Biostatistics
Assignment 4

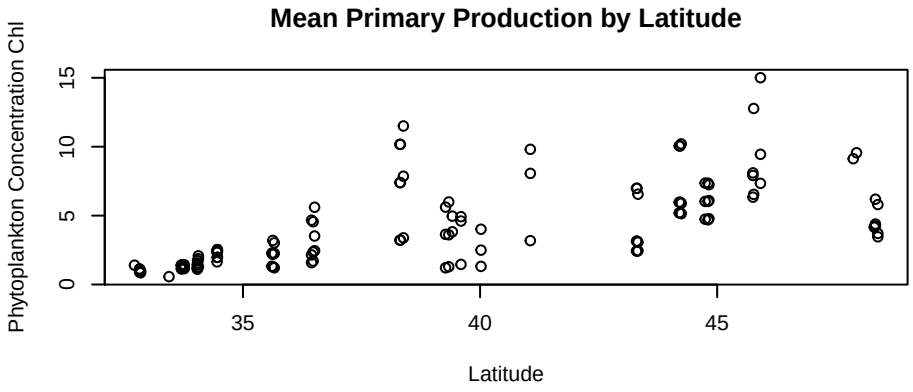
Problem 1: Environmental forcing in Eastern Boundary Current Systems

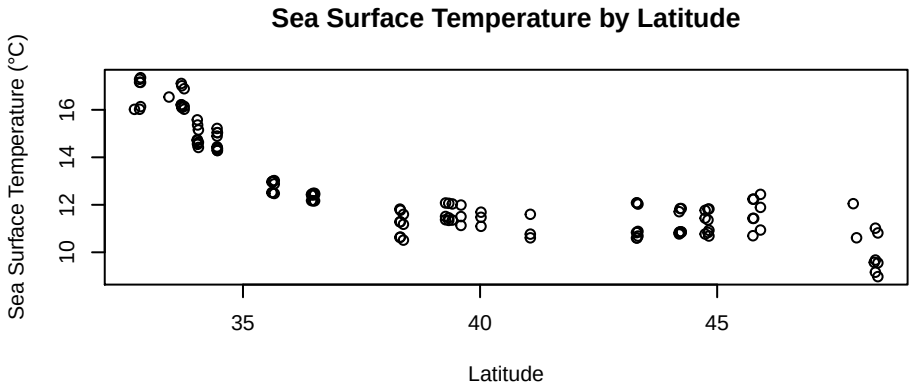
Part 1

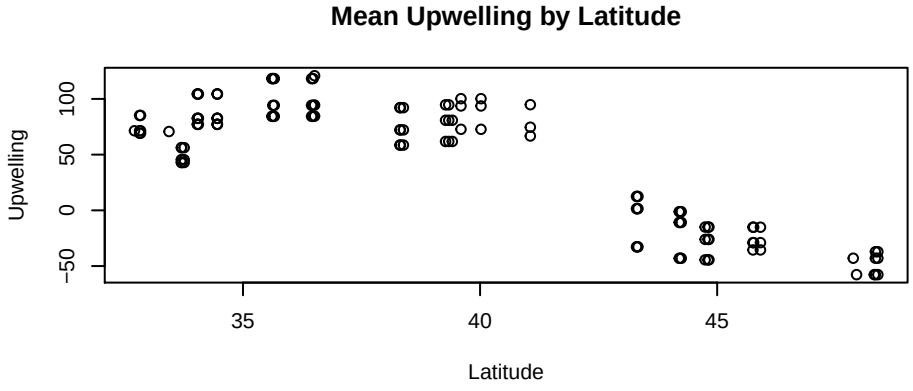
R-code

```
## Download the data
d1 <- read.csv("http://www.northeastern.edu/synchrony/stats/data/assn4_dataset1.csv")

## Use the str function to determine the nature of each variable in the
## dataset
str(d1)

## 'data.frame': 122 obs. of  7 variables:
##  $ latitude    : num  48.4 48.3 48.3 47.9 44.8 ...
##  $ year        : int  2000 2000 2000 2000 2000 2000 2000 2000 2000 2000 ...
##  $ chl         : num  3.46 4.38 4.16 9.56 4.78 ...
##  $ sst         : num  9.56 9.67 9.58 10.61 10.69 ...
##  $ upwelling   : num  -57.7 -57.7 -57.7 -57.7 -44.4 ...
##  $ mussel_cover: num  59.77 21.57 4.83 3.87 65.17 ...
##  $ region      : Factor w/ 2 levels "north","south": 1 1 1 1 1 1 1 1 1 1 ...

## Plot the response variables chl, SST and upwelling as a function of
## explanatory variable latitude on a single figure

## Plot a 3 row, 1 column figure
par(mfrow = c(3, 1))

## Plot w/ response variable chl
plot(d1$latitude, d1$chl, main = "Mean Primary Production by Latitude", xlab = "Latitude",
    ylab = "Phytoplankton Concentration Chl")

## Plot w/ response variable SST
plot(d1$latitude, d1$sst, main = "Sea Surface Temperature by Latitude", xlab = "Latitude",
    ylab = "Sea Surface Temperature (C)")

## Plot w/ response variable Upwelling
```

```
plot(d1$latitude, d1$upwelling, main = "Mean Upwelling by Latitude", xlab = "Latitude",
    ylab = "Upwelling")
```







Interpretation

The resulting figure suggests the following relationships:

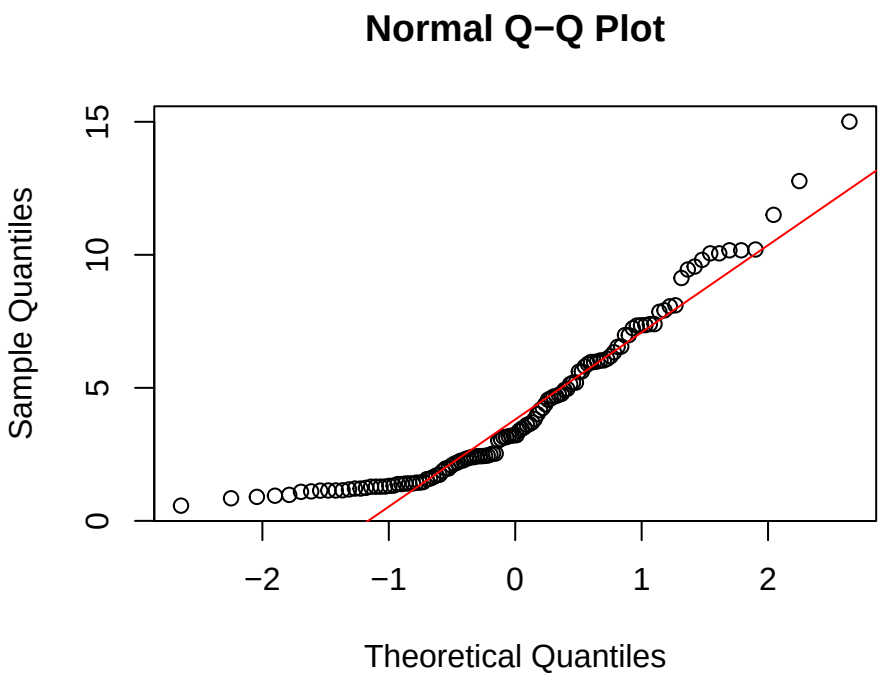
- **Relationship Between Latitude and Mean Primary Production**
  - There is a positive relationship between latitude and mean primary production (chl). According to the figure, there is a general trend of an increase in phytoplankton concentration as latitude increases.
  - The relationship between latitude and mean primary production (chl) appears linear, although it is difficult to be sure due to the large amount of variation in the data.
  - The variation in mean primary production (chl) does not appear to be constant across the latitudinal range. There is less variation at lower latitudes than there is at higher latitudes.
- **Relationship Between Latitude and Sea Surface Temperature**
  - There is a negative relationship between latitude and sea surface temperature. According to the figure, there is a general trend of a decrease in sea surface temperature as latitude increases.
  - The relationship between latitude and sea surface temperature does not appear to be linear.
  - The variation in sea surface temperature appears to be relatively constant across the latitudinal range, but the figure shows a small increase in variation as latitude increases.
- **Relationship Between Latitude and Upwelling**
  - There is a negative relationship between latitude and upwelling. According to the figure, there is a general trend of a decrease in upwelling as latitude increases.
  - The relationship between latitude and upwelling does not appear to be linear, however due to variation and potential holes of data in the latitude range, it is difficult to be sure.
  - The variation in sea surface temperature appears to be relatively constant across the latitudinal range, but the figure shows a potential small decrease in variation at high latitudes.

Part 2

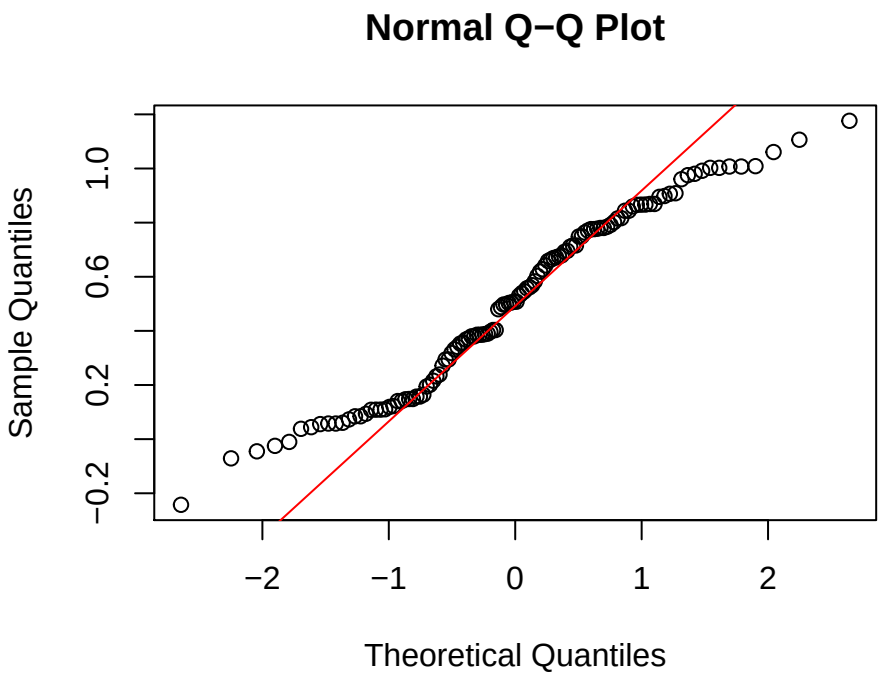
R-code

```
## Use the lm function to fit a multiple regression model that relates chl to
## SST, upwelling, and latitude, transforming the response variable if
## necessary.

## Use function qqnorm to generate a Q-Q plot of chl and then use qqline to
## plot the line in red
```

```
qqnorm(d1$chl)
qqline(d1$chl, col = "red")
```




```
## Use function qqnorm to generate a Q-Q plot of log10(chl) and then use
## qqline to plot the line in red
qqnorm(log10(d1$chl))
qqline(log10(d1$chl), col = "red")
```




```
## A transformation is necessary effect of lat, sst, and upwelling on chl
mod.lm.prob1p2 <- lm(log10(chl) ~ latitude + sst + upwelling, data = d1)
summary(mod.lm.prob1p2)

##
## Call:
## lm(formula = log10(chl) ~ latitude + sst + upwelling, data = d1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.4826 -0.1071  0.0112  0.1141  0.4896
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.359802   0.976832    0.37   0.7133
## latitude     0.024690   0.016752    1.47   0.1432
## sst         -0.064434   0.023614   -2.73   0.0073 **
## upwelling   -0.000221   0.000979   -0.23   0.8221
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 0.208 on 118 degrees of freedom
## Multiple R-squared:  0.605,Adjusted R-squared:  0.595
## F-statistic: 60.2 on 3 and 118 DF,  p-value: <2e-16
```

Interpretation

R^2 is the coefficient of determination, which is the proportion of variation of in the response variable explained by the explanatory variables. In this case, the R^2 value represents the proportion of variation in mean primary production (chl) explained by latitude, sea surface temperature, and upwelling. The p-value represents the probability of the data given that the null hypothesis is true. In this case, the null hypothesis is that there is no relationship (slopes are 0) between mean primary production (chl) and latitude, sea surface temperature, and upwelling.

The R^2 value of 0.6482 and p-value that is $< 2.2 \times 10^{-16}$ means that mean primary production (chl) is significantly related to latitude, sea surface temperature, and upwelling combined. The R^2 value of 0.6482 shows that most of the variation in the data is accounted for through the explanatory variables or that the correlation is a relatively strong fit.

Part 3

Interpretation

Because the plots show the the response variables chl, SST and upwelling as a function of explanatory variable latitude and the multiple regression shows the response variable chl as a result of latitude, SST, and upwelling, a direct comparison cannot be made between the two. However, because the initial graphs show potentially strong relationships between sea surface temperature and latitude, and upwelling and latitude, it is likely that latitude is highly redundant and should be removed for violated the assumption of non-collinearity. This would suggest potential issues with the output of the multiple regression. Because the output of the multiple regression shows non-significant effect on chl for latitude and upwelling, it is therefore possible that if latitude is removed, upwelling may show a significant effect on chl.

Part 4

R -code

```
## Determine whether your data adhere to the assumptions of multiple
## regression by computing the VIF and the correlation between each pair of
## explanatory variables

require(car, quiet = TRUE)
vif(mod.lm.prob1p2)
```

```
##  latitude       sst upwelling
##    19.241     6.906     8.143

## Determine correlation between latitude and sst
summary(lm(latitude ~ sst, data = d1))$coef

##             Estimate Std. Error t value  Pr(>|t|)
## (Intercept)   63.732     1.5911   40.05 2.689e-71
## sst           -1.921     0.1241  -15.48 2.190e-30

## Determine correlation between latitude and upwelling
summary(lm(latitude ~ upwelling, data = d1))$coef

##             Estimate Std. Error t value   Pr(>|t|)
## (Intercept) 42.69479    0.30385  140.51 5.365e-135
## upwelling   -0.07604    0.00436  -17.44  1.071e-34

## Determine correlation between sst and upwelling
summary(lm(sst ~ upwelling, data = d1))$coef

##             Estimate Std. Error t value  Pr(>|t|)
## (Intercept) 11.89854   0.215550  55.201 3.806e-87
## upwelling    0.01756   0.003093   5.678 9.658e-08
```

Interpretation

The use of multiple explanatory variables that are not perfectly independent in regression increases VIF and the standard error of each estimate, thus reducing their statistical significance. The VIF test tests for non-collinearity. The results of the `vif` test show that the explanatory variables violate the assumption of non-collinearity as each variable has a VIF of greater than 5 (Kutner et al., 2004). In addition, because latitude has a strong effect on sea surface temperature (slope= -1.921 and significant p-value) latitude and sea surface temperature are highly redundant and therefore latitude should potentially be removed. (Dormann et al., 2013).

Part 5

R-code

```
## effect of sst and upwelling on chl
mod.lm.prob1p5 <- lm(log10(chl) ~ sst + upwelling, data = d1)
summary(mod.lm.prob1p5)
```

```
##
## Call:
## lm(formula = log10(chl) ~ sst + upwelling, data = d1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.4539 -0.1321  0.0002  0.1410  0.5049
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.788083   0.123383   14.49  < 2e-16 ***
## sst         -0.095879   0.010171   -9.43  4.3e-16 ***
## upwelling   -0.001546   0.000388   -3.98  0.00012 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.209 on 119 degrees of freedom
## Multiple R-squared:  0.598,Adjusted R-squared:  0.591
## F-statistic: 88.3 on 2 and 119 DF,  p-value: <2e-16
```

Interpretation

The results of this model showed a significant effect of both sea surface temperature and upwelling on mean primary production (chl). The previous model that included latitude, however, only showed a significant effect of sea surface temperature on chl. The coefficients did not, however, switch signs. This is a result of the bouncing $\hat{\beta}$ problem which occurs due to strong collinearity between explanatory variables.

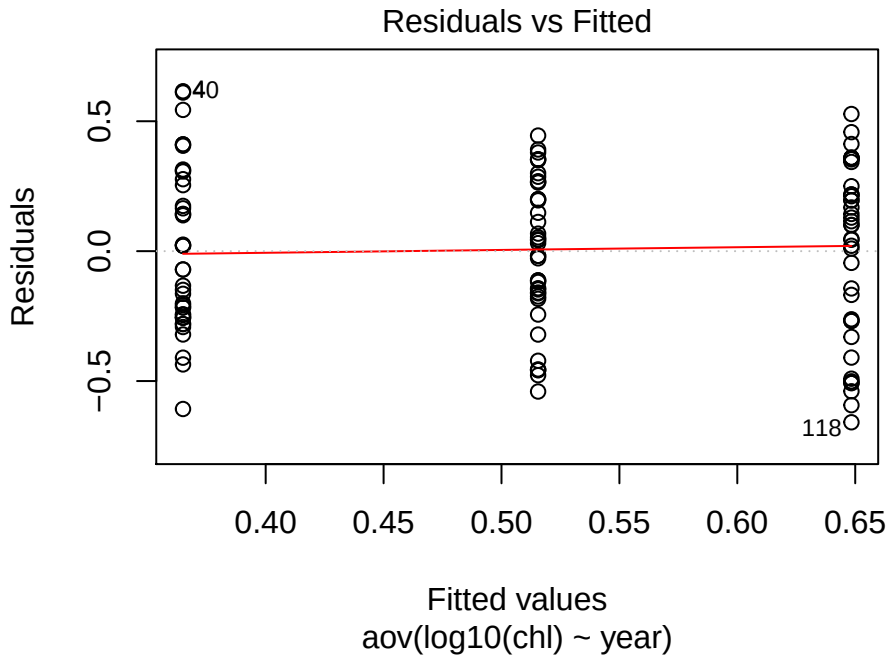
Problem 2: Temporal trends in upwelling and primary production

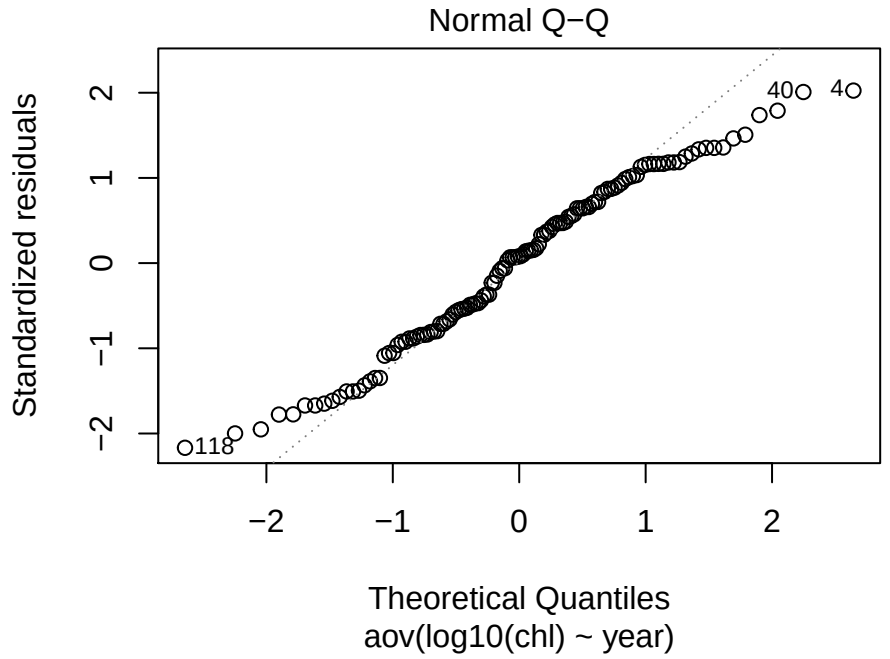
Part 1

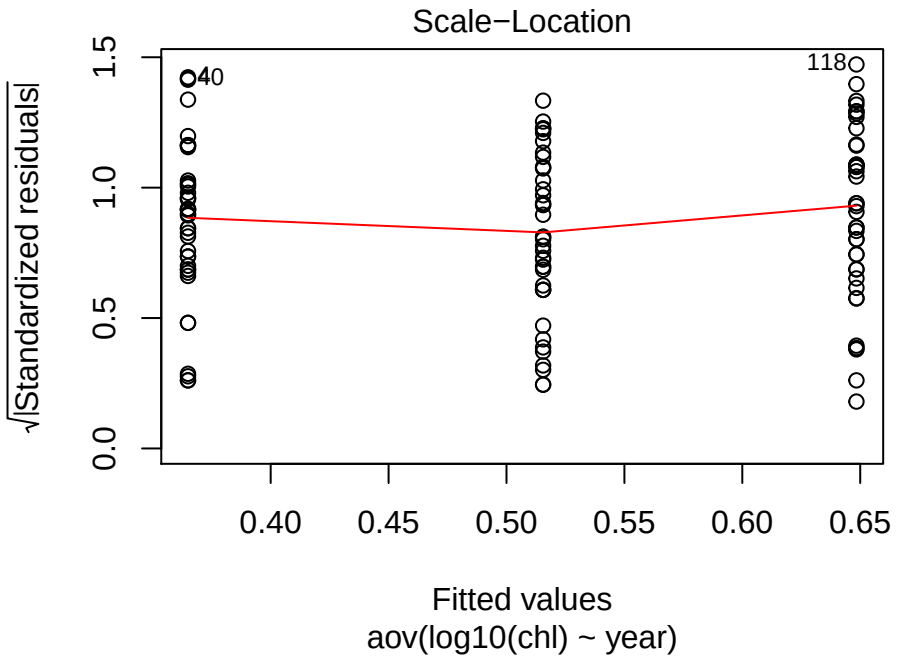
R-code

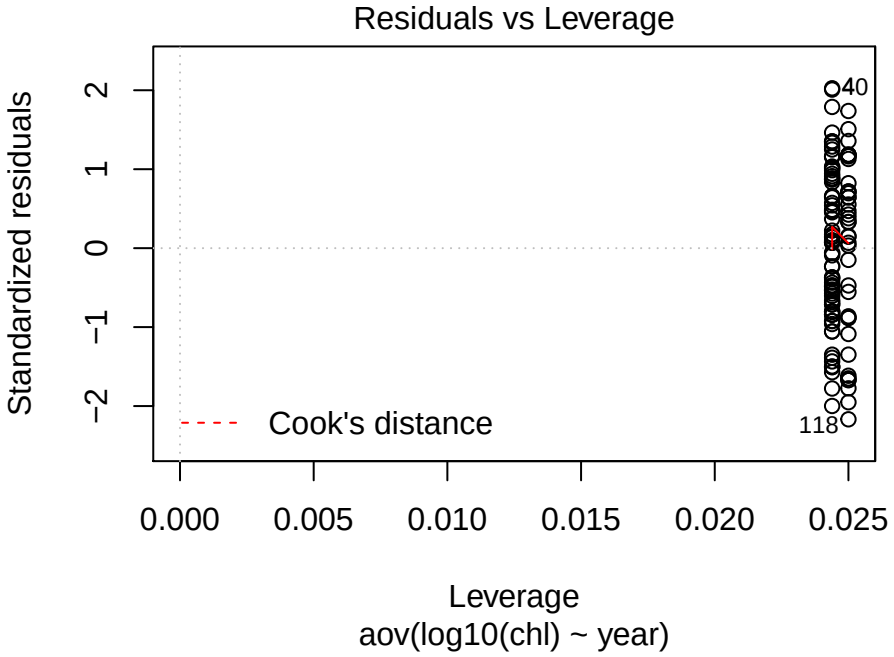
```
## Code year as a factor
d2 <- d1
d2$year <- as.factor(d2$year)
```

```
## Conduct the ANOVA
summary((p2.p1.aov <- aov(log10(chl) ~ year, data = d2)))

##              Df Sum Sq Mean Sq F value  Pr(>F)
## year          2   1.63   0.815    8.61 0.00032 ***
## Residuals   119  11.27   0.095
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## Display diagnostic plots
plot(p2.p1.aov)
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
aov(log10(chl) ~ year)
40
4
118


Scale−Location
√|Standardized residuals|
Fitted values
aov(log10(chl) ~ year)
40
118



Interpretation

The assumptions of ANOVA are as follows:

- random samples
- each population is normally distributed
- each population has the same variance (homoscedasticity)

The diagnostic plots were used to determine whether the data adheres to the assumptions of ANOVA. The Q-Q Plot confirms that the transformed chl values are normally distributed. The residuals plot has no discernible pattern (no evidence of heteroscedasticity). By meeting these assumptions of random samples, normality, and homoscedasticity, we can conduct the ANOVA.

The hypotheses are as follows:

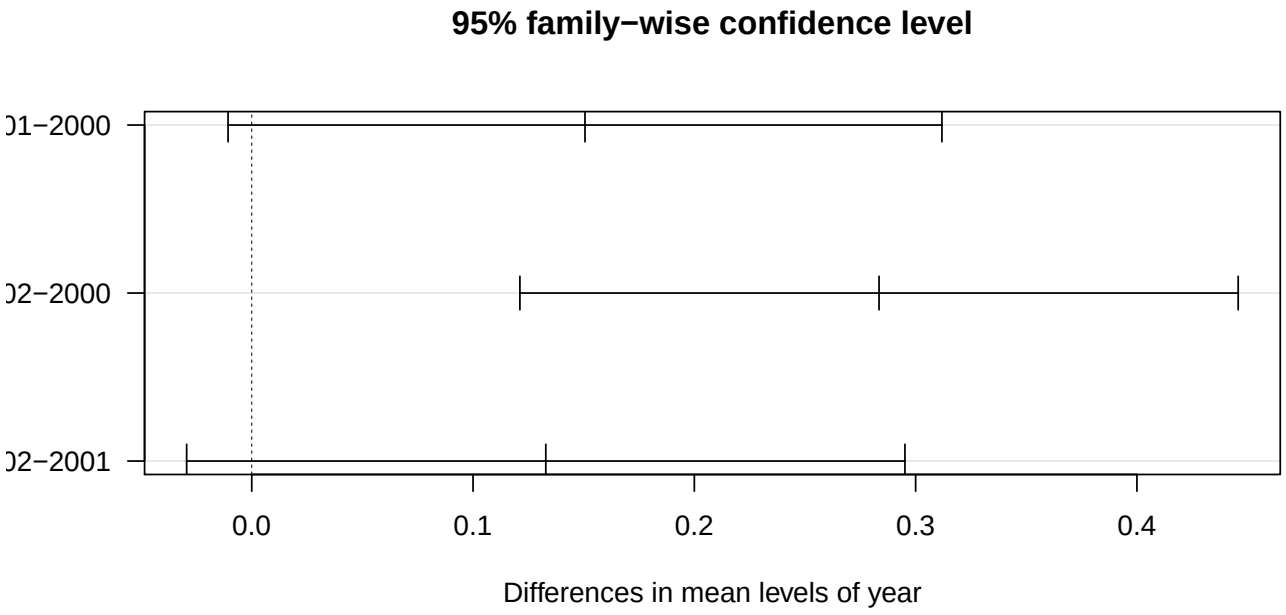
- **Null hypothesis:** There is no difference in mean annual chl by year.
- **Alternative hypothesis:** There is a difference in mean annual chl by year.

According to the results (p-value = 0.000322), we can reject the null hypothesis that there is no difference in mean annual chl by year and accept the alternative that there is a difference in mean annual chl by year at the $\alpha = 0.05$ level.

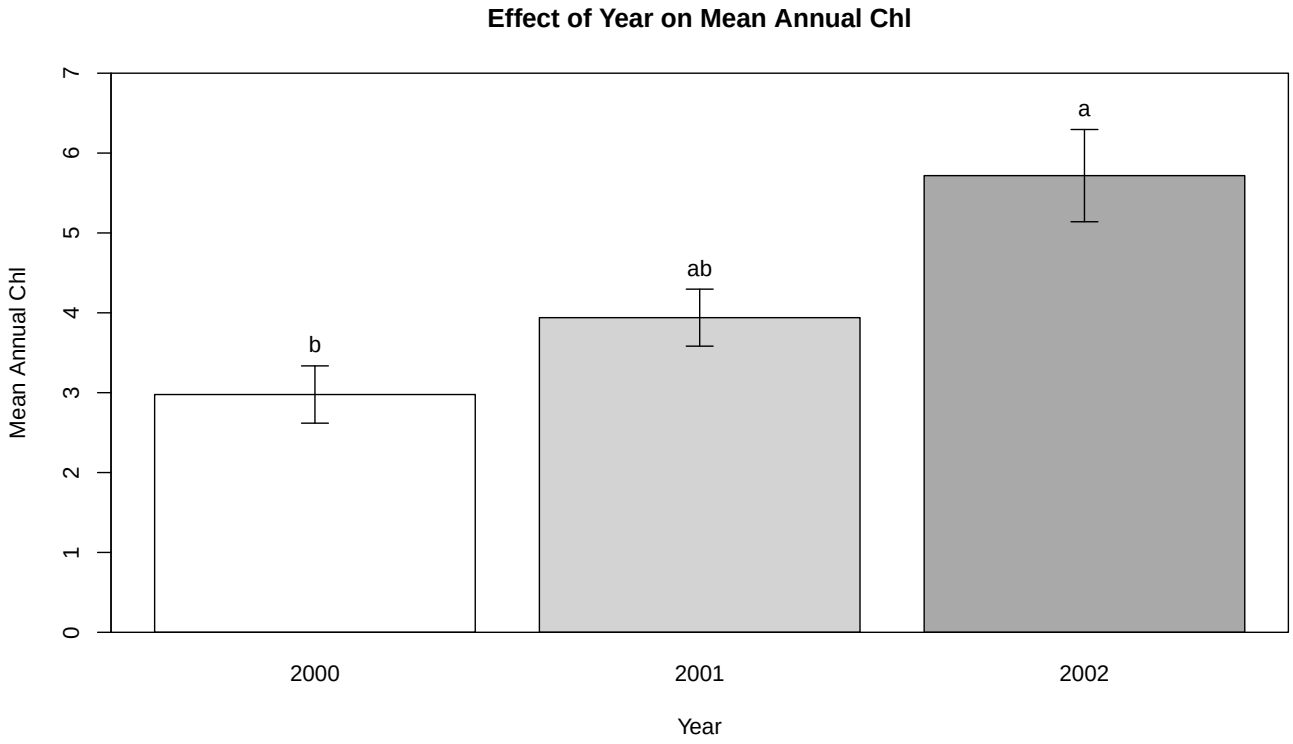
Part 2

R-code

```
## Use function TukeyHSD to perform the post-hoc comparisons on the aov
## object and print the results to the screen
(p2.p1.aov.tukey <- TukeyHSD(p2.p1.aov))

##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = log10(chl) ~ year, data = d2)
##
## $year
##             diff      lwr    upr  p adj
## 2001-2000 0.1506 -0.01070 0.3119 0.0726
## 2002-2000 0.2835  0.12117 0.4458 0.0002
## 2002-2001 0.1329 -0.02942 0.2952 0.1312

## Visualize the results of the post-hoc comparisons
plot(p2.p1.aov.tukey, las = 1)
```



```r
## Plot an even more intuitive figure, using barplot installed:
## install.packages(multcompView)
require(multcompView, quiet = TRUE)

## Assign the labels
labels.p2 <- multcompLetters(p2.p1.aov.tukey$year[, "p adj"])$Letters
## Fix the order of the labels
(labels.p2 <- labels.p2[order(names(labels.p2))])

## 2000 2001 2002
##  "b" "ab"  "a"

# Compute means
means.p2 <- aggregate(chl ~ year, mean, data = d2)

## Compute standard errors
stderrs.p2 <- aggregate(chl ~ year, FUN = function(x) sd(x)/sqrt(length(x)),
    data = d2)

## Plot the figure
p2.bp <- barplot(means.p2$chl, main = "Effect of Year on Mean Annual Chl", beside = TRUE,
    ylab = "Mean Annual Chl", xlab = "Year", ylim = c(0, 7), names = c("2000",
        "2001", "2002"), col = c("white", "lightgrey", "darkgrey"))
box()
## Add error bars
arrows(x0 = p2.bp, y0 = means.p2$chl - stderrs.p2$chl, y1 = means.p2$chl + stderrs.p2$chl,
    code = 3, angle = 90, length = 0.1)

## Add text labels
text(x = p2.bp, y = means.p2$chl + stderrs.p2$chl, labels.p2, pos = 3)
```



Interpretation

Yes, based on the figure, there is evidence that chl is increasing over time as the mean annual chl in 2000 is significantly greater than the mean annual chl in 2002.

Part 3

R-code

```
## Conduct the ANCOVA with year as factor and upwelling as covariate
summary((p2.p3.aocv <- aov(log10(chl) ~ year * upwelling, data = d2)))

##                 Df Sum Sq Mean Sq F value  Pr(>F)
## year             2   1.63    0.81    13.0 8.0e-06 ***
## upwelling        1   4.00    4.00    63.8 1.1e-12 ***
## year:upwelling   2   0.00    0.00     0.0       1
## Residuals      116   7.27    0.06
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Interpretation

The hypotheses are as follows:

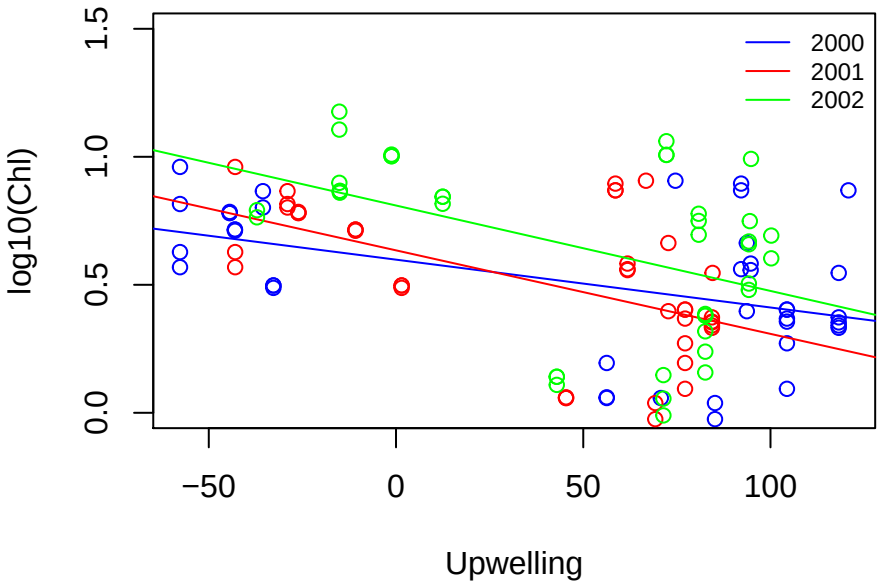
- **Null hypothesis:** There is no difference in mean annual chl by year and upwelling.
- **Alternative hypothesis:** There is a difference in mean annual chl by year and upwelling.

The results show that chl is significantly related to upwelling and that there is a difference in mean annual chl by year at the $\alpha = 0.05$ level. Because there is no significant interaction between year and upwelling (p-value = 0.998), the assumption of homogeneity of slopes is not violated.

Part 4

R-code

```
## Subset d2 by year
d2.2000 <- subset(d2, subset = year == 2000)
d2.2001 <- subset(d2, subset = year == 2001)
d2.2002 <- subset(d2, subset = year == 2002)

## Plot points for 2000
plot(d2.2000$upwelling, log10(d2.2001$chl), col = "blue", ylim = c(0, 1.5),
    xlab = "Upwelling", ylab = "log10(Chl)")
## Plot regression line for 2000
abline(lm(log10(d2.2001$chl) ~ d2.2000$upwelling), col = "blue")
## Plot points for 2001
points(d2.2001$upwelling, log10(d2.2001$chl), col = "red")
## Plot regression line for 2001
abline(lm(log10(d2.2001$chl) ~ d2.2001$upwelling), col = "red")
## Plot points for 2002
points(d2.2002$upwelling, log10(d2.2002$chl), col = "green")
## Plot regression line for 2002
abline(lm(log10(d2.2002$chl) ~ d2.2002$upwelling), col = "green")

## Add legend to plot
legend("topright", c("2000", "2001", "2002"), lty = 1, col = c("blue", "red",
    "green"), bty = "n", cex = 0.75)
```



Part 5

The lines in the figure support the results of the ANCOVA regarding interaction between upwelling and year. Both the figure and the ANCOVA suggest that chl is significantly related to upwelling and that there is a difference in mean annual chl by year.

The lines also support the results of the ANOVA that relates mean annual chl to year as the figure shows that for all values of upwelling, the transformed chl of the year 2002 is greater than that of 2000. Because the lines in the figure for 2000 and 2001 cross, the figure supports the finding that the transformed chl of the year 2000 is not statistically significant from that of 2001.

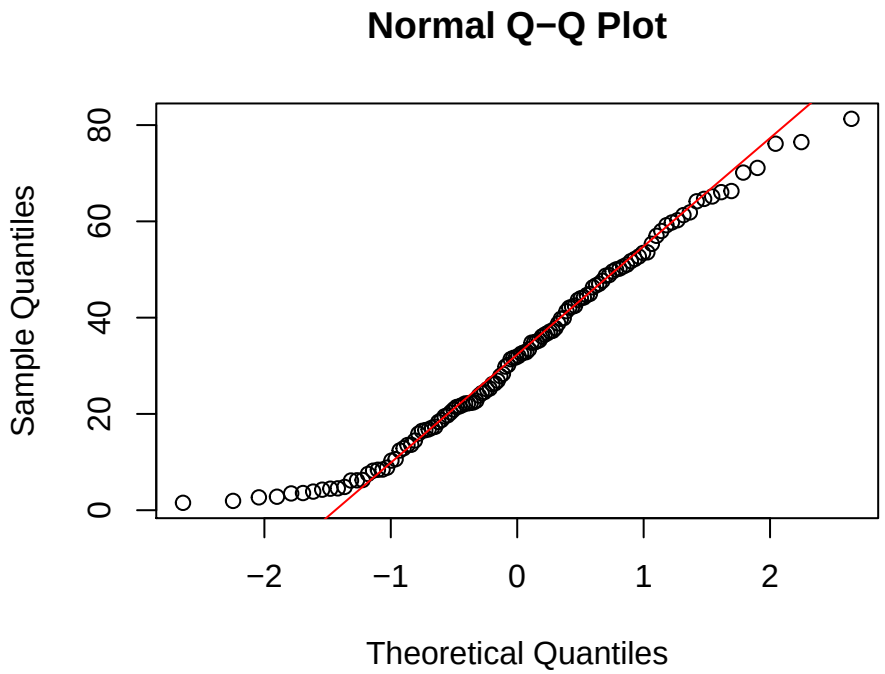
Part 6

The p-value associated with the effect of year greatly decreases (increases in significance) from the ANOVA and the ANCOVA. This is because in the ANOVA the year is treated as a continuous variable while in an ANCOVA the year was able to be treated as a discrete factor. The residual mean squares for the ANOVA was 0.095, while the residual mean squares for the ANCOVA was 0.06. Because in the ANOVA the year is treated as a continuous variable while in an ANCOVA the year was able to be treated as a discrete factor, the ANCOVA was able to explain much more of the variation in the data than the ANOVA was.

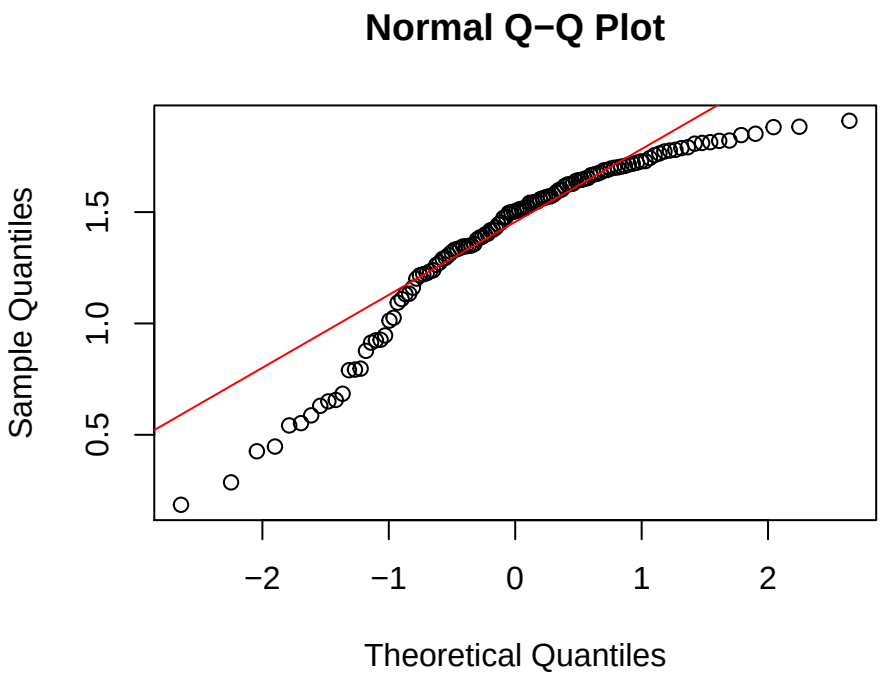
Part 7

R-code

```
## Use function qqnorm to generate a Q-Q plot of mussel abundance and then
## use qqline to plot the line in red
qqnorm(d2$mussel_cover)
qqline(d1$mussel_cover, col = "red")
```




```
## Use function qqnorm to generate a Q-Q plot of log10 transformed mussel
## abundance and then use qqline to plot the line in red
qqnorm(log10(d2$mussel_cover))
qqline(log10(d1$mussel_cover), col = "red")
```




```
## Non-transformed data is more normal

## Null hypothesis: there is no relationship between mussel abundance and
## upwelling

## Alternative hypothesis: there is a relationship between mussel abundance
## and upwelling

mod.lm.p2p7 <- lm(mussel_cover ~ upwelling, data = d2)
summary(mod.lm.p2p7)

##
## Call:
## lm(formula = mussel_cover ~ upwelling, data = d2)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -38.1  -15.1   -0.9   16.0   42.5
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)  36.3947     2.2245   16.36   <2e-16 ***
## upwelling    -0.0837     0.0319   -2.62   0.0098 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19.3 on 120 degrees of freedom
## Multiple R-squared:  0.0542,Adjusted R-squared:  0.0463
## F-statistic: 6.88 on 1 and 120 DF,  p-value: 0.00985
```

Interpretation

Because the data appear normal, mussel cover (the response variable) does not need to be transformed.

The hypotheses of the regression are as follows:

- **Null hypothesis:** There is no relationship between mussel abundance and upwelling.
- **Alternative hypothesis:** There is a relationship between mussel abundance and upwelling.

Although the p-value supports the hypothesis that there is a relationship between mean mussel abundance and upwelling, and the slope sign suggests that it is a negative relationship, the slope magnitude and R^2 value suggest otherwise. The slope magnitude is very small which suggests that there is almost no relationship between mean mussel abundance and upwelling. In addition, the R^2 value is very small which means that the fit explains very little of the variability of the response data around its mean (the regression is not a very strong fit). Therefore, the results of the regression do not support the hypothesis.