictor variable may be useful.

4. As concrete cures, it gains strength. The following data represent the 7-day and 28-day strength (in pounds per square inch) of a certain type of concrete.

7-Day	28-Day
2300	4370
2430	4640
2890	4620
3120	4900
3380	5020
3390	5220

   (a) Using R, plot 28-Day strength versus 7-Day strength. Does it seem appropriate to assume a linear relationship between the two variables?

   (b) Specify algebraically a simple linear regression model for these data.

   (c) Fit this SLR in R and draw the fitted line on your scatterplot.

   (d) Provide a printout of the output with relevant parts highlighted.

   (e) The regression coefficients are given by the R command `coef(fit)`. Verify (by hand!) the slope and intercept parameters, and also estimate (again, by hand!) the model's variance parameter.

   (f) Interpret the parameter estimates.

   (g) Compute the standard errors of the slope and intercept estimates (by hand, and in R).

   (h) Is there evidence of a linear association between mean 28-Day strength and the strength at 7 days?

   (i) Verify the value of the coefficient of determination for this model by hand and interpret it.

5. The data in Life.csv gives the life expectancy and per capita income for various countries. The variable code indicates if the country is industrialised (code=1) or not (code=2). Answer the following questions using R for industrialised countries only and excluding South Africa.

   (a) Draw a scatterplot of life expectancy versus per capita income.

   (b) Fit the simple linear regression of life expectancy on capita income.

   (c) Draw the fitted line on the scatterplot.

   (d) What are the estimates of the slope and intercept and what are their standard errors?

   (e) Is there evidence of a linear association between mean life expectancy and per capita income?

   (f) Find the coefficient of determination for the simple linear regression model and interpret it.