ST303/ST633 Linear Models
Assignment Sheet 3

Due: Fri 2nd Dec, 11:59am.

- Only one, randomly chosen question will be marked.
- For Questions 2 and 3 do all calculations by hand. You can check your working using R.
- Answer Question 4 using R.
- Submit questions 1, 2 and 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 3"
output: pdf_document
author: Jane Doe 1234567
---
```

- 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 3, answer questions and help students getting started with R (finding/reading in data, knitting an Rmd file etc).

1. Which of the following regression models are linear? Give reasons.

   (a) $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \epsilon$

   (b) $y = \beta_0 + \beta_1 10^x + \epsilon$

   (c) $y = (\beta_0 + \beta_1 x)/(\beta_0 + \beta_2 x) + \epsilon$

   (d) $y = \exp(\beta_0 + \beta_1 x + \epsilon)$

   (e) $y = \exp(\beta_0 + \beta_1 x) + \epsilon$

2. Suppose you have the following data

i	y_i	x_{i1}	x_{i2}
1	62	2	6
2	60	9	10
3	57	6	4
4	48	3	13
5	23	5	2

   and want to fit the model $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i$.

   (a) Write down the model in matrix format and specify what each matrix / vector is.

   (b) Calculate the least squares estimates using matrix manipulations.

(c) Calculate the fitted values and residuals.

(d) What is the estimate of σ^2?

3. A candidate model for the data below is

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i1}^2 + \beta_3 x_{i2} + \epsilon_i$$

i	y_i	x_{i1}	x_{i2}
1	6.1	1	0
2	5.4	2	0
3	7.2	3	0
4	8.9	1	1
5	9.1	2	1
6	11.6	3	1

(a) Specify the model in matrix notation writing out each matrix / vector.

(b) Using matrix manipulations, compute the least squares estimates.

(c) Compute the vector of residuals.

(d) Estimate σ^2.

(e) Calculate a 95% confidence interval for the mean response when $x_1 = 1$ and $x_2 = 1$.

(f) Calculate a 95% prediction interval for the response when $x_1 = 1$ and $x_2 = 1$.

(g) Verify the adjusted R^2 value by hand and comment on its usefulness in this context.

4. Accurate measurement of body fat involves underwater weighing and is inconvenient and costly. It is desirable to have easy methods of estimating body fat. Some health books explain how the reader can estimate body fat from tables using their age and various skin-fold measurements using a caliper. Other texts give predictive equations for body fat using body circumference measurement (e.g. abdominal circumference) and / or skin-fold measurements.

   The goal is to come up with a percent body fat estimation method which relies on simple physical measurements.

   The data (bodyfat.txt) consists of observations taken on a sample of 88 males. The following variables were measured: percent body fat, age (years), weight (pounds), height (inches) neck circumference (cm), abdomen circumference (cm), knee circumference (cm) and ankle circumference (cm).

   (a) Fit the regression model with all predictors included.

      i. Comment on your results. Which predictors have a positive impact on percent body fat? Which have a negative impact? Which variables are significant? Which are insignificant?

      ii. Estimate the amount by which percent body fat changes with each year of age. Compute a 95% confidence interval for this quantity and interpret your result.

      iii. Predict the percent body fat and obtain a 95% prediction interval for Tom whose age=49, weight=188, height=68, neck=37, abdomen=90, knee=38, ankle=24. Predict the percent body fat and obtain a 95% prediction interval for Bill whose age=40, weight=220, height=76, neck=40, abdomen=113, knee=34, ankle=20. Whose percent body fat is predicted more precisely? Can you give any reasons for this?

(b) Show (i.e. perform a suitable hypothesis test) that it is appropriate to omit both the predictors weight and ankle. Using the reduced model:

i. Comment on your results. Which predictors have a positive impact on percent body fat? Which have a negative impact? Which variables are significant? Which are insignificant?

ii. Estimate the amount by which percent body fat changes with each year of age. Compute a 95% confidence interval for this quantity and interpret your result. Are these findings about age much different to those found above?

iii. Predict the percent body fat and obtain a 95% prediction interval for Tom and Bill. Whose percent body fat is predicted more precisely? Why is this? Compare these results to those of the intervals found above.

Some code you may find useful is below:

```
bodyfat <- read.table("bodyfat.txt", header = TRUE)
pairs(bodyfat)
fit <- lm(bfat~., data = bodyfat)
summary(fit)
confint(fit)
Tom <- c(49, 188, 68, 37, 90, 38, 24)
Bill <- c(40, 220, 76, 40, 113, 34, 20)
newx <- rbind(Tom, Bill)
newx <- data.frame(newx)
names(newx) <- colnames(bodyfat)[2:8]
predict(fit, newdata = data.frame(newx), interval = c("confidence"),
              level = 0.95, type="response")
```