FIT3152 Data analytics. Tutorial 06: Regression

1. The ‘diamonds’ data set comes packaged with ggplot2 and contains data about the price of diamonds as well as information on size as well as the 4 Cs affecting diamond price: carat (size), cut, colour and clarity.

(a) Taking a random sample using the code below, create a subset of the diamonds data set: ‘dsmall’ to use in the following analysis.

```
install.packages("ggplot2")
library(ggplot2)
set.seed(9999) # Random seed to make subset reproducible
dsmall <- diamonds[sample(nrow(diamonds), 1000), ] # sample of 1000 rows
```

(b) Using the data ‘dsmall’ calculate the regression of ln(price) on ln(carat) and each of the remaining categories (clarity, color and cut) separately. Which of clarity, color or cut has the greatest effect on price? Which has the least? Justify your answer using regression output.

2. The file “body.dat.csv” contains data from a study on the relationship between body dimensions. The study measured 500+ active individuals.

The data was obtained from http://www.amstat.org/publications/jse/jse_data_archive.htm
A related article is http://www.amstat.org/publications/jse/v11n2/datasets.heinz.html

(a) Test the hypothesis that men are taller than women on average. Assume a significance of 5%

(b) Test the hypothesis that men are heavier than women on average. Assume a significance of 1%

(c) BMI is calculated as $\frac{weight\ (kg)}{(height\ (m))^2}$. Test the hypothesis that men have a higher BMI than women on average

(d) Calculate the regression of Height on the other body measurements for men and women separately. Which measurements are the most significant predictors of height for each gender?

3. The data file “Dunnhumby1-20.csv” is a cut down and modified set of test data from the Kaggle competition to predict when consumers would next visit a Dunnhumby supermarket and how much they would spend. See: http://www.kaggle.com/c/dunnhumbychallenge for more information. The current modified data set contains the customer ID, Date of visit, Days since last visit (Delta), and Spend for 20 customers from the test set.

Calculate the regression of Spend vs Delta for each customer and summarize the results in a data frame similar to that below. *Hint: try using “plyr” package and dlply function.*

CustomerID	RegIntercept	RegSlope

4. Using the data from the UCI Machine Learning Repository comment on the factors affecting red wine quality. Data site is: http://archive.ics.uci.edu/ml/datasets/Wine+Quality The file name is: winequality-red.csv.

5. Install the “ISLR” library. Using the “Carseats” data, calculate the regression equation predicting Sales (child car seat sales) as a function of the input variables. Which variables are significant predictors?

6. The text, G. James et al., An Introduction to Statistical Learning: with Applications in R (ISLR) uses the “Advertising” data set to illustrate a number of different learning models. A description of the data (p15) follows: “The Advertising data set consists of the sales of that product in 200 different markets, along with advertising budgets for the product in each of those markets for three different media: TV, radio, and newspaper. A copy of the data was downloaded from: https://www.kaggle.com/ashydv/advertising-dataset and is on Moodle.

Using the Advertising data, answer the following questions (taken from pp59-60 ISLR):

(a) Is there a relationship between advertising budget and sales?
(b) How strong is this relationship?
(c) Is the relationship linear?
(d) Which media contribute to sales?
(e) How accurately can we estimate the effect of each medium on sales?
(f) (Extension) Is there synergy (interactions) among the advertising media?

Potential ways of addressing these questions using regression models and extensive discussion of regression can be found on pages 59-82 of ISLR.