BIO782P - Week 1 Assignment

Ana Penedos (UID 1932)

14th December 2018

packages used: ggplot2 car

Dataset 1: Marine microbial diversity

- 2 cruises: Jan and Aug
- sampled 20l of seawater
- 10 separate locations
- equatorial (0-8 degrees N) and temperate (48-55 degrees N) waters
- samples were filtered for microbiome enrichment
- DNA extracted, WGS sequenced, and sequence reads identified using BLASTN
- Microbial diversity assessed using UniFrac, expressed relative to reference

Data import

The data was imported into Microsoft Excel and checked for obvious issues such as missing values. As none were found, we can import it into R.

```
marine <- read.table("part_1_student_1932.tdf",
                     sep = "\t",
                     header = TRUE)
head(marine) # check if there are obvious issues with data import
```

```
##   UniFracInd season latitude
## 1  3.0116243    Jan tropical
## 2  3.1859106    Jan tropical
## 3 -0.3233971    Jan tropical
## 4 -0.9165660    Jan tropical
## 5  0.3896491    Jan tropical
## 6 -0.0169074    Jan tropical
```

```
str(marine) # check number of observations, variables and their type
```

```
## 'data.frame':    40 obs. of  3 variables:
##  $ UniFracInd: num  3.012 3.186 -0.323 -0.917 0.39 ...
##  $ season    : Factor w/ 2 levels "Aug","Jan": 2 2 2 2 2 2 2 2 2 2 ...
##  $ latitude  : Factor w/ 2 levels "temperate","tropical": 2 2 2 2 2 2 2 2 2 2 ...
```

Season and latitude are factors with two levels each, matching the described experiment.

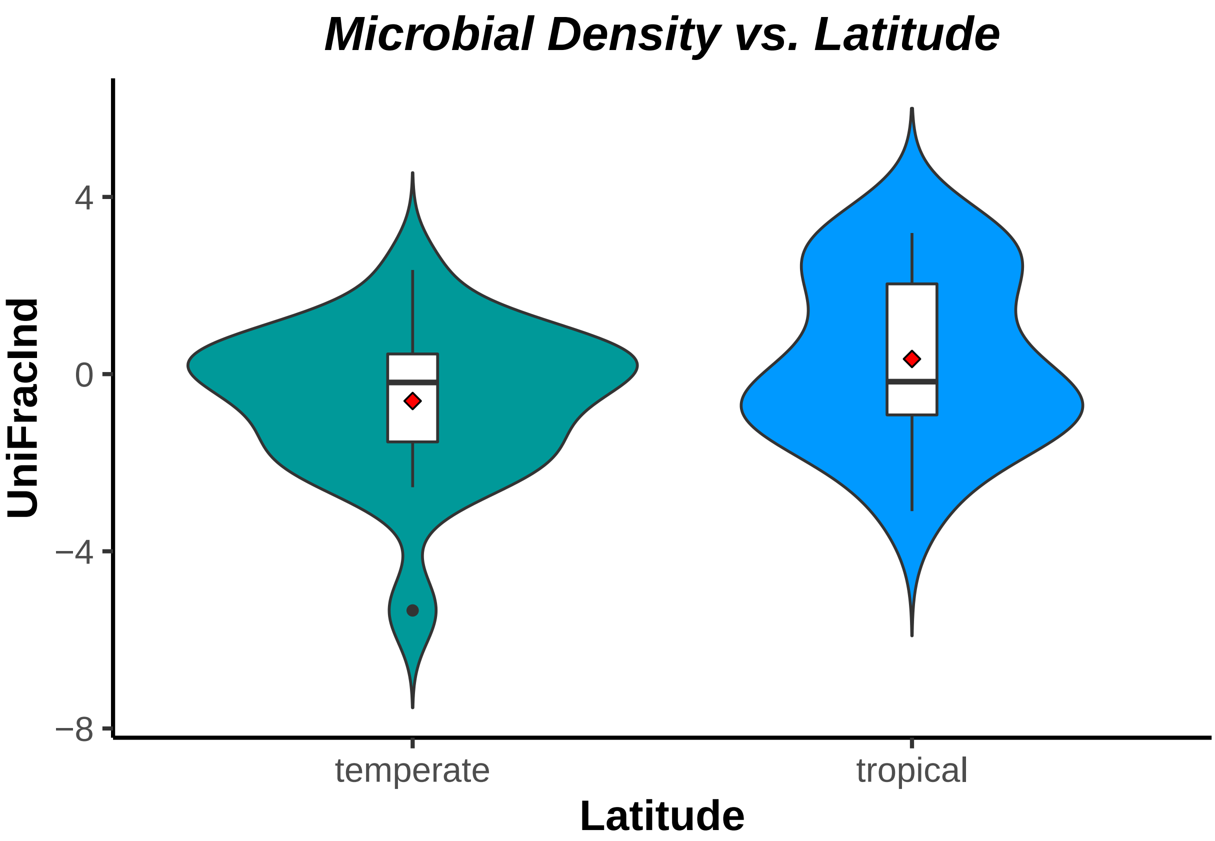
```
# Defining variables to simplify dataframe references:
diversity.col <-marine$UniFracInd
latitude.col <- marine$latitude
season.col <- marine$season
```

1. How does microbial diversity change with latitude?

```r
library(ggplot2)
# set up plot-specific elements
# ggplot object plotting UniFracInd against latitude
marine.latitude.plot <- ggplot(marine,
                             aes(factor(latitude.col),
                                 diversity.col,
                                 fill = latitude.col))
# specific plot title and axes labels
marine.latitude.titles <- labs(title = "Microbial Density vs. Latitude",
                               x = "Latitude",
                               y = "UniFracInd")
# change fill from the default colours
latitude.fill <- scale_fill_manual(values=c("#009999", "#0099FF"))

# set up re-usable elements to edit violin plots
# create a violin plot with count on the y axis and including outliers
violin.plot <- geom_violin(scale = "count",
                           trim = FALSE)
## add sample points
#geom_jitter(height = 0, width = 0.2) +
# add a boxplot
add.boxplot <- geom_boxplot(width = 0.1,
                            fill = "white")
# add the mean of values
add.mean <- stat_summary(fun.y = mean,
                         geom = "point",
                         shape = 23,
                         size = 2,
                         fill = "red")
# change the plot theme to classic (no grey plot area or gridlines)
set.plot.theme <- theme_classic(base_size = 16)
# set variable for the style of plot titles;
# hjust is the horizontal position of the title
title.style <- element_text(face = "bold.italic",
                            size = 18,
                            hjust = .5)
# set variable for the style of plot axes
axes.labels.style <- element_text(face = "bold")
# apply title and axes format to the plot theme and remove legend
title.axes.theme <- theme(plot.title = title.style,
                          axis.title = axes.labels.style,
                          legend.position="none")

# bring plot and formatting together
marine.latitude.plot + marine.latitude.titles + violin.plot + add.boxplot +
  add.mean + latitude.fill + set.plot.theme + title.axes.theme
```



```
mean(diversity.col[latitude.col=="temperate"])
```

```
## [1] -0.6068768
```

```
sd(diversity.col[latitude.col=="temperate"])
```

```
## [1] 1.685375
```

```
mean(diversity.col[latitude.col=="tropical"])
```

```
## [1] 0.3421579
```

```
sd(diversity.col[latitude.col=="tropical"])
```

```
## [1] 1.898183
```

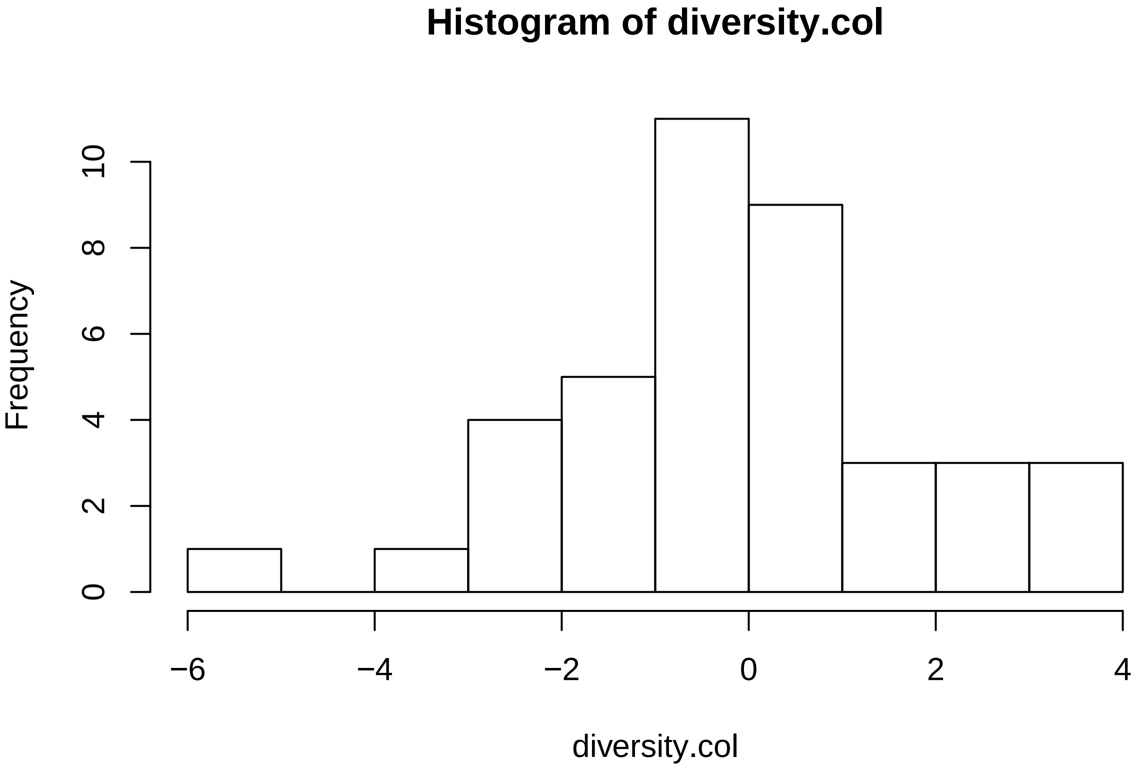
The violin plot shows that although the median of UniFracInd is identical between samples taken in equatorial waters and those obtained in tropical waters, the distribution of sample microbiological diversity is slightly different in each type of environment. While in temperate waters, the samples with a UniFracInd above the median tend to aggregate close to the median value, with the remaining values more spread below the mid point's value, in tropical waters this trend is reversed, with the samples with higher diversity than the median value more spread and those with lower than median diversity showing UniFracInd values closer to the median value. This is illustrated by the fact that the means of each type of sample (red losanges) difer by a larger margin than the medians, with samples collected in temperate waters showing a lower mean of microbial diversity index (UniFracInd) than those from tropical waters. In order to test whether these differences are significant we could perform a t-test. The null hypothesis would be: H0: There is no difference in mircobial diversity between samples collected in temperate and tropical regions. The alternative hypothesis would be: Ha: There is a difference in mircobial diversity between samples collected in temperate and tropical regions. Significance level, a=0.05 t-test assumptions:

1. The data is measured in a continuous scale - UniFracInd is a continuous variable

2. The samples collected are random and representative - 20l of water from different locations
3. The data is normally distributed
4. Large size sample
5. Homogeneous variance

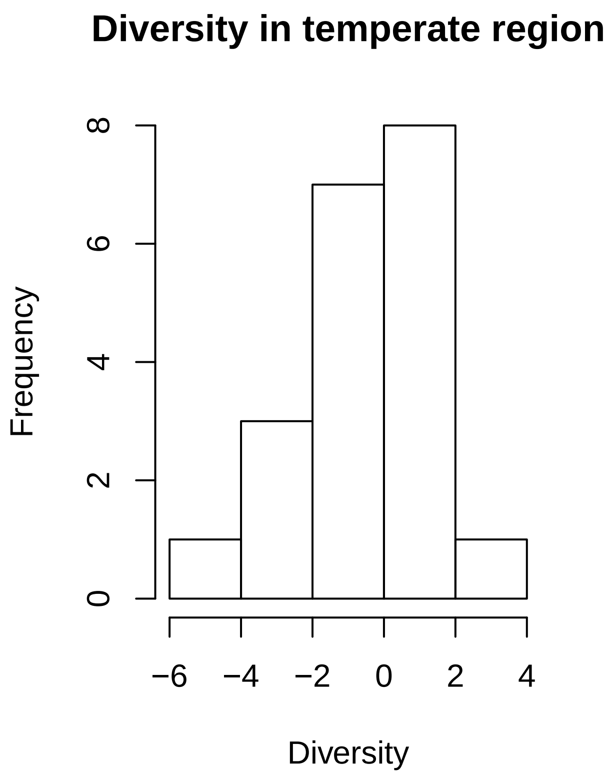
From the histograms below, the total data is close to the normal distribution:

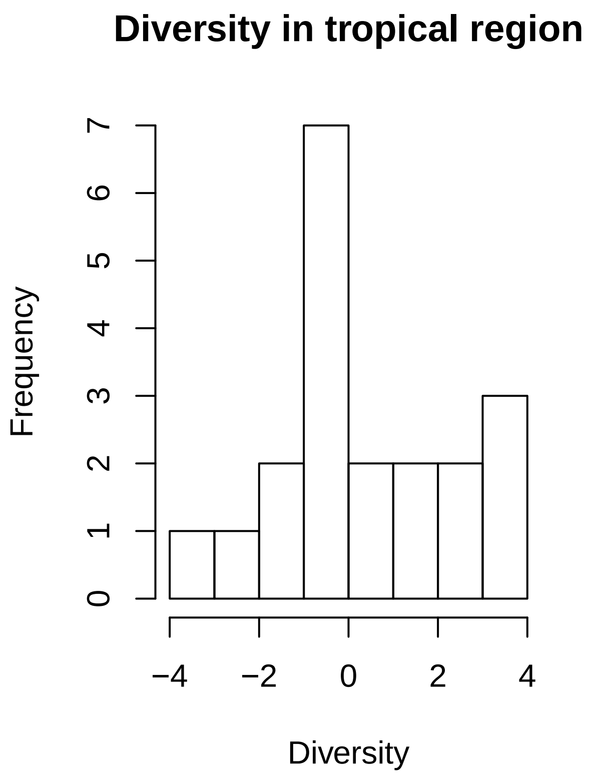
```r
hist(diversity.col)
```



However, the subsets of data corresponding to the temperate and tropical samples appear farther from the normal distribution (assumption 3).

```r
par(mfrow = c(1,2))
hist(diversity.col[latitude.col == "temperate"],
     breaks = 5,
     main = "Diversity in temperate region",
     xlab = "Diversity")
hist(diversity.col[latitude.col == "tropical"],
     breaks = 5,
     main = "Diversity in tropical region",
     xlab = "Diversity")
```





We can perform a normality test on these:

```
shapiro.test(diversity.col[latitude.col=="temperate"])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  diversity.col[latitude.col == "temperate"]
## W = 0.93387, p-value = 0.1833
```

```
shapiro.test(diversity.col[latitude.col=="tropical"])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  diversity.col[latitude.col == "tropical"]
## W = 0.93339, p-value = 0.1794
```

The p-values are above the recommended for this test (<0.1; ethz R manual, which means we cannot exclude the null hypothesis for the Shapiro-Wilk test that the sample is normally distributed, so we can proceed to assess whether the remaining assumptions of the t-test are met.

We can determine the number of samples in each group:

```
length(latitude.col[latitude.col=="temperate"])
```

```
## [1] 20
```

```
length(latitude.col[latitude.col=="tropical"])
```

```
## [1] 20
```

There are 20 samples in each group, which should be sufficient if the other assumptions are met.

Homoscedasticity of variance can be tested with Levene's test, which is a variance homogeneity test that can be used when there is a slight departure from normality R cookbook:

```
library(car)
```

```
## Loading required package: carData
```

```
leveneTest(diversity.col ~ latitude.col)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  1  0.7528  0.391
##       38
```

Given that the null hypothesis for the Levene's homoscedasticity test is that the variances are homogeneous, the p-value of 0.391 is not sufficient to reject homoscedasticity.

Given that we can assume all t-tests assumptions are met, we can then perform a t-test as below:

```
t.test(diversity.col[latitude.col=="temperate"], diversity.col[latitude.col=="tropical"])
```

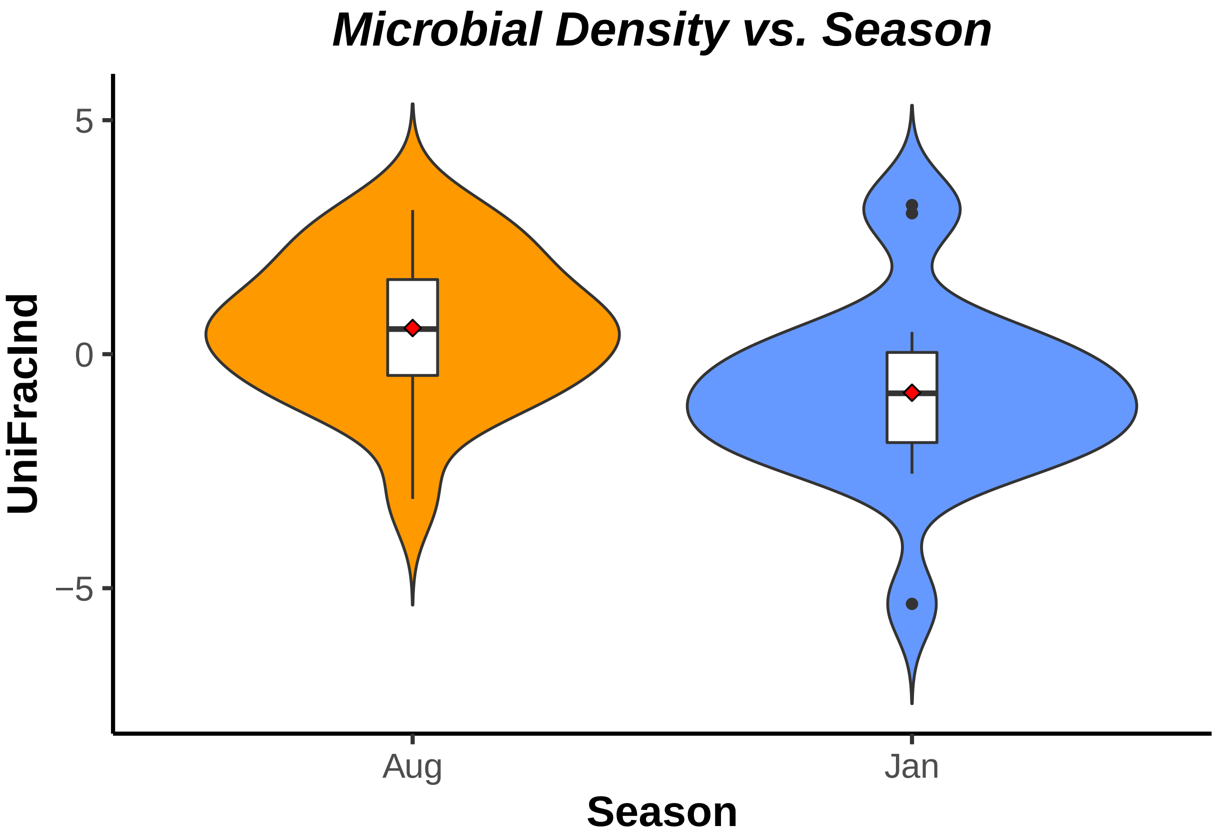
```
## 
##  Welch Two Sample t-test
## 
## data:  diversity.col[latitude.col == "temperate"] and diversity.col[latitude.col == "tropical"]
## t = -1.672, df = 37.475, p-value = 0.1029
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -2.0986268  0.2005576
## sample estimates:
##  mean of x  mean of y 
## -0.6068768  0.3421579
```

The p-value of 0.1209 would indicate that the likelihood of observing the current samples if H0 is true is above the set significance level of 0.05 and hence we cannot reject the null hypothesis. From these analyses, I would conclude there is no statistically significant difference in microbial diversity at different latitudes at the set level of significance (a=0.05).

2. How does microbial diversity change with time of year?

```
# set up plot-specific elements
# ggplot object plotting UniFracInd against season
marine.season.plot <- ggplot(marine,
                             aes(factor(season.col),
                                 diversity.col,
                                 fill = season.col))
# specific plot title and axes labels
marine.season.titles <- labs(title = "Microbial Density vs. Season",
                             x = "Season",
                             y = "UniFracInd")
# change fill from the default colours
season.fill <- scale_fill_manual(values=c("#FF9900", "#6699FF"))

# bring plot and formatting together
marine.season.plot + marine.season.titles + violin.plot + add.boxplot +
  add.mean + season.fill + set.plot.theme + title.axes.theme
```



```
mean(diversity.col[season.col=="Aug"])
```

```
## [1] 0.5588926
```

```
sd(diversity.col[season.col=="Aug"])
```

```
## [1] 1.551889
```

```
mean(diversity.col[season.col=="Jan"])
```

```
## [1] -0.8236116
```

```
sd(diversity.col[season.col=="Jan"])
```

```
## [1] 1.871713
```

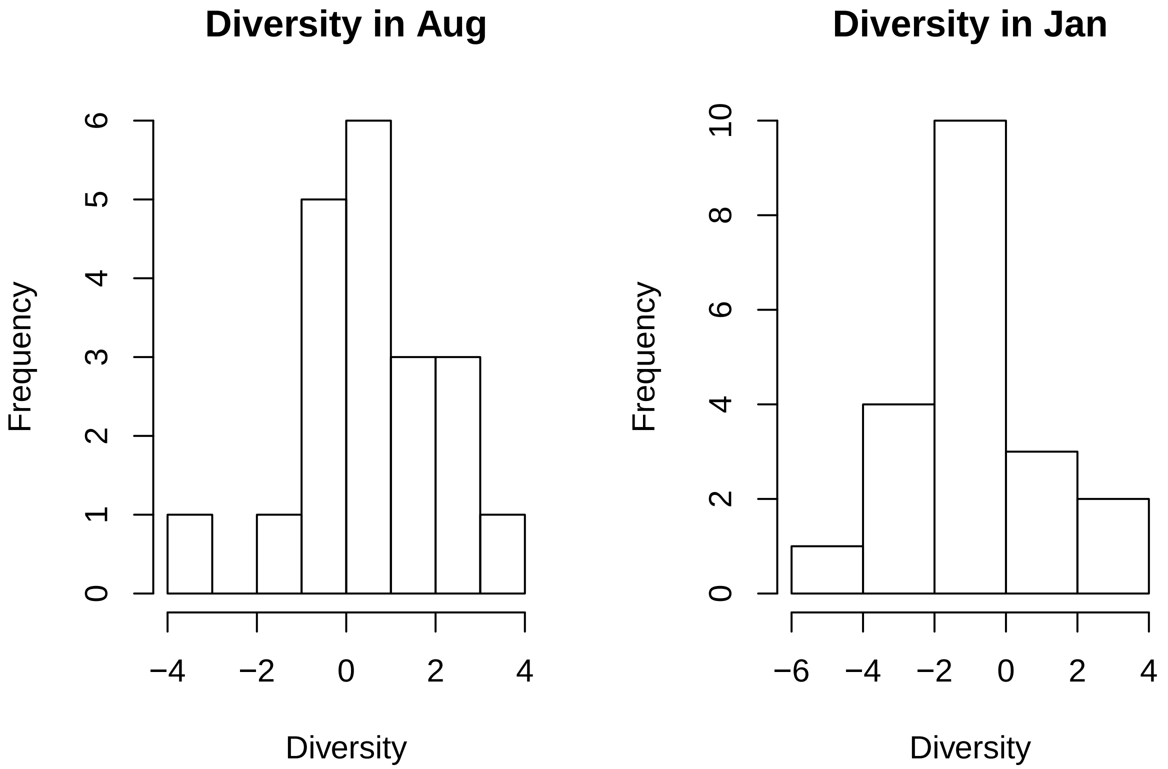
Higher microbial diversity is observed in the August cruise than in January. The values for UniFracInd are more normally distributed around the median value than those measured by latitude. As such, median and mean values coincide, being higher in August. The spread of UniFracInd values measured in January is smaller than that seen in August.

We can perform a t-test to assess if these differences are significant. H0: There is no difference in mircobial diversity between samples collected in Aug and Jan seasons. Ha: There is a difference in mircobial diversity between samples collected in Aug and Jan seasons. Significance level, a=0.05

As for the latitude, we can assume that conditions 1 and 2 (continuous variable being measured and random representative samples) for a t-test are verified. Normality:

```
par(mfrow=c(1,2))
hist(diversity.col[season.col=="Aug"],
     main="Diversity in Aug",
     xlab="Diversity")
```

```
hist(diversity.col[season.col=="Jan"],
     main="Diversity in Jan",
     xlab="Diversity")
```



```
shapiro.test(diversity.col[season.col=="Aug"])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  diversity.col[season.col == "Aug"]
## W = 0.97725, p-value = 0.8939
```

```
shapiro.test(diversity.col[season.col=="Jan"])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  diversity.col[season.col == "Jan"]
## W = 0.92805, p-value = 0.1416
```

These two groups of observations appear to be normally distributed (p>0.1).

Large size sample:

```
length(season.col[season.col=="Aug"])
```

```
## [1] 20
```

```
length(season.col[season.col=="Jan"])
```

```
## [1] 20
```

Again, there are are 20 samples in each group, which should be sufficient for a t-test if the remaining assumptions are met.

Homoscedasticity:

```
leveneTest(diversity.col ~ season.col)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  1  0.0836 0.7741
##       38
```

The probability of observing the current variances in the two groups if they are homoscedastic is very high (>0.7) by Levene's test, hence we can assume there is homogeinity in the variances.

We can thus perform a t-test:

```
t.test(diversity.col[season.col=="Aug"], diversity.col[season.col=="Jan"])
```

```
##
##  Welch Two Sample t-test
##
## data:  diversity.col[season.col == "Aug"] and diversity.col[season.col == "Jan"]
## t = 2.5429, df = 36.74, p-value = 0.01534
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.2806488 2.4843596
## sample estimates:
##  mean of x  mean of y
##  0.5588926 -0.8236116
```

This time, we obtain a p-value < 0.02, below the sgnificance level set (a=0.05). In this context, we would reject the null hypothesis that there is no difference in microbial diversity in each season, and accept the alternative hypothesis that states that there is a statistically significant difference between microbial diversity in Jan and Aug. From the violin plot, we can see that there is higher diversity in Aug than in Jan.

3. Is there an interaction between the season and location?

To statistically assess whether there is an interaction of season and latitude, we can fit a linear model which takes into account interactions of season and latitude and then do an analysis of variance to determine which factors better explain changes in diversity.

```
# Correlation of diversity with an interaction of latitude and season
corr.seas.lat <- lm(diversity.col ~ season.col * latitude.col)
anova(corr.seas.lat)
```

```
## Analysis of Variance Table
##
## Response: diversity.col
##                         Df Sum Sq Mean Sq F value  Pr(>F)
## season.col               1 19.113 19.1132  7.0867 0.01153 *
## latitude.col             1  9.007  9.0067  3.3395 0.07594 .
## season.col:latitude.col  1  6.222  6.2218  2.3069 0.13753
## Residuals               36 97.093  2.6970
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the ANOVA test of the model fitted taking into acount the interaction of latitude and season on the measure microbial diversity we conclude that, as was apparent from the simpler t-tests, season has a larger impact on microbial diversity than latitude. The term acounting for interaction between season and latitude has a p-valua > 0.1, which is above the 0.05 significance level. This suggests that the main impact on microbial diversity comes from the season when the samples are collected and that we cannot exclude the null hypothesis that there is no interaction between season and location.

Dataset 2: Pairwise nucleotide substitutions and RNA expression levels

- luciferase
- RNA-seq project
- over 900 coding loci from 208 species of plants seq'd
- found 30 genes with homology to luciferase
- measured exprssion levels in each species
- phylogenetic tree
- computed pairwise genetic distances each species' gene and that of closely-related *Arabidopsis thaliana*

Data import

No issues were observed in Excel.

```
luciferase <- read.table("part_2_student_1932.tdf",
                         sep = "\t",
                         header = TRUE)
head(luciferase) # quick check of correct data import
```

```
##   expression_fold distance
## 1      0.09552236 2.418824
## 2      1.88550850 3.664110
## 3      1.72791380 2.272160
## 4      1.34275436 3.938938
## 5      0.20025704 4.698375
## 6      1.71045880 4.503878
```

```
names(luciferase) # check names of columns for dataset
```

```
## [1] "expression_fold" "distance"
```

```
str(luciferase) # check that factors and variables were assigned correctly
```

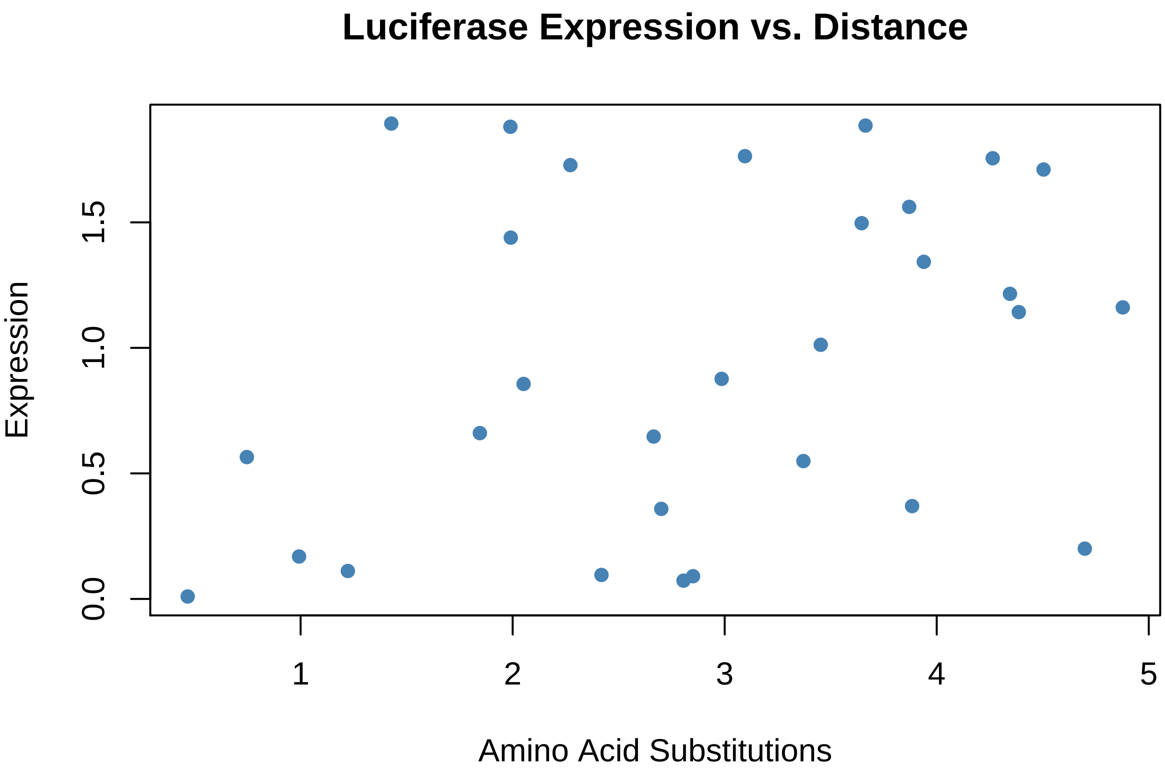
```
## 'data.frame':    30 obs. of  2 variables:
##  $ expression_fold: num  0.0955 1.8855 1.7279 1.3428 0.2003 ...
##  $ distance       : num  2.42 3.66 2.27 3.94 4.7 ...
```

```
# simplify dataframe references
luciferase.expression <- luciferase$expression_fold
aa.substitutions <- luciferase$distance
```

1. How does the putative 'luciferase' homologue expression change with genetic distance (amino acid substitutions)?

We can start by visualising these data using a scatter plot:

```r
plot(aa.substitutions, luciferase.expression,
     main="Luciferase Expression vs. Distance",
     xlab="Amino Acid Substitutions",
     ylab="Expression",
     pch=16,
     col="steelblue")
```



There appears to be no significant correlation of luciferase expression with amino acid substitutions.

```r
expr.dist.model <- lm(luciferase.expression ~ aa.substitutions)
expr.dist.model
```

```
##
## Call:
## lm(formula = luciferase.expression ~ aa.substitutions)
##
## Coefficients:
##      (Intercept)  aa.substitutions
##           0.4529            0.1719
```

If we try to fit a regression to these data, we obtain the following relationship: luciferase.expression = 0.4529 + 0.1719 * aa.substitutions + errors

```r
summary(expr.dist.model)
```

```
##
## Call:
## lm(formula = luciferase.expression ~ aa.substitutions)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.06047 -0.51345 -0.04952  0.47323  1.19504
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       0.45291    0.30141   1.503   0.1441
## aa.substitutions  0.17193    0.09551   1.800   0.0826 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6333 on 28 degrees of freedom
## Multiple R-squared:  0.1037, Adjusted R-squared:  0.07172
## F-statistic: 3.241 on 1 and 28 DF,  p-value: 0.08261
```

When we analyse the statistical parameters associated with this model, it is clear that the p-value is above the significance level of 0.05 (P=0.08261). The fit of the curve is also low, with an R-squared of 0.1037.

There is no obvious curve that would fit the data better than the linear regression. I have attempted to transform the data to represent expression in terms of log and log2, but it seems likely that the values presented have already been transformed in this way. I have also attempted a power of two of the distance. None of these transformations yield more promising plots than the linear regression model. As such, I would conclude that there is no statistically significant correlation between gene expression and distance in the dataset analysed.

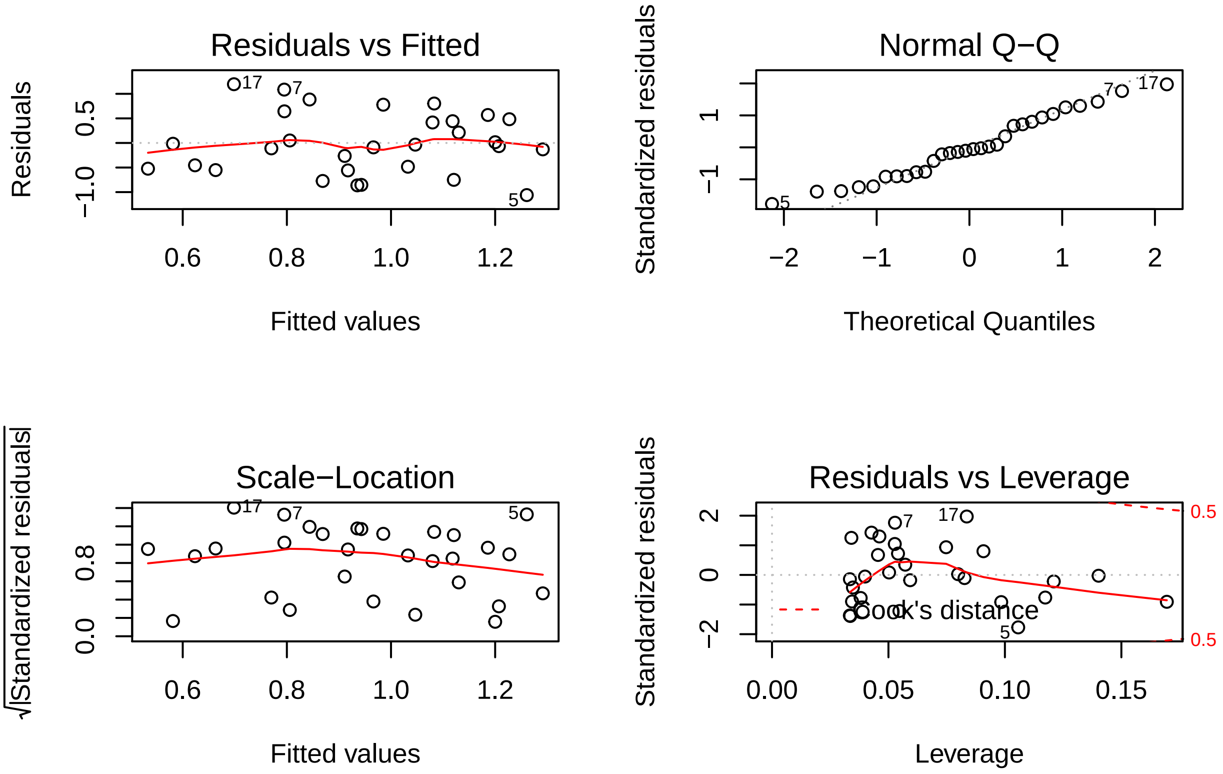
2. Comment on whether the model assumptions are valid.

The assumptions behind a regression are that:

- the errors are normally distributed
- the residuals are homoscedastic
- there is a linear structure in the data

As seen above, there is no statistically significant linear structure in the data. However, we can look at whether the remaining assumptions are met using the model diagnosis plots.

```
par(mfrow=c(2,2))
plot(expr.dist.model)
```



There is no obvious non-linear relationship between the residuals and the predictor (top left). The Normal Q-Q plot indicates that the residuals are normally distributed. The residuals are also randomly distributed around an approximately horizontal line in the Scale-Location plot (bottom left), which is consistent with homoscedasticity. On the final plot, Residuals vs Leverage (bottom right), there are no obvious outliers which disproportionally influence the model. All these observations are consistent with an acceptable model, if not for the low significance level of the regression obtained and the lack of a linear structure to the data.

3. Assuming your model is *statistically* valid, can you guess what effect is responsible for the relationship you've found?

If we were assume the model was statistically valid, i.e., if we had observed a linear structure in our data and obtained a higher R-squared value, this model would suggest that there was a positive correlation between luciferase expression and the distance to the luciferase putative homologue in *Arabidopsis thaliana*. Given this is an unexpected finding, I would first double-check the data (e.g., contamination in the RNA seq run or samples). If no issues were found in the data, the positive correlation would suggest that this putative gene produces a limited amount of luciferase, so species in which the homologue is further away from the model organism's one are the ones where the gene may be upregulated.

Dataset 3: HIV viral load and within-patient population dynamics

- HIV may evade drug therapy and persist in immunoprotected tissues, especially CNS (no CD4+ T-cells)
- HIV+ patient consented to viral load sampling over 40 weeks in a year
- sample taken from the brain or spinal cord
- Chip assay to determine viral 'load', expressed as log10(number of viral particles per ml)
- Average Shannon population diversity
- mean pairwise genetic distance from individual viruses present in each weekly sample to a reference sequence was assessed using single-copy PCR amplification of the viral Env gene

- relationship between population size, diversity, evolutionary distance and tissue

Data import

The data was imported following check on Excel.

```
# import data from tab-delimitted file
hiv <- read.table("part_3_student_1932.tdf",
                  sep = "\t",
                  header = TRUE)
tail(hiv)  # quick check that data was correctly imported
```

```
##         VLoad CD4     tissue score_shannon score_distance
## 35 -4.4746426  lo spinalCord      2.919199      0.5845748
## 36 -0.3372384  lo spinalCord      2.312107      1.4827633
## 37 -3.5201091  lo spinalCord      4.487376      0.3197251
## 38 -1.3449078  lo spinalCord      1.893682      1.2252042
## 39 -7.3452547  lo spinalCord      4.218892      1.9217303
## 40 -1.6384461  lo spinalCord      2.231115      0.6860380
```

```
names(hiv)  # dataframe variable names
```

```
## [1] "VLoad"          "CD4"            "tissue"         "score_shannon"
## [5] "score_distance"
```

```
str(hiv)  # dataframe structure
```

```
## 'data.frame':    40 obs. of  5 variables:
##  $ VLoad         : num  1.025 -0.428 -1.286 -3.191 -2.332 ...
##  $ CD4           : Factor w/ 2 levels "hi","lo": 1 1 1 1 1 1 1 1 1 1 ...
##  $ tissue        : Factor w/ 2 levels "brain","spinalCord": 1 1 1 1 1 1 1 1 1 1 ...
##  $ score_shannon : num  3.49 4.49 1.44 2.09 3.44 ...
##  $ score_distance: num  0.848 1.475 0.79 1.819 0.809 ...
```

The data was imported correctly.

```
# set up dataframe call vaariables
viral.load <- hiv$VLoad
cd4.level <- hiv$CD4
site <- hiv$tissue
shannon <- hiv$score_shannon
distance <- hiv$score_distance
```

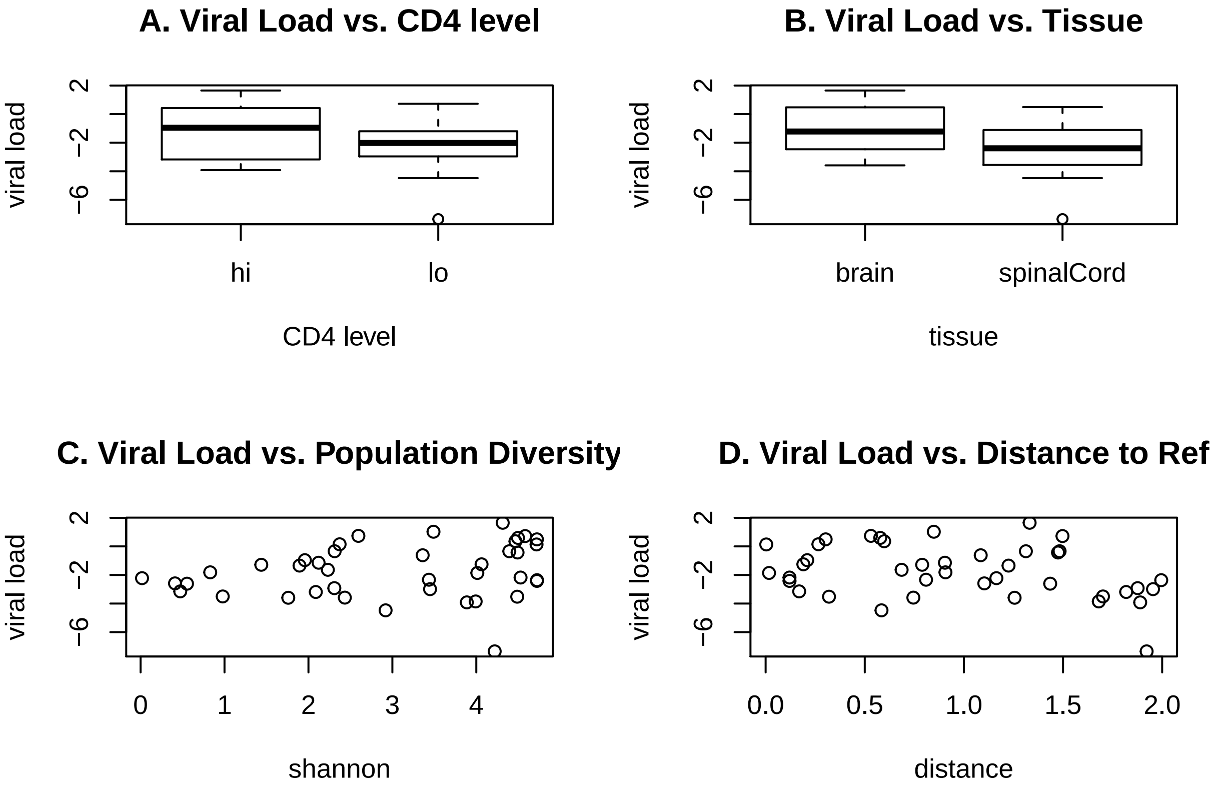
Fit a model explaining viral load in terms of the other variables, using any combination of variables you see fit, and any model selection procedure

- The rows in the data table are in chronological order
- Likelihood of CD4 cell penetration into CNS may vary between patients
- viral load (log10)
- CD4: hi or lo
- 2 tissues: spinal cord and brain
- Shannon pop diversity

- distance to ref env seq

We can start by performing a quick visualisation of the data to get a sense of the most relevant variables.

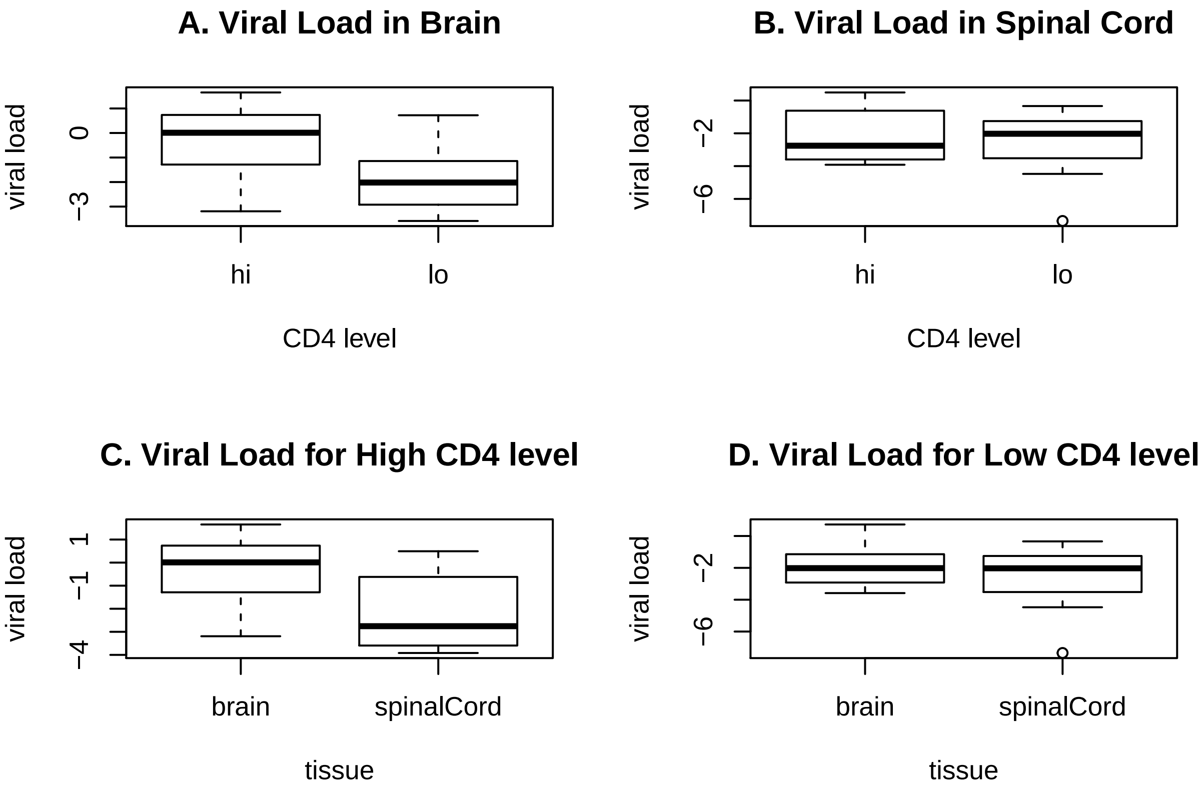
```
par(mfrow=c(2, 2))
plot(cd4.level, viral.load,
     main = "A. Viral Load vs. CD4 level",
     xlab = "CD4 level",
     ylab = "viral load")
plot(site, viral.load,
     main = "B. Viral Load vs. Tissue",
     xlab = "tissue",
     ylab = "viral load")
plot(shannon, viral.load,
     main = "C. Viral Load vs. Population Diversity",
     xlab = "shannon",
     ylab = "viral load")
plot(distance, viral.load,
     main = "D. Viral Load vs. Distance to Ref",
     xlab = "distance",
     ylab = "viral load")
```



The viral load appears to be higher when the level of CD4-positive cells is higher. When the samples are collected from the spinal cord, the values for viral load have a lower median than for brain biopsies. I cannot detect a clear relationship between viral load and either the Shannon measurement of population diversity or the distance to a reference sequence.

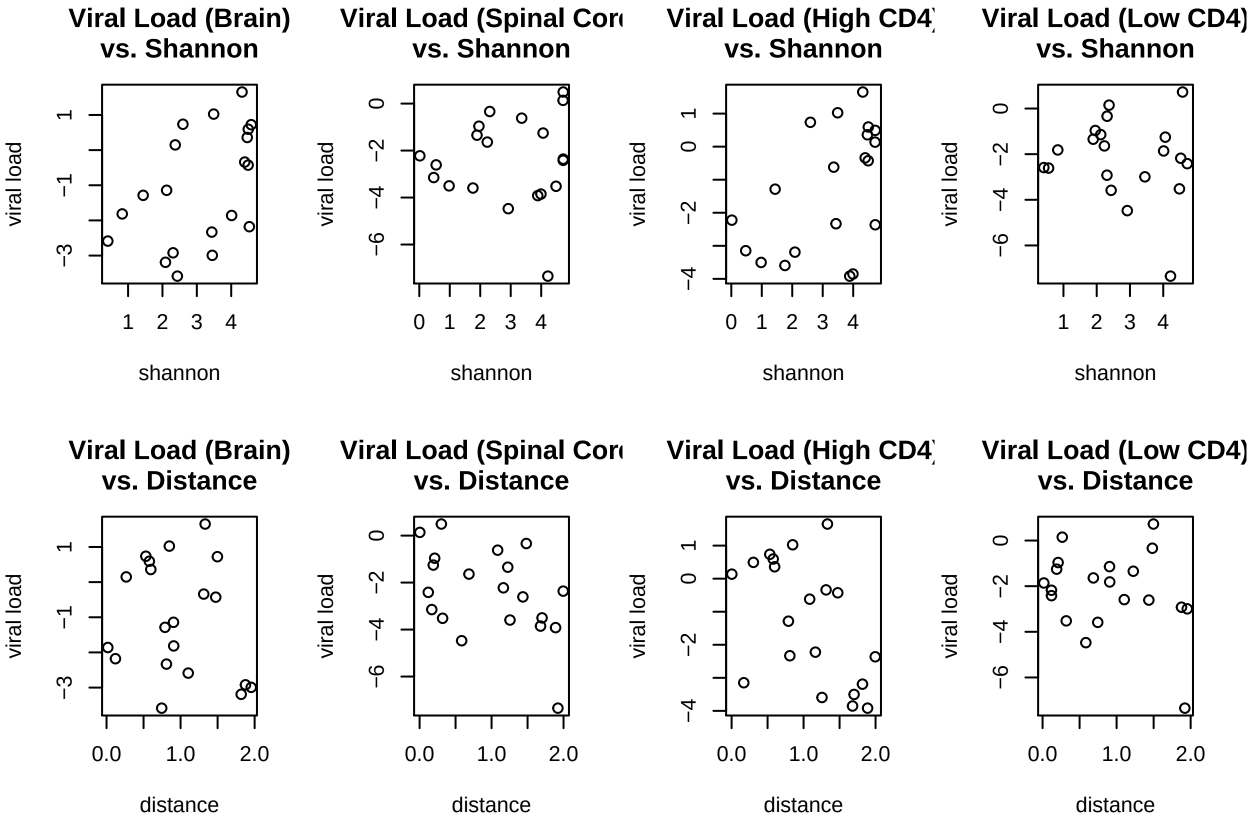
We can look at the combined effects of CD4 levels and tissue on viral load.

```
par(mfrow=c(2, 2))
plot(cd4.level[site=="brain"], viral.load[site=="brain"],
     main = "A. Viral Load in Brain",
     xlab = "CD4 level",
     ylab = "viral load")
plot(cd4.level[site=="spinalCord"], viral.load[site=="spinalCord"],
     main = "B. Viral Load in Spinal Cord",
     xlab = "CD4 level",
     ylab = "viral load")
plot(site[cd4.level=="hi"], viral.load[cd4.level=="hi"],
     main = "C. Viral Load for High CD4 level",
     xlab = "tissue",
     ylab = "viral load")
plot(site[cd4.level=="lo"], viral.load[cd4.level=="lo"],
     main = "D. Viral Load for Low CD4 level",
     xlab = "tissue",
     ylab = "viral load")
```

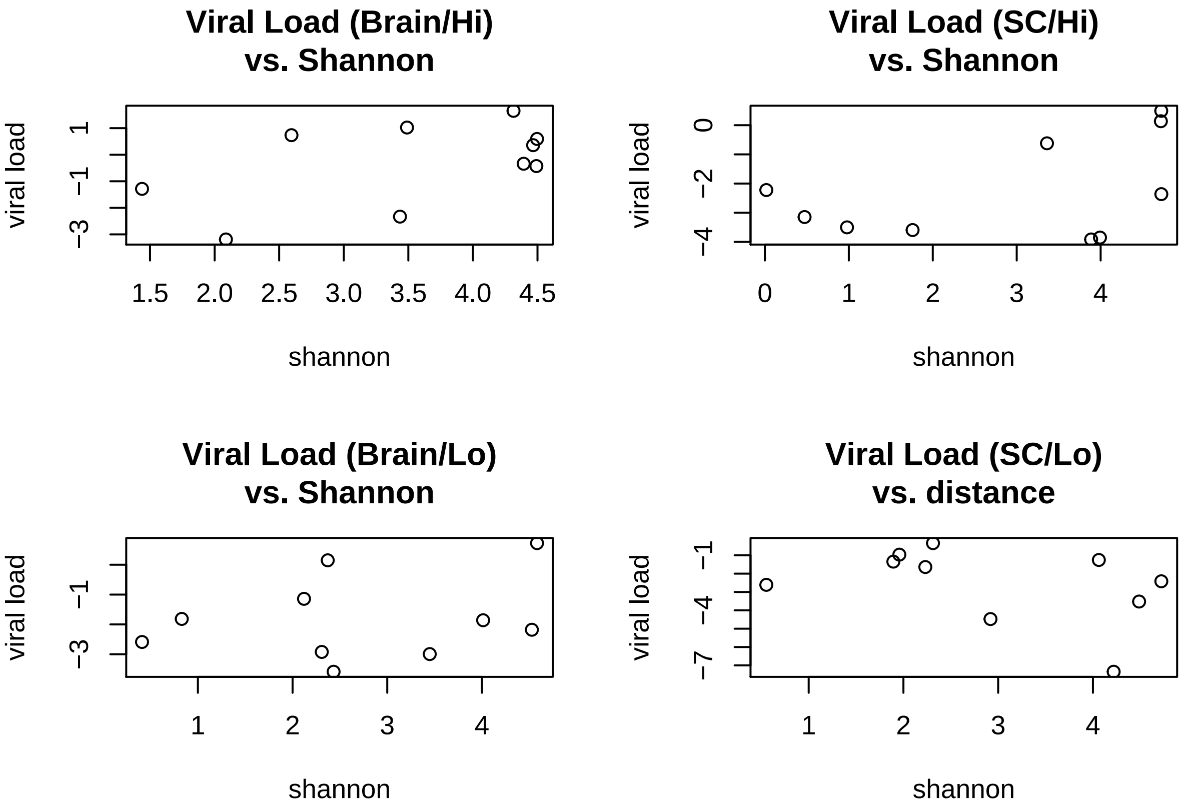


```
par(mfrow=c(2, 4))
plot(shannon[site=="brain"], viral.load[site=="brain"],
     main = "Viral Load (Brain)\nvs. Shannon",
     xlab = "shannon",
     ylab = "viral load")
plot(shannon[site=="spinalCord"], viral.load[site=="spinalCord"],
     main = "Viral Load (Spinal Cord)\nvs. Shannon",
     xlab = "shannon",
     ylab = "viral load")
plot(shannon[cd4.level=="hi"], viral.load[cd4.level=="hi"],
     main = "Viral Load (High CD4)\nvs. Shannon",
     xlab = "shannon",
     ylab = "viral load")
```

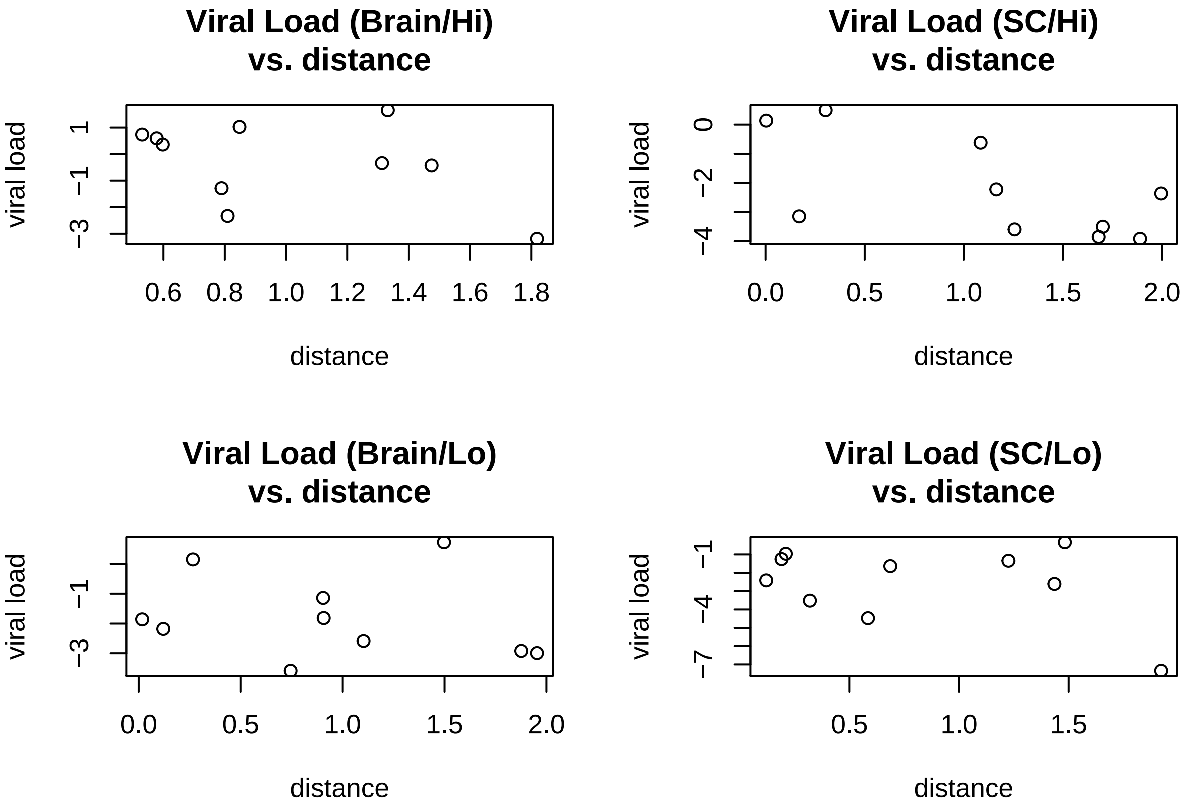
```r
plot(shannon[cd4.level=="lo"], viral.load[cd4.level=="lo"],
     main = "Viral Load (Low CD4)\nvs. Shannon",
     xlab = "shannon",
     ylab = "viral load")
plot(distance[site=="brain"], viral.load[site=="brain"],
     main = "Viral Load (Brain)\nvs. Distance",
     xlab = "distance",
     ylab = "viral load")
plot(distance[site=="spinalCord"], viral.load[site=="spinalCord"],
     main = "Viral Load (Spinal Cord)\nvs. Distance",
     xlab = "distance",
     ylab = "viral load")
plot(distance[cd4.level=="hi"], viral.load[cd4.level=="hi"],
     main = "Viral Load (High CD4)\nvs. Distance",
     xlab = "distance",
     ylab = "viral load")
plot(distance[cd4.level=="lo"], viral.load[cd4.level=="lo"],
     main = "Viral Load (Low CD4)\nvs. Distance",
     xlab = "distance",
     ylab = "viral load")
```



```r
par(mfrow=c(2, 2))
plot(shannon[site=="brain"][cd4.level=="hi"],
     viral.load[site=="brain"][cd4.level=="hi"],
     main = "Viral Load (Brain/Hi)\nvs. Shannon",
     xlab = "shannon",
     ylab = "viral load")
plot(shannon[site=="spinalCord"][cd4.level=="hi"],
     viral.load[site=="spinalCord"][cd4.level=="hi"],
     main = "Viral Load (SC/Hi)\nvs. Shannon",
     xlab = "shannon",
     ylab = "viral load")
plot(shannon[site=="brain"][cd4.level=="lo"],
     viral.load[site=="brain"][cd4.level=="lo"],
     main = "Viral Load (Brain/Lo)\nvs. Shannon",
     xlab = "shannon",
     ylab = "viral load")
plot(shannon[site=="spinalCord"][cd4.level=="lo"],
     viral.load[site=="spinalCord"][cd4.level=="lo"],
     main = "Viral Load (SC/Lo)\nvs. distance",
     xlab = "shannon",
     ylab = "viral load")
```



```
par(mfrow=c(2, 2))
plot(distance[site=="brain"][cd4.level=="hi"],
     viral.load[site=="brain"][cd4.level=="hi"],
     main = "Viral Load (Brain/Hi)\nvs. distance",
     xlab = "distance",
     ylab = "viral load")
plot(distance[site=="spinalCord"][cd4.level=="hi"],
     viral.load[site=="spinalCord"][cd4.level=="hi"],
     main = "Viral Load (SC/Hi)\nvs. distance",
     xlab = "distance",
     ylab = "viral load")
plot(distance[site=="brain"][cd4.level=="lo"],
     viral.load[site=="brain"][cd4.level=="lo"],
     main = "Viral Load (Brain/Lo)\nvs. distance",
     xlab = "distance",
     ylab = "viral load")
plot(distance[site=="spinalCord"][cd4.level=="lo"],
     viral.load[site=="spinalCord"][cd4.level=="lo"],
     main = "Viral Load (SC/Lo)\nvs. distance",
     xlab = "distance",
     ylab = "viral load")
```



The relationship between population diversity or distance to the reference sequence to the viral load determined appears to depend on the site of sampling or the level of CD4-positive cells. We could then fit a model that takes into consideration the interaction between tissue and CD4 level and then the remaining, non-interacting, variables of Shannon and distance. Our start model could be the following:

```
start.model <- lm(viral.load ~ site + cd4.level + shannon + distance + site:cd4.level)
summary(start.model)
```

```
##
## Call:
## lm(formula = viral.load ~ site + cd4.level + shannon + distance +
##     site:cd4.level)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.7784 -1.0768  0.2047  0.9586  3.0342
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  0.2473     0.9386   0.264   0.7937
## sitespinalCord              -1.7270     0.7150  -2.416   0.0212 *
## cd4.levello                 -1.4608     0.7208  -2.027   0.0506 .
## shannon                      0.1381     0.1789   0.772   0.4454
## distance                    -1.0449     0.4144  -2.522   0.0165 *
## sitespinalCord:cd4.levello   0.7990     1.0119   0.790   0.4352
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.575 on 34 degrees of freedom
## Multiple R-squared:  0.3662, Adjusted R-squared:  0.273
## F-statistic:  3.93 on 5 and 34 DF,  p-value: 0.006417
```

We can attempt to improve the fitting with the step function:

```r
my.model = step(start.model,
                scope=c(lower=~site,
                        upper=~site * cd4.level * shannon * distance),
                direction="both")
```

```
## Start:  AIC=41.86
## viral.load ~ site + cd4.level + shannon + distance + site:cd4.level
##
##                        Df Sum of Sq     RSS    AIC
## + cd4.level:shannon   1   11.9988  72.384 37.724
## - shannon             1    1.4794  85.863 40.555
## - site:cd4.level      1    1.5473  85.931 40.586
## <none>                             84.383 41.860
## + site:shannon        1    3.0376  81.346 42.393
## + cd4.level:distance  1    2.4304  81.953 42.691
## + site:distance       1    1.5629  82.820 43.112
## + shannon:distance    1    0.1525  84.231 43.787
## - distance            1   15.7805 100.164 46.717
##
## Step:  AIC=37.72
## viral.load ~ site + cd4.level + shannon + distance + site:cd4.level +
##     cd4.level:shannon
##
##                        Df Sum of Sq    RSS    AIC
## + site:shannon        1    7.5995 64.785 35.288
## - site:cd4.level      1    0.9779 73.362 36.261
## <none>                            72.384 37.724
## + site:distance       1    1.9333 70.451 38.642
## + cd4.level:distance  1    1.3281 71.056 38.984
## + shannon:distance    1    0.2182 72.166 39.604
## - cd4.level:shannon   1   11.9988 84.383 41.860
## - distance            1   19.6306 92.015 45.323
##
## Step:  AIC=35.29
## viral.load ~ site + cd4.level + shannon + distance + site:cd4.level +
##     cd4.level:shannon + site:shannon
##
##                            Df Sum of Sq    RSS    AIC
## - site:cd4.level          1    0.1866 64.972 33.403
## <none>                                64.785 35.288
## + site:distance           1    3.0711 61.714 35.345
## + site:cd4.level:shannon  1    2.1211 62.664 35.956
## + cd4.level:distance      1    0.9548 63.830 36.694
## + shannon:distance        1    0.2746 64.510 37.118
## - site:shannon            1    7.5995 72.384 37.724
## - cd4.level:shannon       1   16.5607 81.346 42.393
## - distance                1   20.7887 85.574 44.420
##
## Step:  AIC=33.4
## viral.load ~ site + cd4.level + shannon + distance + cd4.level:shannon +
##     site:shannon
##
##                      Df Sum of Sq    RSS    AIC
## + site:distance       1    3.2364 61.735 33.359
## <none>                            64.972 33.403
## + cd4.level:distance  1    0.9576 64.014 34.809
## + shannon:distance    1    0.2849 64.687 35.227
## + site:cd4.level      1    0.1866 64.785 35.288
## - site:shannon        1    8.3909 73.362 36.261
## - cd4.level:shannon   1   17.2360 82.207 40.815
## - distance            1   21.3647 86.336 42.775
##
## Step:  AIC=33.36
## viral.load ~ site + cd4.level + shannon + distance + cd4.level:shannon +
##     site:shannon + site:distance
##
##                      Df Sum of Sq    RSS    AIC
## <none>                            61.735 33.359
## - site:distance       1    3.2364 64.972 33.403
## + cd4.level:distance  1    0.2362 61.499 35.206
## + shannon:distance    1    0.2090 61.526 35.223
## + site:cd4.level      1    0.0213 61.714 35.345
## - site:shannon        1    9.3375 71.073 36.993
## - cd4.level:shannon   1   17.9851 79.720 41.586
```

It looks like the interaction of the Shannon score and CD4 level contribute to a lower AIC. This would suggest that the interaction between population diversity and the presence of CD4-positive cells have an effect on the HIV viral load. This seems to work intuitively, so I will keep this interaction in the model. On the other hand, it appears that population diversity (Shannon) on its ownn is not good at explaining viral load, so we can drop the term from the model.

If we were to use the final model as a start for a backwards step optimisation we could find if we can drop some of these factors to simplify the model.

```
my.model2 <- step(my.model,
                  direction="backward")
```

```
## Start:  AIC=33.36
## viral.load ~ site + cd4.level + shannon + distance + cd4.level:shannon +
##     site:shannon + site:distance
##
##                     Df Sum of Sq    RSS    AIC
## <none>                           61.735 33.359
## - site:distance      1    3.2364 64.972 33.403
## - site:shannon       1    9.3375 71.073 36.993
## - cd4.level:shannon  1   17.9851 79.720 41.586
```

No terms to remove. From the stepwise optimisation in both directions it looked like the cd4.level:shannon term was the one that led to the biggest reduction in AIC. I can try to start building a model from this term only.

```
my.model3 <- step(lm(viral.load ~ cd4.level:shannon),
                  scope = ~site * cd4.level * shannon * distance,
                  direction = "forward")
```

```
## Start:  AIC=46.99
## viral.load ~ cd4.level:shannon
## 
##             Df Sum of Sq    RSS    AIC
## + distance   1   21.3641 102.63 43.690
## + site       1   17.4777 106.52 45.177
## + shannon    1   12.5427 111.45 46.989
## <none>                   124.00 49.254
## + cd4.level  1    0.4756 123.52 51.101
## 
## Step:  AIC=41.15
## viral.load ~ distance + cd4.level:shannon
## 
##             Df Sum of Sq     RSS    AIC
## + site       1   17.2634  85.367 38.323
## + shannon    1   11.0122  91.619 41.150
## <none>                   102.631 43.690
## + cd4.level  1    0.1306 102.500 45.639
## 
## Step:  AIC=35.95
## viral.load ~ distance + site + cd4.level:shannon
## 
##                 Df Sum of Sq    RSS    AIC
## + shannon        1    8.8336 76.534 35.954
## <none>                       85.367 38.323
## + site:distance  1    2.3596 83.008 39.202
## + cd4.level      1    0.3304 85.037 40.168
## 
## Step:  AIC=35.95
## viral.load ~ distance + site + shannon + cd4.level:shannon
## 
##                    Df Sum of Sq    RSS    AIC
## + site:shannon      1    4.9833 71.550 35.261
## <none>                          76.534 35.954
## + cd4.level         1    3.1714 73.362 36.261
## + site:distance     1    2.3133 74.220 36.726
## + shannon:distance  1    0.2653 76.268 37.815
## 
## Step:  AIC=35.26
## viral.load ~ distance + site + shannon + shannon:cd4.level +
##     site:shannon
## 
##                    Df Sum of Sq    RSS    AIC
## + cd4.level         1    6.5790 64.972 33.403
## <none>                          71.550 35.261
## + site:distance     1    3.0118 68.539 35.541
## + shannon:distance  1    0.3149 71.236 37.085
```

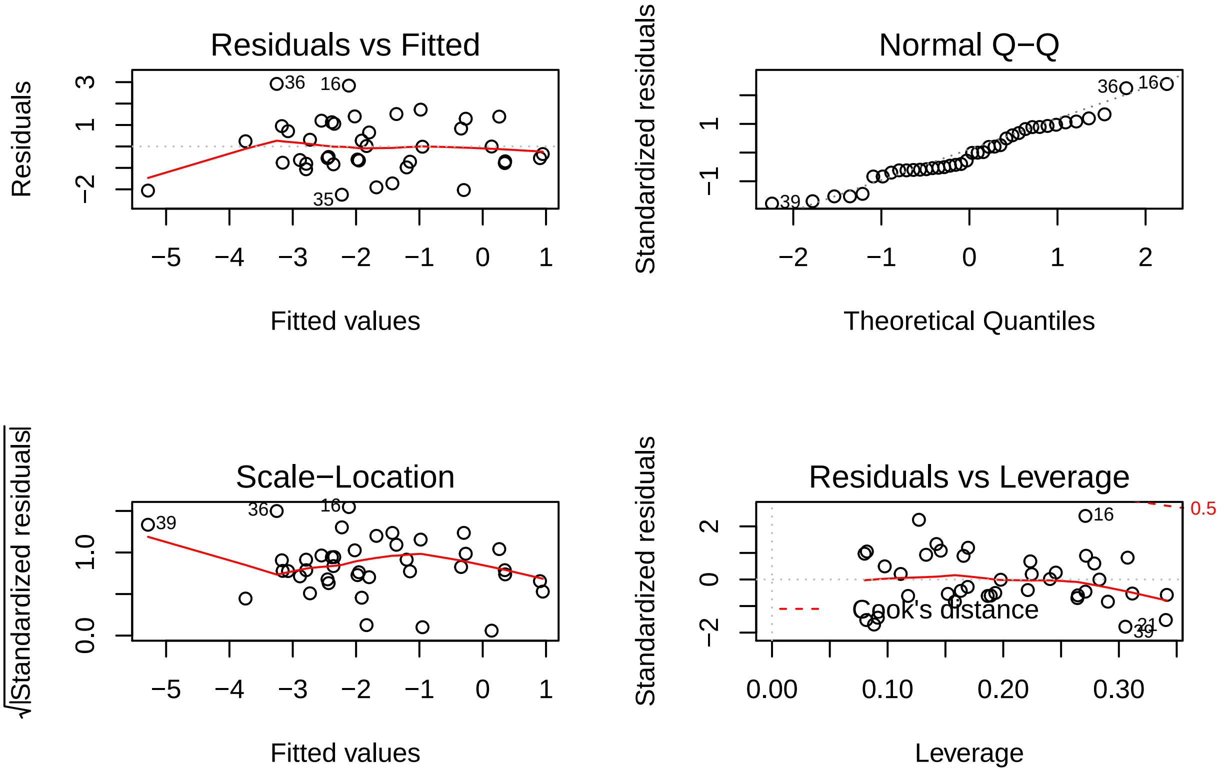
```
##
## Step:  AIC=33.4
## viral.load ~ distance + site + shannon + cd4.level + shannon:cd4.level +
##     site:shannon
##
##                      Df Sum of Sq    RSS    AIC
## + site:distance       1    3.2364 61.735 33.359
## <none>                            64.972 33.403
## + cd4.level:distance  1    0.9576 64.014 34.809
## + shannon:distance    1    0.2849 64.687 35.227
## + site:cd4.level      1    0.1866 64.785 35.288
##
## Step:  AIC=33.36
## viral.load ~ distance + site + shannon + cd4.level + shannon:cd4.level +
##     site:shannon + distance:site
##
##                      Df Sum of Sq    RSS    AIC
## <none>                            61.735 33.359
## + cd4.level:distance  1  0.236238 61.499 35.206
## + shannon:distance    1  0.209042 61.526 35.223
## + site:cd4.level      1  0.021305 61.714 35.345
```

The resulting model is the same as obtained for the start.model testing in both directions. Given that I have tried searching for a model in both directions, and starting from 3 different models, it seems likely that this is the best fitting model with this data.

```
final.model <- my.model3
summary(final.model)
```

```
##
## Call:
## lm(formula = viral.load ~ distance + site + shannon + cd4.level +
##     shannon:cd4.level + site:shannon + distance:site)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.2501 -0.7629 -0.1836  0.9760  2.9160
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)              -3.3027     1.2711  -2.598  0.01405 *
## distance                 -0.6635     0.5683  -1.167  0.25164
## sitespinalCord            2.0050     1.3708   1.463  0.15331
## shannon                   1.0309     0.3149   3.274  0.00255 **
## cd4.levello               2.0448     1.0889   1.878  0.06953 .
## shannon:cd4.levello      -1.0007     0.3277  -3.053  0.00453 **
## sitespinalCord:shannon   -0.7248     0.3294  -2.200  0.03514 *
## distance:sitespinalCord  -0.9513     0.7345  -1.295  0.20451
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.389 on 32 degrees of freedom
## Multiple R-squared:  0.5363, Adjusted R-squared:  0.4349
## F-statistic: 5.288 on 7 and 32 DF,  p-value: 0.0004355
```

```
par(mfrow = c(2,2))
plot(final.model)
```



The plots for the model's residuals raise no concerns with the normality of residual distribution and the homoscedasticity of errors.

viral.load ~ distance + site + shannon + cd4.level + shannon:cd4.level + site:shannon + distance:site + errors

According to this model, HIV viral load is best explained by taking into account:

- distance, how much the viruses have evolved from a reference sequence. Given the viral load is negatively correlated with the distance, suggesting the larger the distance, the lower the viral load this could be seen as a fitness loss associated with the evolutionary pressure of replicating within the host;
- site, the tissue type of the sample. This will affect the environment of replication as well as the accessbility of the site to CD4-positive cells;
- Shannon score, a measurement of the viral population diversity: more diverse populations may have adaptive advantages in the response to the immune system or drugs. On the other hand, specific viruses can be randomly selected through bottleneck effects;
- CD4 level, how many CD4-positive cells were found in the sample: this would be an indicator of an ongoing immune response;
- Shannon score interaction with CD4 levels: this could be giving an indicator of effective vs. non-effective immune-responses as more diverse populations may evolve under selective pressure or be restricted to a few evading viruses;
- site interaction with Shannon score: could indicate how successfully viruses are replicating or being eliminated from a given tissue;
- distance to reference interaction with site: once the virus infects a tissue, it may need to adapt to that tissue's environment and hence it will evolve towards or away from the genome reference used.

All these appear to be sensible factors in explaining the viral load in a given sample.