Biostatistics 5MT013 - - Written Assignment for **Zhouhui Qi**

- This is an individual assignment. You are being given a dataset which is unique to you and one other student. You are encouraged to discuss with this person, but you must write your own report/use your own words (this will be checked).
- You should answer the questions written in your individual assignment. Your answers should be written as a short report. You should keep to the length requirements assigned to each question.
- Reports should be written using font size 11 and figures should be included in the main report. The figures should be exported to jpeg images before they are included in the report, so that your report does not become too large (the file size should be smaller than 5MB).
- You will need to use R to analyse the given data. Include an appendix with the R code you use. The appendix (which should include only code) does not count towards the page limit.
- Your report should be uploaded to Canvas (under Assignments).

Your data set is: **assignment8_KW-2449_RUNX1.RData**

Your data represents a selection from the course data and includes three objects *clinical_data*, *drug_data*, *gex*. We have restricted the data to only include the individuals included in the tests using **KW-2449** (n=**292**). We have also filtered the gene expression data (it contains data for **6452** genes).

1. (max. 1 side A4) [10 points]

   (i) For individuals with/without (positive/negative for) mutations in **RUNX1**, separately, describe in words the distributions of AUC for **KW-2449** - refer to, and include, a table or graphical representations.

   (ii) Compute 95% confidence intervals for AUC for **KW-2449** for patients with and without mutations in **RUNX1** – use central limit/normal distribution assumptions.

   (iii) Do you think that the assumptions for the confidence intervals calculated above are reasonable? Answer in words and refer to a plot (which you should include) that explicitly examines the assumption(s). If the assumptions are not (or if they had not been) reasonable, what alternative approach could you have used?

2. (max. 1 side A4) [14 points]

   (i) Identify genes that are associated with AUC for **KW-2449**, at a suitable threshold for false discovery. State the threshold and report how many genes are selected.

   (ii) What is the null hypothesis of each test used to identify whether a gene expression is associated with drug response?

   (iii) Summarise the distribution of p-values graphically, and in words.

   (iv) Select the gene with the strongest evidence of association, together with mutation status (of **RUNX1**), and use these two variables as covariates in a linear regression model for AUC (for **KW-2449**). Is there any evidence that **RUNX1** status confounds the association between the expression level of this gene and drug response/AUC? Is **RUNX1** a potential confounder for the association between other gene expressions and AUC? Explain carefully.

3. (max. $1\frac{1}{4}$ sides A4) [10 points]

   (i) Based on gene expressions (use all genes), perform a cluster analysis of individuals. Describe the distance measure and clustering approach you have used. Select the clusters at level 5 of the dendrogram. How many individuals/samples are found in each of the clusters? You don't need to include the dendrogram as a figure in this part, since you are asked to incorporate it in the next part of this question.

   (ii) Produce an informative heatmap of the expression data that shows the clustering of individuals and genes (two dendrograms) and additionally provides annotation that shows how (if at all) **RUNX1** mutation status relates to the clusters. Do you see a relationship between the 5 clusters that you chose (above) and **RUNX1** mutation status?

4. (max. $1\frac{1}{4}$ sides A4) [16 points]

   (i) Select the 100 genes with the highest variance. Construct a linear regression model to predict the AUC for **KW-2449** using the expression data from these 100 genes simultaneously. Include the gene expressions on their original scale (no transformations) and do not include interaction terms. For individual **BA2409**, what are the observed and predicted AUCs? Include a plot of the in-sample predictions against the observed values (all individuals) in your report. Do you think the assumptions of the model are met? In answering this question, include an additional plot and describe the plot in words.

   (ii) Use a correlation coefficient to summarise how well the prediction performs – report first the in-sample correlation coefficient (i.e. applying the prediction model to the training data) and then the correlation coefficient using leave-one-out cross-validation. Describe how/why the two estimates of correlation differ (how they should be used).