Queen Mary
University of London

Specific Learning Difficulties Cover Note

Student ID: 170857044

Advice for assessors and examiners

Guidelines for markers assessing coursework and examinations of students diagnosed with Specific Learning Difficulties (SpLDs) –
As far as the learning outcomes for the module allow, examiners are asked to mark exam scripts sympathetically, ignoring the types of errors that students with SpLDs make and to focus on content and the student's understanding of the subject. Specific learning difficulties such as Attention Deficit Disorders, dyslexia and or dyspraxia may affect student performance in the following ways:

- The candidate's spelling, grammar and punctuation may be less accurate than expected
- The candidate's organisation of ideas may be confused, affecting the overall structure of written work
- The candidate's proof reading may be weak with some errors undetected, particularly homophones and homonyms which can avoid spell checkers

Under examination conditions, these difficulties are likely to be exacerbated. Errors are likely to become more marked towards the end of scripts.

Useful approaches can include:

- Reading the passage quickly for content
- Including positive/constructive comments amongst the feedback so that students can work with specialist study skills tutors on developing new coping strategies
- Using clear English and when correcting; explain what is wrong and give examples
- Using non-red coloured pens for comments/corrections

Colleagues in schools are asked to ensure that students with specific learning difficulties access the support provided by the Disability and Dyslexia Service.

For more information regarding marking guidelines see DDS webpage http://www.dds.qmul.ac.uk/staffinfo/index.html and the Institutional Marking Practices for Dyslexic Students

Disability and Dyslexia Service
Room 3.06, Francis Bancroft Building
Queen Mary, University of London
Mile End Road
London
E1 4NS

T: +44 (0) 20 7882 2756
F: +44 (0) 20 7882 5223
E: dds@qmul.ac.uk
W: www.dds.qmul.ac.uk

Alteration or misuse of this document will result in disciplinary action

Dataset 1 – Marine Microbial Diversity.

Code:

```
#Find and open dataset 1
Mar_Micro <- read.table(file.choose(new = FALSE ), header = TRUE)#choose the part_1_student_1493.tdf
str(Mar_Micro)
attach(Mar_Micro) #do this to attach it inorder to utalise the variables within the table without having to type Mar_Micro$ everytime.
hist(UniFracInd) #to see if data is normally distributed.
plot(UniFracInd~latitude, col="pink") #auto produces a boxplot with the axis labels.
plot(UniFracInd~season, col="pink")
#get the means for each to create a table of 4 values
Mean_jan_trop <- sum(UniFracInd[1:10])/10
Mean_aug_trop <- sum(UniFracInd[11:20])/10
Mean_jan_temp <- sum(UniFracInd[21:30])/10
Mean_aug_temp <- sum(UniFracInd[31:40])/10

#table for the means to visulise
SeasonXLat <- matrix(c(Mean_jan_trop, Mean_aug_trop,Mean_jan_temp,Mean_aug_temp), nrow = 2)
rownames(SeasonXLat) <- c("January","August")
colnames(SeasonXLat) <- c("Tropical","Temperate")

model1 <- lm(UniFracInd ~ latitude * season)
anova(model1)
summary(model1)
par(mfrow = c(2,2))
plot(model1)
par(mfrow = c(1,1))
```

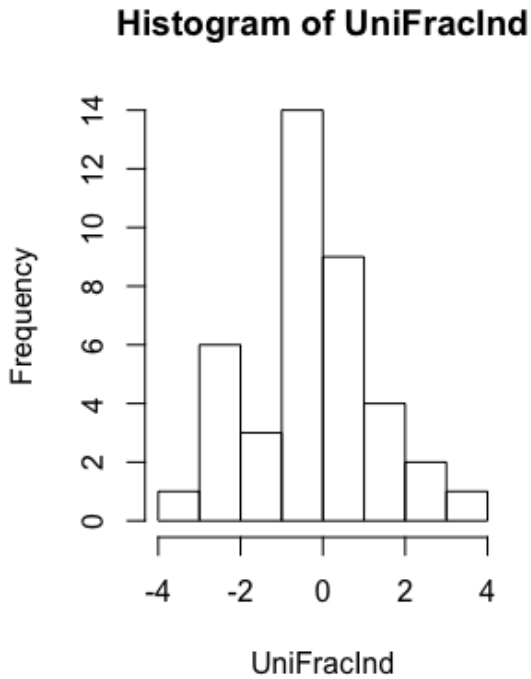

Figure 1. Histogram of the expression of microbial diversity within sea water using UniFrac, relative to the reference sample from Plymouth. From this histogram we can see that the microbial diversity values are normally distributed and thus statistically viable for analysis.

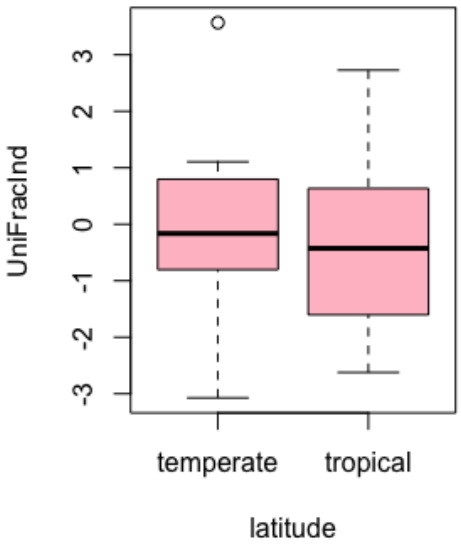

Figure 2. Boxplot of the effects of latitude against microbial diversity in sea water. Within this figure we can see that there is one data point for the temperate latitude that falls outside of the boxplot and the standard deviation bars. However, this has no real effect on the result.

It is likely that latitude does have an effect on the microbial diversity. The temperate latitude shows a greater diversity compared to the tropical latitude as there are more positive values despite there being a smaller inter quartile range. The impact of this effect is nominal as the mean value of diversity for each latitude is fairly similar to one another.

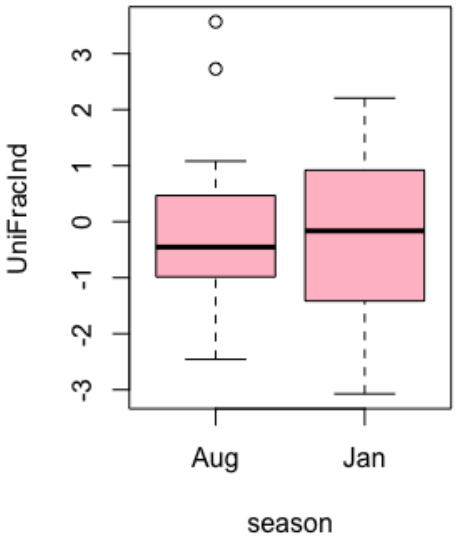

Figure 3. Boxplot of the effects of season against microbial diversity in sea water. Within this figure we can see that the August (Aug) season has two outlying data points that are much greater than the rest of the points which could possibly increase the mean for that data set.

The difference in season when samples were collected has an effect on the microbial diversity seen within these samples. The samples taken in January (Jan) show more positive values within the boxplot. The mean of this data set (0) is also much higher than that of the August data set (-0.5) indicating that in January the microbial diversity is greater compared to that in August.
Although in contrast the boxplot has a greater inter quartile range and a mean of 0, indicating that compared to the reference sample there was not much change in diversity.

Table 1. Summary of one-way ANOVA, testing the effects of latitude and season on microbial diversity using a linear model.

	DF	SUM SQUARES	MEAN SQUARES	F VALUE	P VALUE
LATITUDE	1	0.521	0.52131	0.2221	0.6403
SEASON	1	0.206	0.20640	0.0879	0.7685
LATITUDE:SEASON	1	0.612	0.61194	0.2607	0.6128
RESIDUALS	36	84.512	2.34754		
TOTAL	39	85.851			

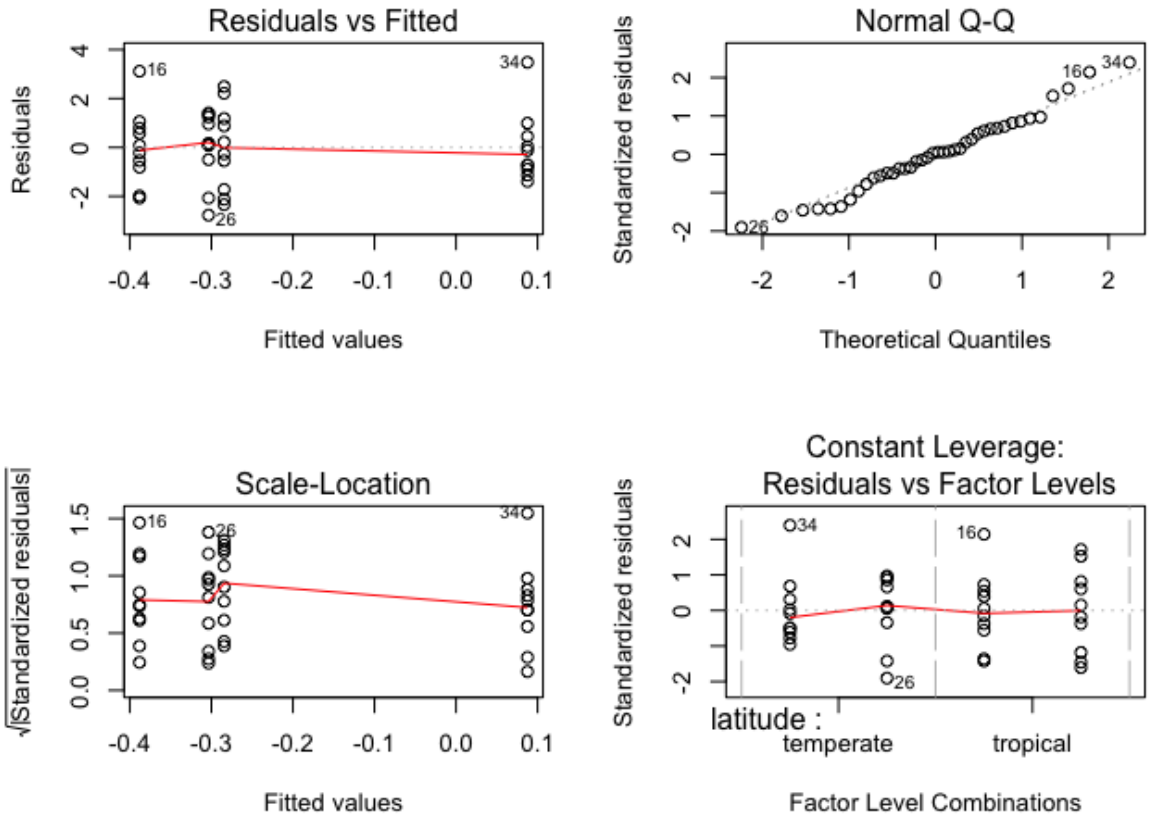

Figure 4. Diagnostic plots of residuals from the linear model UniFracInd against latitude and season. There are a few residuals that look extreme, residual 16 and 34. In general for this model they are fairly well fitted in all plots. The Residuals vs Fitted plot shows homoscedasticity but there are some points to consider and take note of at approximately -0.4, -2. The Q-Q plot shows that most of the residuals are normally distributed apart from the aforementioned extremes.

For the ANOVA conducted in *Table 1* the interaction between latitude and season (latitude: season) shows a p-value of 0.6128, which is greater than the alpha 0.05. Due to this we can ascertain that there is no interaction between latitude and season. The other p-values within the table are also greater than the alpha value, this indicates that the means of all the variables present have no statistical significance.
For completeness, another model was also tested (UniFracInd ~ season * latitude). This model produced the same outcome as the first with the p-values for all variables and interactions being greater than the alpha 0.05 value. Therefore, it was concluded that the means of the data present had no statistical significance as well as there being no interaction between the variables.

Dataset 2 – Pairwise Nucleotide Sub and RNA expression.

Code:

```
#find an open dataset 2
Nucleo_Sub <- read.table(file.choose(new = FALSE ), header = TRUE)#choose the part_1_student_1493.tdf
str(Nucleo_Sub) #check the structure of the file to ensure it has been loaded in correctly.
attach(Nucleo_Sub) #attach the Nucleo_Sub object, so you can utalise the factors within the file.
plot(distance,expression_fold, ylab = "Expression Level", xlab = "Genetic Distance", main = "Change in Luciferase Homolog Expression with Genetic Distance", pch = 16, col = "steelblue")
#this plots the expression variable as the dependent on the y axis and distance as the independent variable on the x axis.
#this in turn allows us to see how the genetic distance changes expression levels.
model2 <- lm(expression_fold ~ distance) #the variables for the model are in this order as expression id dependent on distance which is independent.
summary(model2) #check the model but more importantly find the coefficents for the line y=a+bx
abline(model2, col="red") #plot the liner line of the model against the data points.

par(mfrow = c(2,2))
plot(model2) #these plots will be used to check the assumptions of the linear model.
par(mfrow = c(1,1))
#the QQ-plot for normality of errors looks as if it is normally distributed.
hist(model2$residuals) #nother visulaition to see if the errors are normally distributed.
shapiro.test(model2$residuals) #formal normality test of residuals to check normal distribution.
#accept null hyp - noramlly dist. however it is likely there isnt enough data to see minute changes
```

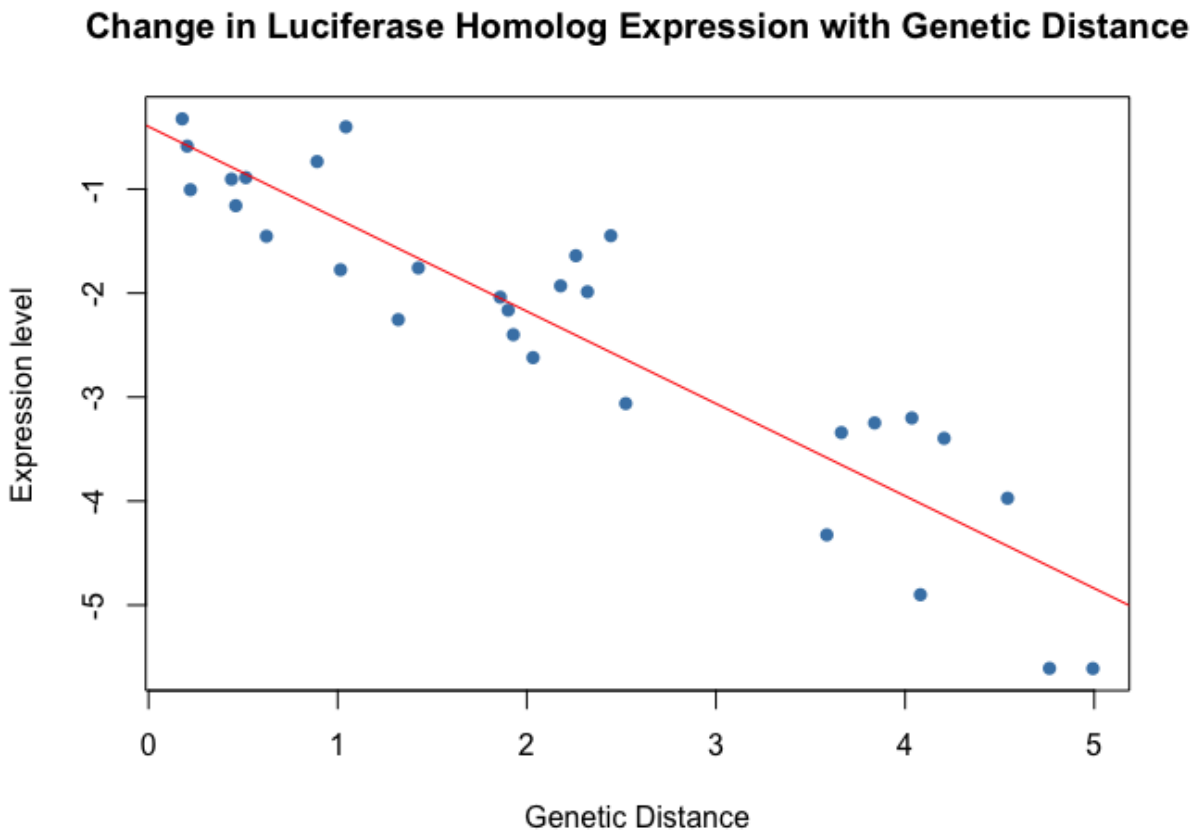

Figure 5. Change in Luciferase homolog expression depending on genetic distance. As genetic distance increases the level of expression decreases, therefore there is a negative correlation between these variables. The line seen in red is the regression line y=a+bx for the linear model expression fold ~ distance, this creates the line y= -0.398 + -0.888x. Note expression level is actually a fold (or power) of an x value.

Putative Luciferase homolog expression decreases as the genetic distance (amount of amino acid substitutions) increases. This can be seen by the negative correlation shown in *Figure 5* This is likely due to the amino acid substitutions effecting the production of the enzyme, inhibiting the process in which it is expressed.

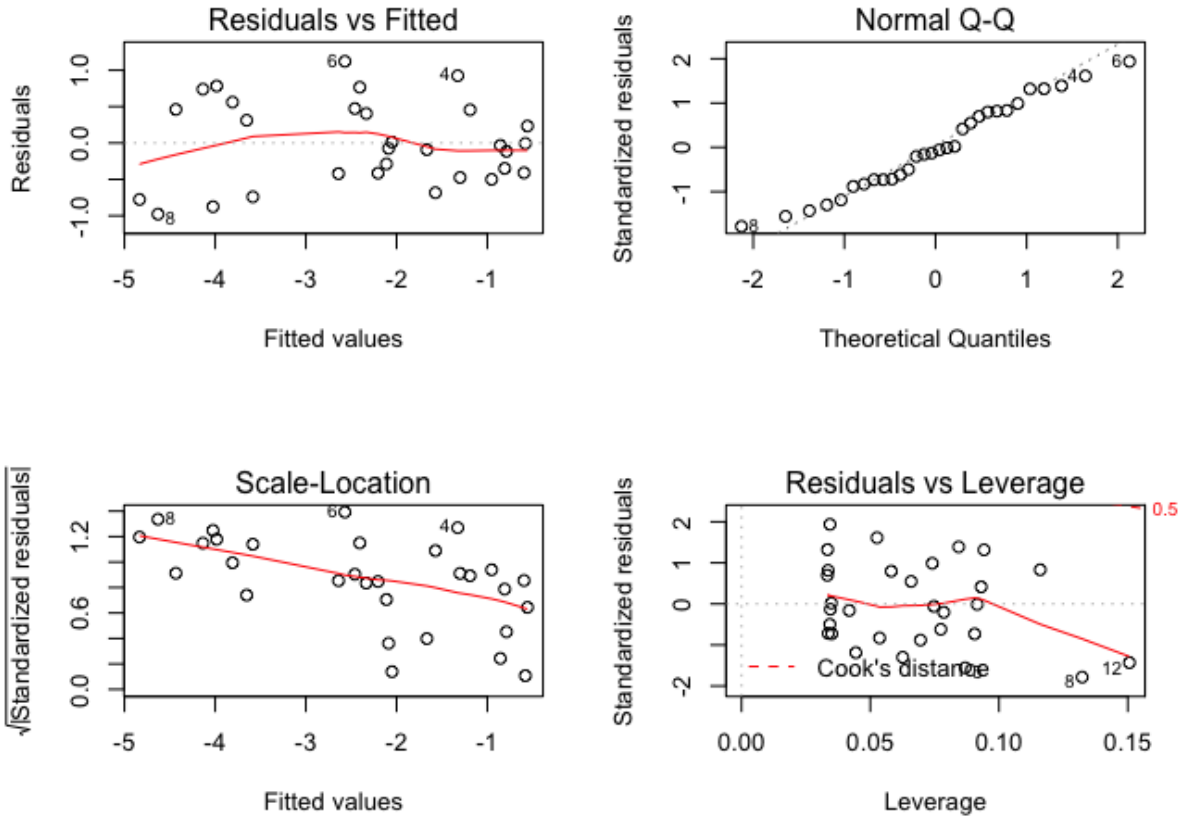

Figure 6. Diagnostic plots of residuals from the linear model expression fold ~ distance. Within this model there are two residuals that stand out from the others, residue 4 and 6. Apart from these extremes the other residuals look well fitted. The residuals are homoscedastic as seen in the Residuals vs Fitted plot, as no obvious pattern in present. the residuals are also normally distributed and linear, this can be seen in the normal Q-Q and Scale-Location plots respectively.

For general linear models there are four assumptions that the model must adhere too. The first is that the data points are all independent of one another and thus hold no influence on the next/previous data point. This is true for the data for this statistical test and can be seen in *Figure 5*.

The next assumption is that the errors for the residuals are normally distributed. One way to see if this is true is to use the Normal Q-Q plot in *Figure 6* the data points are close to the normal distribution line, indicating that the errors are normally distributed. Another way to visualise this is to plot a histogram of the errors. To ensure that the errors are normally distributed a Shapiro-Wilks test was conducted. For this test the null hypothesis is that the errors are normally distributed (SAS Institute inc 2016). The test gave a W-value of 0.96695 and a p-value of 0.45 therefore we accept the null hypothesis as the p-value is larger than the 0.05 threshold. Another indication of normality in the distribution is that the W value for the test is close to 1 (Kirkegaard, E.O.W 2014).

Next the data was carefully looked at to ensure it was homoscedastic in nature. Utilising the Residuals vs Fitted plot in *Figure 6* data points are fairly symmetrical with no clear pattern present indicating homoscedasticity. Finally using the Scale-Location plot of *Figure 6*. it is seen that the residuals are linear. Thus, the model assumptions for this linear model are statistically valid.

Genetic diversity/evolution is a possible effect that is responsible for the relationship seen between expression levels of luciferase and genetic distance. The gene in which amino acid substitutions occur is being mutated via this process, this in turn effects the expression of the protein the gene codes for. The extent of the effect is dependent on the amino acid that is substituted. If the amino acids are similar then a lesser effect is seen, where as if they are very different e.g. Gly substituted for a Leu then the effect can potentially be catastrophic (Davis and Hunter 1987).

Within the context of luciferase and homologs the substitutions are having a lesser effect as luciferase is still being produced however its expression is lessening, indicating the substitutions are similar amino acids. It is likely that these substitutions are having an effect on the structure of the protein causing an inhibition of expression, but further studies would need to be done to support this hypothesis.

Another factor for the substitution is the environment, as the bioluminescence from luciferase expression is likely disadvantageous and therefore selected against. Therefore, over generations and evolution more substitutions occur, increasing genetic distance and decreasing expression. This in turn increases survival of the species, further ingraining the decreasing expression (Clancy 2008: Valiadi, et al 2012).

Dataset 3 – HIV Viral Load and Population Dynamics.

Code:

```
#find and open dataset 3
HIV <- read.table(file.choose(new = FALSE ), header = TRUE)#choose the part_1_student_1493.tdf
str(HIV)
attach(HIV)
hist(VLoad)
hist(score_distance)
hist(score_shannon)

backwards_final <- step(lm(VLoad ~ CD4 * tissue * score_distance * score_shannon), direction =
"backward")
forwards_final <- step(lm(VLoad ~ CD4*tissue), scope = (~CD4*tissue*score_distance*score_shannon),
direction = "forward")

#best both direction
both_final <- step(lm(VLoad~score_shannon*score_distance), scope = c(lower=~score_distance,
upper=~score_shannon*score_distance*tissue), direction = "both")
```

My proposed model to explain viral load in the terms of the other variables present is: vload ~ score_shannon + score_distance + tissue.

Table 2. Model choice using Akaike Information Criterion (AIC) in a stepwise algorithm. The input model for this was VLoad ~ score_shannon * score_distance using an upper scope of score_shannon * score_distance * tissue. Using the upper scope every potential model is included in the table. + indicating the variable is added to the model for that run, - indicating the variable has been removed and tested and none, meaning no change has occurred to the model.

VARIABLE	DF	SUM OF SQUARES	RSS	AIC
NONE			148.22	60.392
+ SCORE_DISTANCE:TISSUE	1	5.025	143.19	61.012
+ SCORE_SHANNON:SCORE_DISTANCE	1	1.134	147.08	62.085
+ SCORE_SHANNON:TISSUE	1	0.247	147.97	62.325
-TISSUE	1	15.956	164.17	62.482
-SCORE_SHANNON	1	159.845	308.06	87.657

Akaike Information Criterion (AIC) is a value that is used to assess the quality of statistic models for given data. One key factor is that AIC is able to estimate quality relative to other models in the table. The lower the AIC score the "better" the model is however, no model will be truly perfect as Box and Draper stated, "Essentially all models are wrong, but some are useful." Therefore, although not perfect the best possible model is still useful to us (Aho, Derryberry and Peterson 2014).

Note that the variable CD4 has not been included within this model. This is due to the AIC scores for models including CD4 where much higher than the proposed model, for example the model VLoad ~ CD4 + score_distance + score_shannon + CD4:score_shannon gave a AIC of 63.42. There was one model that had a similar AIC score to the above model; VLoad ~ CD4 + tissue + score_shannon + score_distance + CD4:tissue + CD4:score_distance + tissue:score_distance with an AIC of 60.86. As all the AIC scores are above the proposed model, we can infer that CD4 is not needed for the best fitting model.

References:

Aho, K., Derryberry, D., and Peterson, T., (2014). Model selection for ecologists: the worldviews of AIC and BIC. *Ecology, Ecological Society of America*, 95 (3), pp. 631 – 636.

Boisbunon, A., Canu, S., Strawderman, W., and Wells, M.T., (2014). Akaike's Information Criterion, Cp and estimators of loss for elliptically symmetric distributors. *International Statistical Review*, 82 (3), pp. 422 – 439.

Box, G., and Draper, N., (1987). Empirical model-building and response surfaces.

Clancy, S., (2008). A single base change can create a devastating genetic disorder or a beneficial adaptation, or might have no effect. How do mutations happen, and how do they influence the future of a species. *Genetic Mutation, Nature Education* 1 (1).

Davis, G., and Hunter, E., (1987). A charged amino acid substitution within the transmembrane anchor of the rus sarcoma virus envelope glycoprotein affects surface expression but not intracellular transport. *The Journal of Cell Biology*, 105, pp. 1191 – 1203.

Kirkegaard, E.O.W., (2014). W values from the Shapiro-Wilk test visualised with different datasets [online]. Available at: http://emilkirkegaard.dk/en/?p=4452 [Accessed 23rd November 2018].

SAS Institute inc., (2016). JMP statistical discovery from SAS [online]. Avalible at: http://www.jmp.com/support/notes/35/406.html [Accessed 23rd November 2018].

Valiadi, M., Iglesias-Rodriguez, M.D., and Amorim, A., (2012). Distribution and genetic diversity of the luciferase gene within marine dinoflagellates. *Journal of phycology*, 48 (3), pp. 826 – 836.