COMP257 2017 Quiz 1

Answer the questions in the spaces provided.
This exam has 7 questions, for a total of 85 points.

Name: _____

Student ID: _____

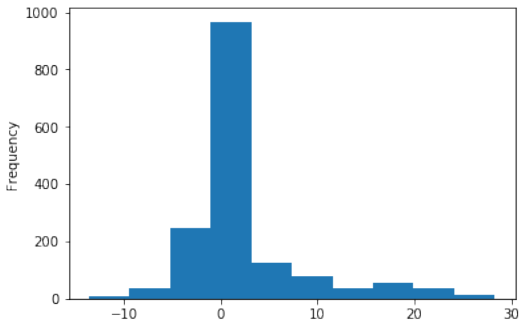
1. (15 points) In the CRIPS-DM methodology, there is a back-link between the Evaluation phase and the first phase of the process, Business Understanding. Explain what the Evaluation stage should involve and how it might inform the next iteration through Business Understanding.

Solution: The evaluation stage involves evaluating the performance of the model that has been developed and trained on the data. This will be different depending on the kind of model. It will probably involve testing the model on unseen data to see how well it performs on data that it has not been trained on. This gives an indication of how well the model will perform in the real world.

In other cases the evaluation may be about how well the model helps to explain patterns in the data - for example whether we were able to produce a visualisation of the data that helps to understand what is happening.

The link back to the BU phase is there because things we find out during evaluation might raise new questions relevant to the business that we want to follow up in the next cycle. For example it could be that the performance of the model is not good and we need to collect new data to add in, or a new category might be identified through visualisation that is worth following up.

2. (10 points) The following plot shows the distribution of some data. How would you describe this distribution? Point out any significant features that you observe.



Solution: The mean of the data is close to zero but there is a large tail of data to the right - the data is skewed to higher values and shows high kurtosis as the data is spread out a lot away from a normal distribution.

There is a secondary peak in the data around 18 suggesting that there might be two distinct populations represented here. These should be checked to see if they are outliers. We should try to see if this second group is identified by some other variable.

3. (15 points) I'm looking at some data from a drug trial, trying to find out whether the people treated with the drug had different outcomes from those that have not been treated. The null hypothesis is that the two groups have the same outcomess. I perform a *Fisher Test* on the data and get a *p-value* of 0.0322.

   Explain what the *p-value* means in this context. What is the answer to my question (is the drug effective) given the *p-value* of 0.0322?

   Solution: The p-value is the probability that the observed difference in the means of the two samples would occur by chance if the samples were drawn randomly from the population. We reject the null hypothesis if the probability that this distribution is observed is very low. The typical level of probability is 0.05 but some people suggest that 0.01 or even lower is more appropriate.

   A p-value of 0.0322 means that this distribution would only occur 3.22% of the time in samples drawn from the population. This is below the traditional 5% threshold so I would reject the null hypothesis and accept that the drug has made a difference.

4. (10 points) When the current Same Sex Marriage postal survey was announced, there were some concerns voiced that the survey would be biased because young people are not familiar with using the postal system. Assuming this to be true (young people don't know how to post letters) explain why this might lead to bias in the result of the survey. Under what circumstances would this not lead to bias?

   (Note this is not a political question – it's about surveys.)

   Solution: If the younger population were under-represented in the survey and if on average the views of younger people were different to those of older people, then the end result of the survey would not contain as many of these responses and would contain a bias.

   For example, if more young people would vote Yes than in the general population, and fewer of them were represented in the result, the overall Yes vote would be reduced and therefore not representative of the views of the overall population.

   This would not lead to bias if the responses of younger voters had the same distribution as that of the population in general - that is, if age were not a predictor of the outcome of the survey.

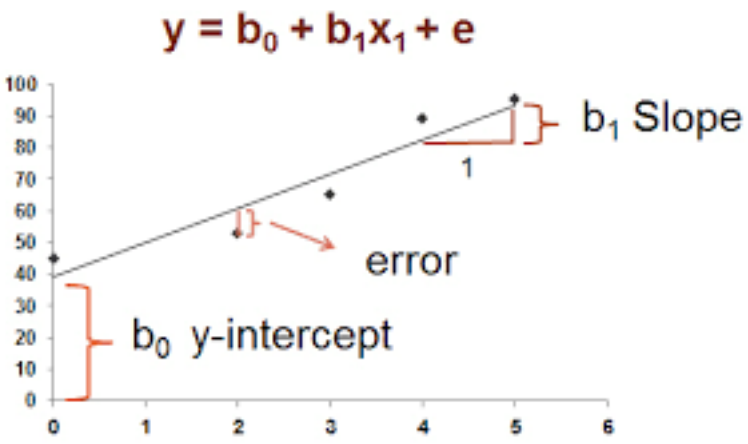
5. (10 points) Explain what the terms in the standard static linear regression model equation below mean?

$$y_i = b_0 + b_1 x + e$$

Solution: y_i - estimated or predicted y value b_0 - estimate of the intercept b_1 - estimate of the regression slope or gradient x - the independant variable e - the error term

6. (10 points) Draw a graph showing the simplest representation of the terms in $y_i = b_0 + b_1 x + e$.

Solution:



7. (15 points) Explain the steps you would take in constructing a static linear regression model with two variables taking into account any data anomalies?

Solution: 1)Conduct preliminary analyses : i.Examine descriptive statistics of the continuous variables ii.Check the normality assumption by examining histograms of the continuous variables iii.Check the linearity assumption by examining correlations between continuous variables and scatter diagrams of iv.the dependent variable versus independent variables.

2)Conduct multiple linear regression analysis i. Run model with dependent and independent variables

3) Check Model i. Examine collinearity diagnostics to check for multicollinearity ii. Examine residual plots to check error variance assumptions (i.e., normality and homogeneity of variance) iii. Examine influence diagnostics (residuals, dfbetas) to check for outliers iv. If outliers and /or leverage points are valid data points, consider modifications to the model otherwise exclude v. Examine significance of coefficient estimates to trim the model

4. Revise the model and rerun the analyses based on the results of steps i-iv. 5. Write the final regression equation and interpret the coefficient estimates