hw2

Jake VanCampen

November-10-17

Problem 1

Multiple Linear Regression

To analyze the effect of various environmental variables on the abundance of ARID plants, multivariate linear regression was performed on the dataset.

```
library(tidyverse)
library(car)
library(knitr)
library(broom)
library(Hmisc)

multi_data <- read_tsv('multivariate-1.tsv')
multi_data
```

```
## # A tibble: 73 x 7
##     ARID   MAP   MAT JJAMAP DJFMAP   LONG   LAT
##    <dbl> <int> <dbl>  <dbl>  <dbl>  <dbl> <dbl>
##  1  0.65   199  12.4   0.12   0.45 119.55 46.40
##  2  0.65   469   7.5   0.24   0.29 114.27 47.32
##  3  0.76   536   7.2   0.24   0.20 110.78 45.78
##  4  0.75   476   8.2   0.35   0.15 101.87 43.95
##  5  0.33   484   4.8   0.40   0.14 102.82 46.90
##  6  0.03   623  12.0   0.40   0.11  99.38 38.87
##  7  0.00   259  14.5   0.47   0.17 106.75 32.62
##  8  0.02   969  15.3   0.30   0.14  96.55 36.95
##  9  0.05   542  13.9   0.44   0.13 101.53 35.30
## 10  0.05   421   8.5   0.31   0.14 104.60 40.82
## # ... with 63 more rows
```

```
# 73 X 7 matrix
```

The response variable in this dataset is the abundance of ARID plants, while the predictor variables are the amounts of percipitation that fall throughout different parts of the year.

Correlation Matrix

It is necessary to analyze the assumptions of a multiple linear model. An important assumption is that predictor variables do not show colinearity. To address this we can display a correlation matrix of the predictors.

```
# filter out the response variable to only look at predictors
predictors_data <- multi_data[2:7]
predictors_matrix <- cor(predictors_data)

kable(predictors_matrix)
```

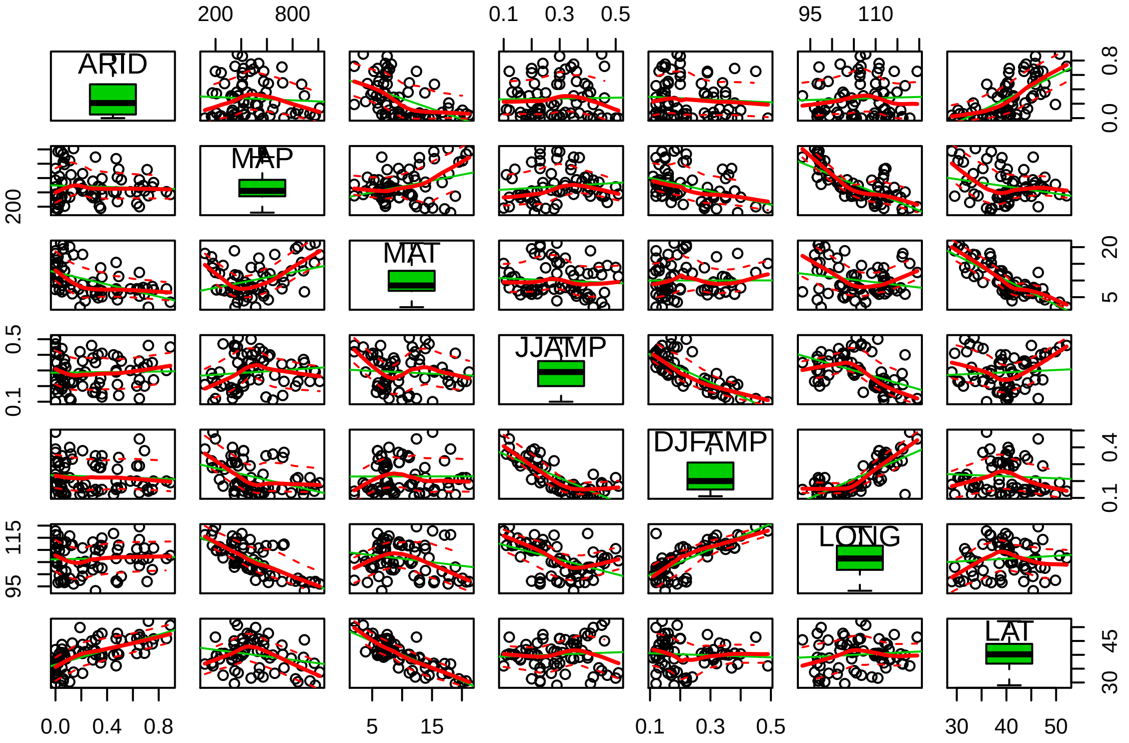
	MAP	MAT	JJAMAP	DJFMAP	LONG	LAT
MAP	1.0000000	0.3550908	0.1122590	-0.4045124	-0.7336870	-0.2465058
MAT	0.3550908	1.0000000	-0.0807713	0.0014780	-0.2131091	-0.8385904
JJAMAP	0.1122590	-0.0807713	1.0000000	-0.7915404	-0.4915577	0.0741750
DJFMAP	-0.4045124	0.0014780	-0.7915404	1.0000000	0.7707440	-0.0651248
LONG	-0.7336870	-0.2131091	-0.4915577	0.7707440	1.0000000	0.0965528
LAT	-0.2465058	-0.8385904	0.0741750	-0.0651248	0.0965528	1.0000000

The correlation matrix shows strong correlations between MAP and LONG (-0.73), MAT and LAT (-0.84), JJAMAP and DJFMAP (-0.79), and DJFMAP with LONG (0.77). This evidence of colinearity will need to be furthur explored.

Scatterplot Matrix

These relationships can be visualized with a scatterplot matrix.

```
scatterplotMatrix(~multi_data$ARID+multi_data$MAP+multi_data$MAT+
                    multi_data$JJAMAP+
                    multi_data$DJFMAP+
                    multi_data$LONG+
                    multi_data$LAT,
                    diagonal = 'boxplot',  var.labels =
                    c('ARID', 'MAP', 'MAT', 'JJAMP', 'DJFAMP', 'LONG', 'LAT'))
```



Again we see strong linearity between MAP and LONG, MAT and LAT, DJFAMP and LONG. DJFAMP and JJAMP appear to have a weaker colinearity, consistent with the correlation matrix.

Analysis of the model fit

The full additive multivariate linear regression model for these data can be describes as follows:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \beta_6 x_{i6} + \epsilon_i$$

An additive multiple linear model fit to the data show the following result:

```
MLM <- lm(multi_data$ARID ~ multi_data$MAP+multi_data$MAT+
                    multi_data$JJAMAP+
                    multi_data$DJFMAP+
                    multi_data$LONG+
                    multi_data$LAT)

summary(MLM)
```

```
##
## Call:
## lm(formula = multi_data$ARID ~ multi_data$MAP + multi_data$MAT +
##     multi_data$JJAMAP + multi_data$DJFMAP + multi_data$LONG +
##     multi_data$LAT)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.35531 -0.15057  0.01143  0.11100  0.46400
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -2.3957310  1.0279698  -2.331   0.0228 *
## multi_data$MAP     0.0002443  0.0001804   1.354   0.1804
## multi_data$MAT     0.0074275  0.0096448   0.770   0.4440
## multi_data$JJAMAP -0.1750540  0.3941150  -0.444   0.6584
## multi_data$DJFMAP -0.5048251  0.5942319  -0.850   0.3987
## multi_data$LONG    0.0100024  0.0083232   1.202   0.2338
## multi_data$LAT     0.0393018  0.0082364   4.772 1.05e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.198 on 66 degrees of freedom
## Multiple R-squared:  0.4728, Adjusted R-squared:  0.4249
## F-statistic: 9.865 on 6 and 66 DF,  p-value: 9.495e-08
```

The model shows the only predictor for which we reject the null hypothesis is LAT. Adjusted and multiple R-squared less than 0.5 show a poor fit of the model. Additional analysis of colinarity can be determined from the tolerance values of the predictor variables.

Tolerance analysis

```
tol <- 1/vif(MLM)
kable(tol)
```

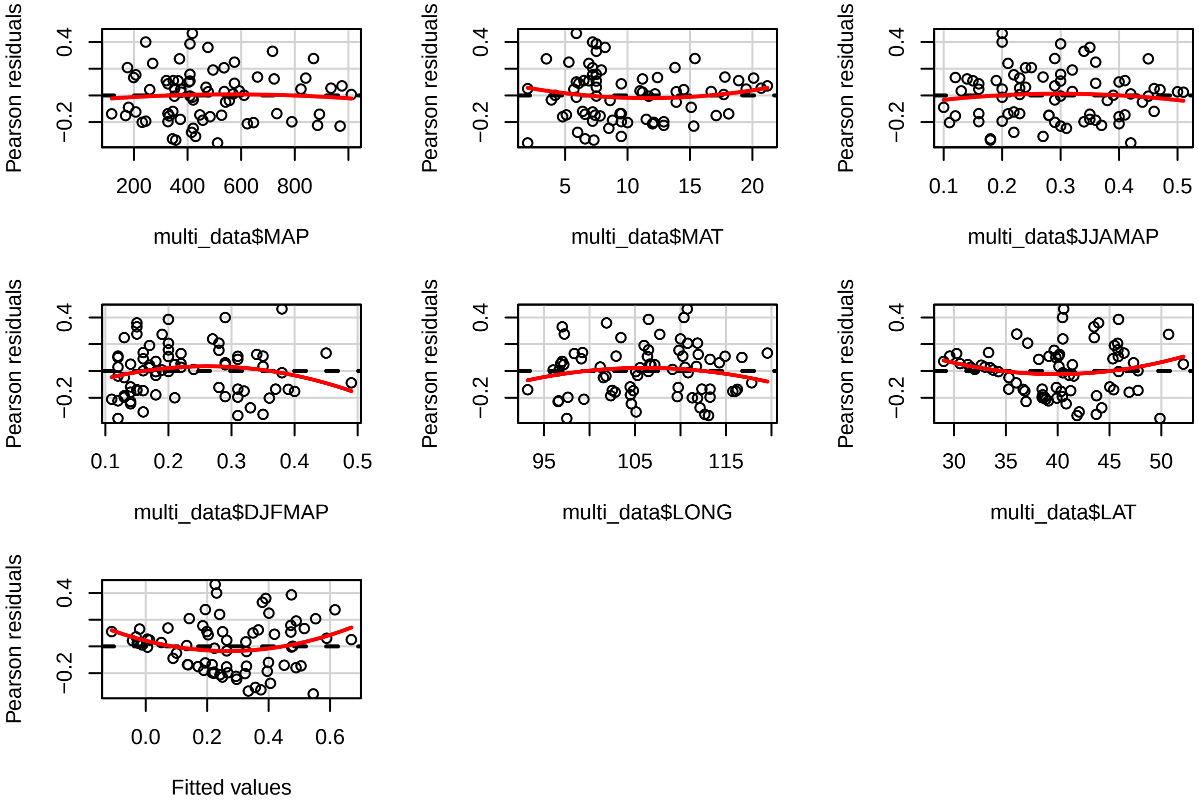
multi_data$MAP	0.3572159
multi_data$MAT	0.2671810

multi_data$JJAMAP	0.3161340
multi_data$DJFMAP	0.1751217
multi_data$LONG	0.1898391
multi_data$LAT	0.2854914

The low tolerance values for especially DJFMAP (0.17) and LONG (0.189) provide more evidence for colinearity.

Residual Analysis

```
residualPlots(MLM)
```



```
##                      Test stat Pr(>|t|)
## multi_data$MAP          -0.462    0.646
## multi_data$MAT           0.915    0.364
## multi_data$JJAMAP       -0.738    0.463
## multi_data$DJFMAP       -1.420    0.160
## multi_data$LONG         -1.291    0.201
## multi_data$LAT           1.635    0.107
## Tukey test               2.294    0.022
```

The residual plots show non-linearity of the predictors, furthur decreasing the model fit.

Conclusions

These data do not meet the major assumptions for using hypothesis tests derived from the additive multiple linear model. For one, scatterplot matrices show poor linearity of the response variable to the predictors.

Secondly, the variance of the residuals show heterogeneity. Thirdly, there is strong evidence for multicolinearity of the predictor variables shown in tolerance values < 0.2 for LONG and LAT, and the correlation matrix of the predictors. This model likely needs to be reduced, broken apart to more appropriately fit the data and make assumptions about the predictors.

Problem 2

Analyze a fabricated dataset relating concentration of bacteria in a biofilm to a four-level categorical variable using single-factor ANOVA model.

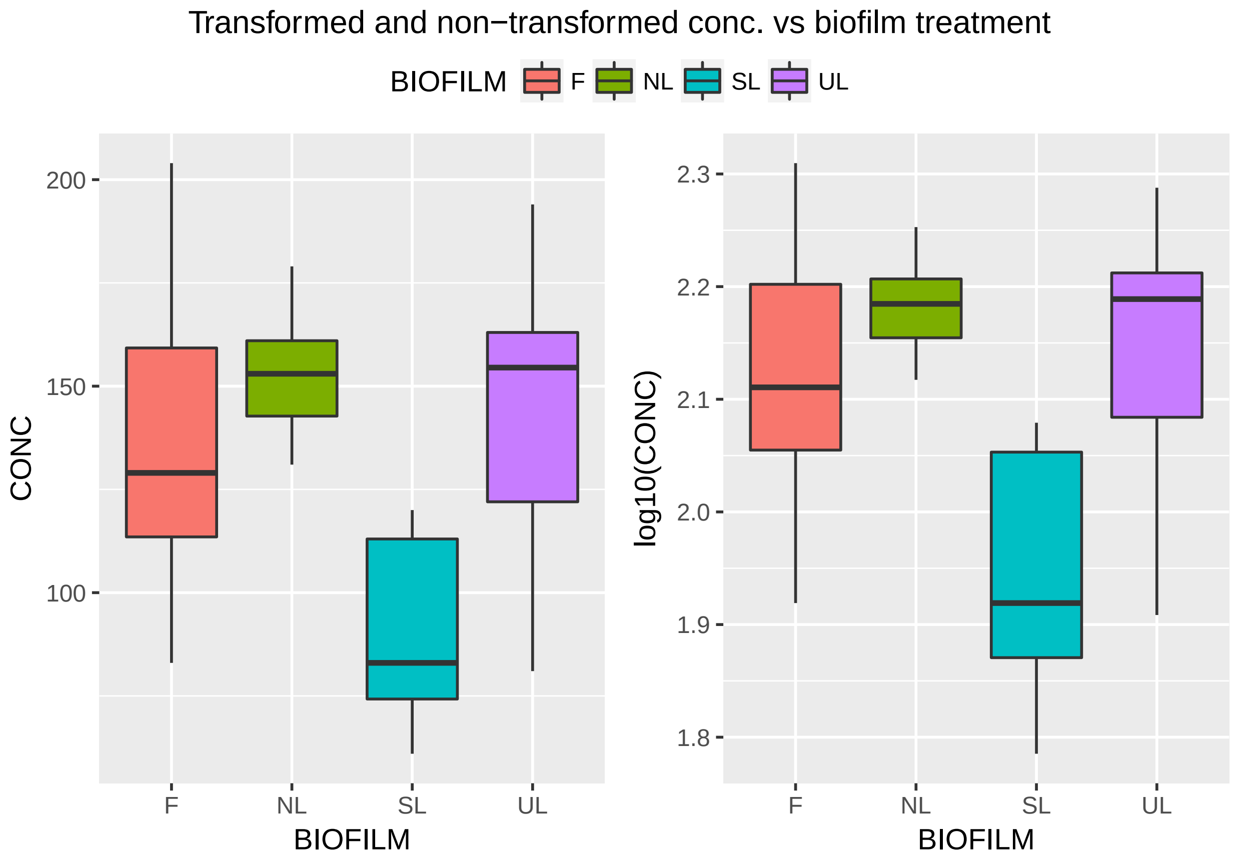
Exploratory Data Analysis

Bacteria were sampled from different known 'levels' of soil type in this experiment. We can consider these levels fixed variables, as the results from each soil type cannot be extrapolated to different soil types.

```r
library(ggpubr)

biofilm <- read_tsv('biofilm-1.tsv')
biofilm
```

```
## # A tibble: 80 x 2
##    BIOFILM  CONC
##      <chr> <int>
##  1      SL    61
##  2      SL   113
##  3      SL   120
##  4      SL    75
##  5      SL    72
##  6      SL    83
##  7      SL    95
##  8      SL    66
##  9      SL   113
## 10      SL   119
## # ... with 70 more rows
```

```r
log10 <- ggplot(biofilm) +
  geom_boxplot(aes(BIOFILM, log10(CONC), fill = BIOFILM))

non_trans <- ggplot(biofilm) +
  geom_boxplot(aes(BIOFILM, CONC, fill = BIOFILM))

ggarrange(non_trans, log10, common.legend = TRUE) %>% annotate_figure( top = 'Transformed and non-trans
```



These data show some skewness in their distributions suggesting potential variance heterogeneity.

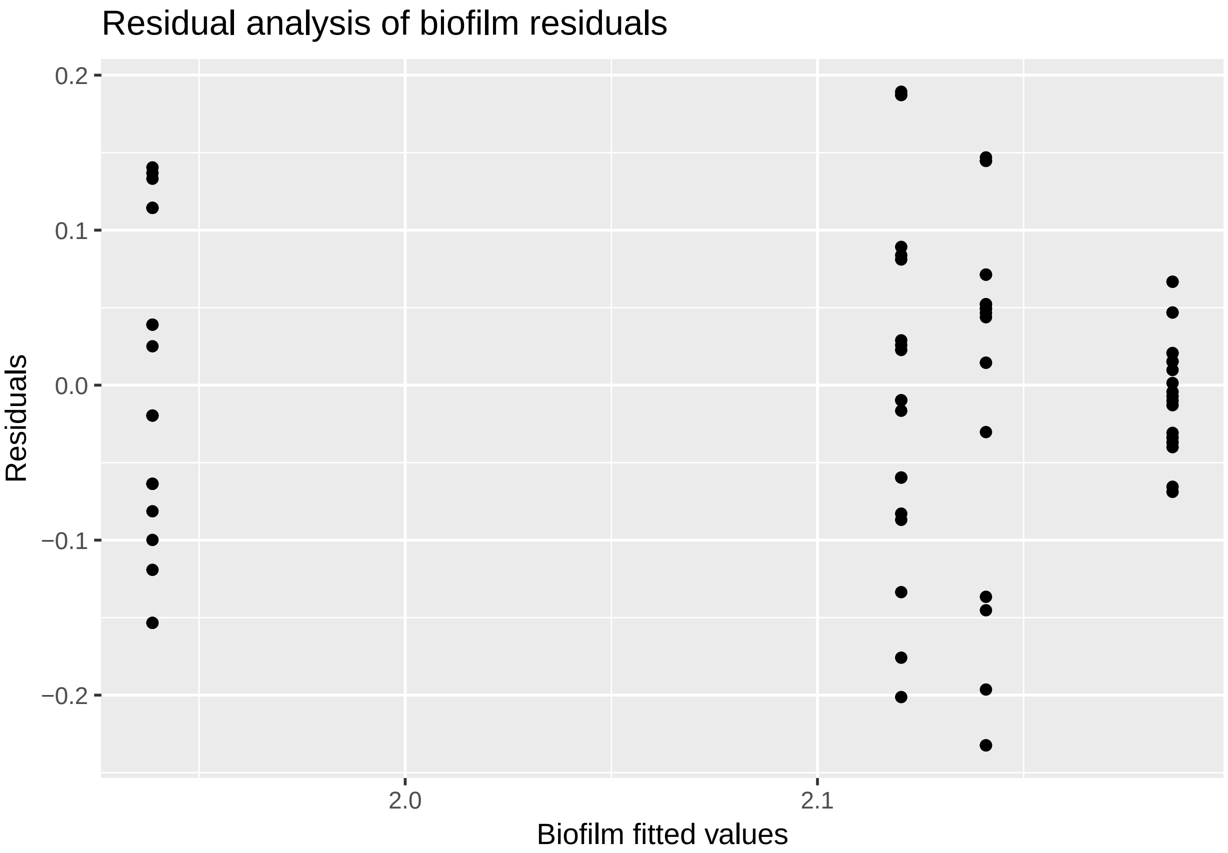
Single Factor Anova

To assess how these data vary over categorical variables, a single-factor ANOVA can be used to analyze the data, and the residuals analyzed for homogeneity of variance.

```
library(multcomp)

biofilm_aov = aov(log10(biofilm$CONC) ~ as.factor(biofilm$BIOFILM))

summary(biofilm_aov)
```

```
##                            Df Sum Sq Mean Sq F value   Pr(>F)
## as.factor(biofilm$BIOFILM)  3 0.7094  0.2365   24.14 4.48e-11 ***
## Residuals                  76 0.7445  0.0098
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
ggplot(biofilm_aov) +
  geom_point(aes(biofilm_aov$fitted.values, biofilm_aov$residuals)) +
  ylab('Residuals') +
  xlab('Biofilm fitted values') +
  ggtitle('Residual analysis of biofilm residuals')
```



Residual analysis shows relatively homogenous variance, meeting an assumption of single factor ANOVA. This model rejects the null hypothesis of no difference in means between biofilm environments (fixed effects, single factor ANOVA: F 3, 76 = 24.14, p < 0.001)

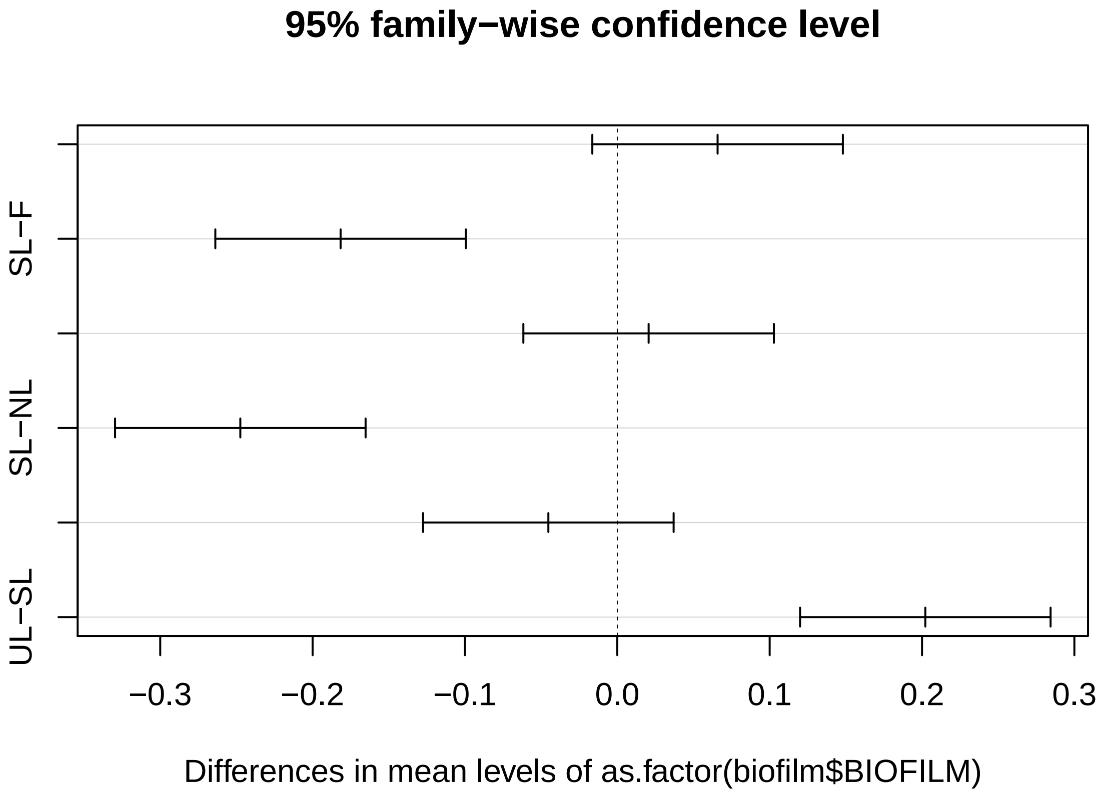
Post-hoc Comparisons

To look at the difference in means between individual effects, a Tukey's post-hoc means test was used.

```
tmeans <- TukeyHSD(biofilm_aov)

# unplanned comparisons
plot(tmeans)

# table of differences
kable(tmeans$`as.factor(biofilm$BIOFILM)`)
```

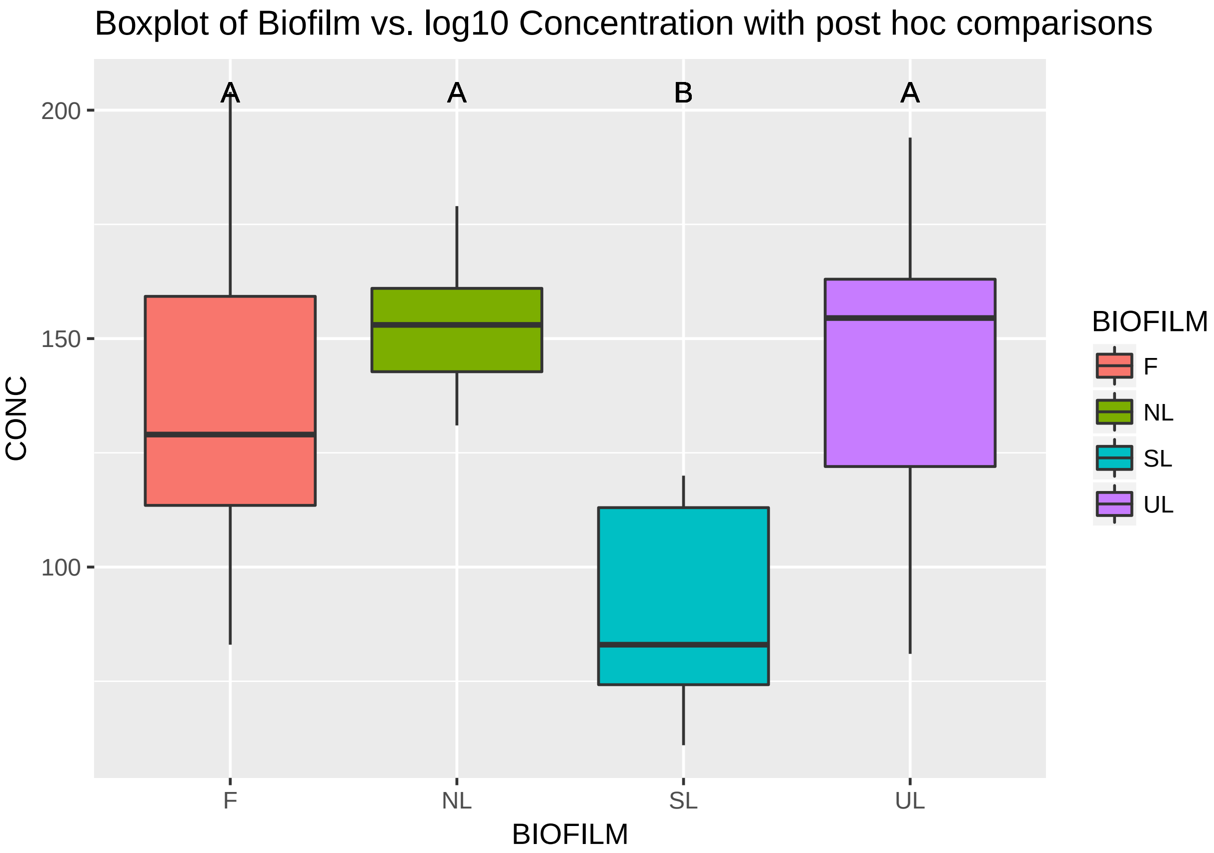
	diff	lwr	upr	p adj
NL-F	0.0658102	-0.0164051	0.1480255	0.1615567
SL-F	-0.1816196	-0.2638349	-0.0994043	0.0000008
UL-F	0.0205500	-0.0616653	0.1027653	0.9129190
SL-NL	-0.2474298	-0.3296451	-0.1652145	0.0000000
UL-NL	-0.0452602	-0.1274755	0.0369551	0.4750032
UL-SL	0.2021696	0.1199543	0.2843849	0.0000001

```
plot(tmeans)
```



It is clear that the the significant differences are between UL and SL, SL and NL, and SF with F ($p < 0.0001$). The group differences in means can then be used to annotate the original boxplot of the log10(CONC) vs the biofilm condition.

```
biofilm$posthoc[biofilm$BIOFILM == 'SL'] = 'B'
biofilm$posthoc[biofilm$BIOFILM == 'NL'] = 'A'
biofilm$posthoc[biofilm$BIOFILM == 'UL'] = 'A'
biofilm$posthoc[biofilm$BIOFILM == 'F'] = 'A'

biofilm = biofilm %>% mutate_if(.predicate = is.character, .funs = as.factor)

ggplot(biofilm) +
  geom_boxplot(aes(BIOFILM, CONC, fill = BIOFILM)) +
  geom_text(data = biofilm, aes(x = BIOFILM, y = max(CONC), label = posthoc)) +
  ggtitle('Boxplot of Biofilm vs. log10 Concentration with post hoc comparisons')
```



Planned Comparisons

Test that the mean of UL = NL, and that SL is 2 X F.

```
# define the contrasts for comparison
biofilm_contrasts <- contrasts(biofilm$BIOFILM) <- cbind(c(0,-1,0,1), c(2,0,-1,0))

# verify orthogonality
crossprod(biofilm_contrasts)
```

```
##      [,1] [,2]
## [1,]    2    0
## [2,]    0    5
```

```
# contrast labels
contrasts_list <- list(BIOFILM = list('NL vs UL' = 1, 'F vs SL' = 2))

summary(aov(log10(CONC) ~ BIOFILM, data = biofilm), split = contrasts_list)
```

```
##                      Df Sum Sq Mean Sq F value   Pr(>F)    
## BIOFILM               3 0.7094 0.23647  24.139 4.48e-11 ***
##   BIOFILM: NL vs UL   1 0.0205 0.02048   2.091    0.152    
##   BIOFILM: F vs SL    1 0.1777 0.17770  18.140 5.81e-05 ***
## Residuals            76 0.7445 0.00980                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The soil type has a significant effect on the mean log10 concentration of bacteria in the sample (F3,76 =

24.14, $p < 0.001$). No difference in means was found between NL and UL soiltypes (F1,76 = 2.1, $p = 0.152$), but SL level mean was found to be twice that of F level mean (F1,76 = 18.14, $p < 0.001$).