Auriel Fournier

Biometry Final

Spring 2014

May 6, 2014

I choose to complete questions 1,3 and 5.

**Question 1**

Background

There is interest in understanding the relationship of cholesterol levels in different states to see if further research into environmental variables is worthwhile.

Statement of Hypothesis

I hypothesize that there is a difference between the mean cholesterol levels of women in Iowa and Nebraska, regardless of their age.

Assumptions

Normality

I assume that the response variable and residuals have a normal distribution for every level of X. I examine this graphically through histograms before running the models, and if necessary transform the data. It can also be examined after examining the normal QQ plot and seeing if the points deviate from the line run the model.

Homogeneity of Variance

I assume that the response and residuals have an equal variance for each level of X. I can examine this by graphically examining the data through histograms and box plots before running the models and if necessary transform the data. Homogeneity of Variance can also be examined after the model is run by examining the residuals plot and seeing if the residuals are clumped or form wedges.

Independence

I assume that the response and residuals are independent of each other. That is that the response for a given value of x does not influence the response for another value of x. Good experimental design should take into account independence.

Linearity

I assume the relationship between the response and X is each group is linear. I can graphically examine this using scatter plots.

Covariate values similar across groups

I assume that each covariate has the same distribution and most importantly a similar range of values across each group. This means that there is no collinearity between the categorical and continuous predictor. I can examine this by running an ANOVA on the predictors against each other, and if they are significant I cannot continue with an ANCOVA.

Fixed Covariate (there is no error in X)

I assume that the covariate X was collected without error. In most biological settings I would have to ‘ignore’ this assumption, but if I can assume that all the women were honest about their age and state of origin then I can assume that this covariate is fixed. There is no way to test for this or graphically examine it.

Methods

I chose to perform an ANCOVA on the data so that I could look at the difference between the states, while also incorporating age as a predictor variable. ANCOVA generates two linear relationships, one for each level of the ‘state’ variable and compares their slopes to see if they are significantly different.

Results

Verification of Assumptions

Normality

I graphically examined the data with histograms and corrected cholesterol with a square root transformation to make it better fit a normal distribution.

Homogeneity of Variances

I graphically examined the data with box plots and saw that there was similar variance between each state

Independence

I assume that the data was collected in such a way that it is independent.

Linearity

I graphically examined the data in a scatter plot and saw that there was a generally positive trend between both states when comparing age and cholesterol level (Figure 1)

Covariates values similar across groups

The ANOVA on the two predictors versus each other (Table 1) was not significant, showing that there is no relationship between age and state and allowing us to preform an ANCOVA

Fixed Covariate

I am assuming that the women correctly reported their age and state.

Analysis/Results

First I graphically examined the data to make sure that it met all the assumptions (see above). Then I ran an ANOVA looking at cholesterol level and state, and found that it was not significant (Table 2). I graphically examined the residuals and found that there was slight clumping and the data was not perfectly normally distributed. With only n=30 this is not surprising and transformations will likely not be able to completely correct for this.

Then I ran another ANOVA looking at the predictor variables vs. each other (Table 1). This was also not significant, which showed that there is no relationship between age and state, meeting the subject of no collinearity and allowing me to continue with the ANCOVA

I ran the ANCOVA and found that only age was a significant predictor of cholesterol level, there was not a significant difference in state (Table 3). The R2 value was decent (0.525) which means this model explains just over half the variation in cholesterol.

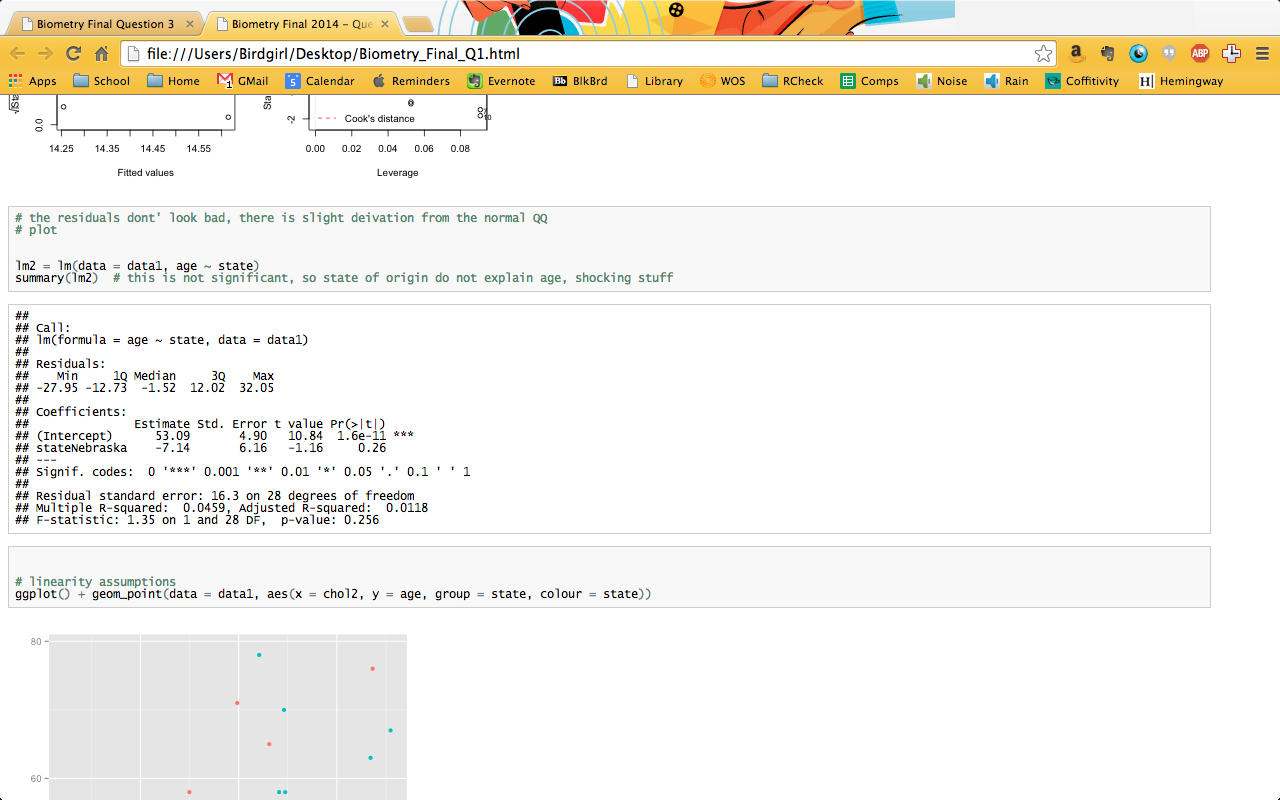
I ran another ANCOVA including the interaction between state and age to check for parallel slopes and found that the interaction was not significant (Table 4). This means that the slopes are parallel, though because state is not significant in Table 3, the difference between the two states is still not significant.

Discussion

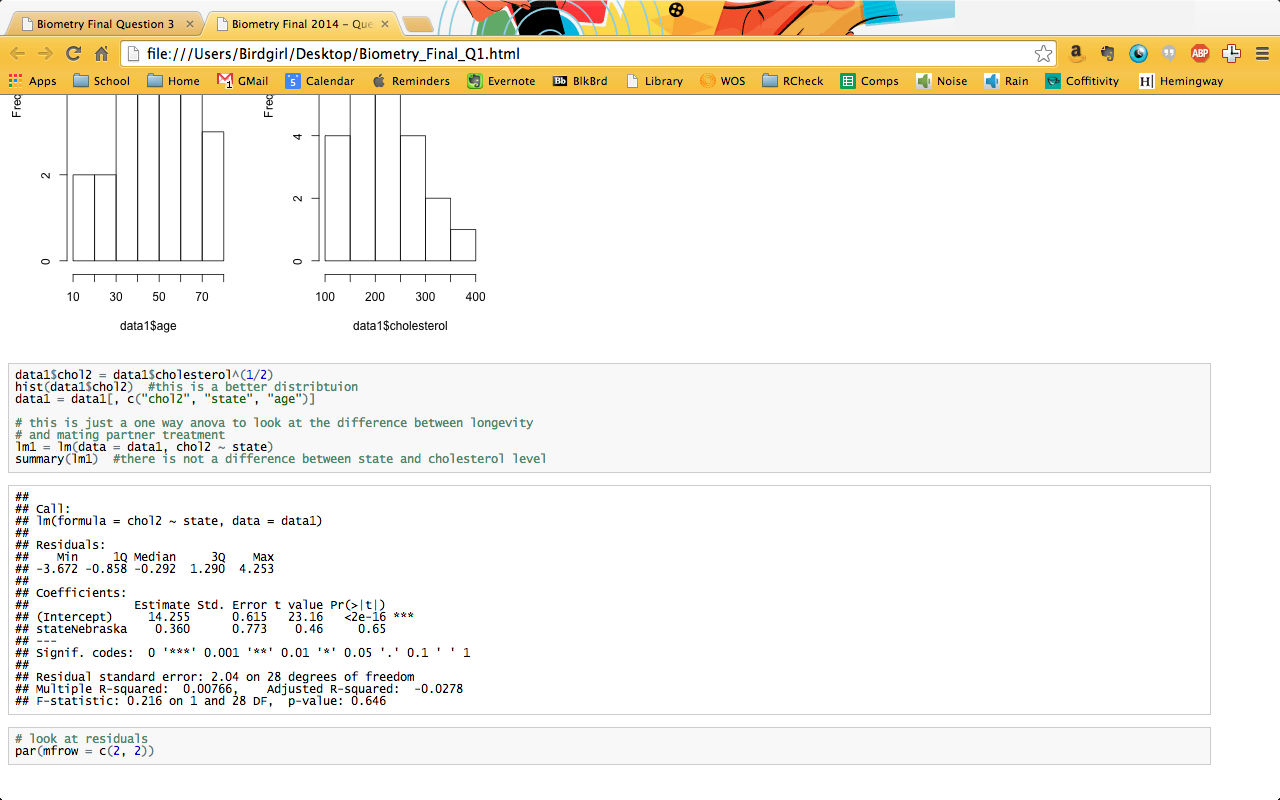
I fail to reject our null hypothesis of no difference. This means that running an ANCOVA did not improve our ability to explain the variation in cholesterol level, and I would be better of running a simple regression on age and cholesterol level, instead of an ANCOVA.

This suggests that there are not environmental factors, at least at the state level, that are influencing cholesterol levels in women. Examination of the effects of environmental variables at a smaller scale (county or city) might be worth investigating.

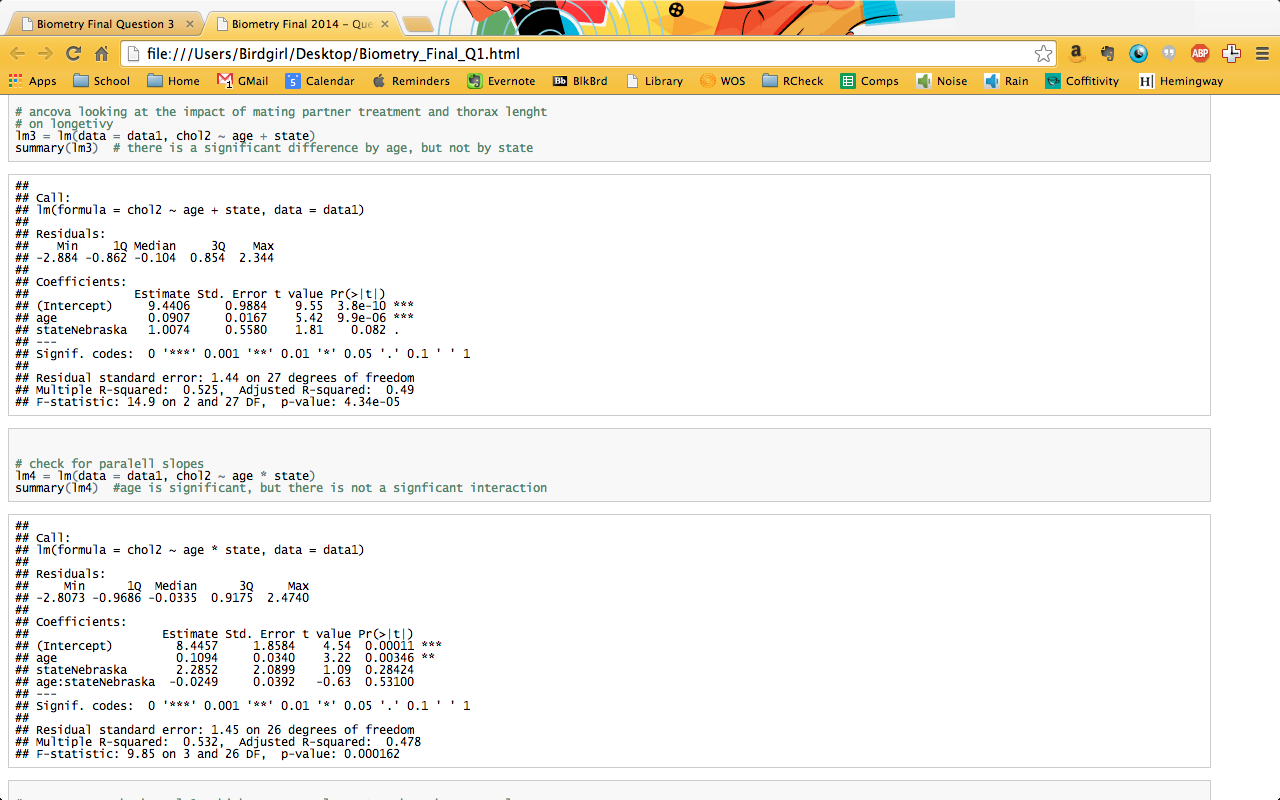
**Table 1 – ANOVA on predictor variables state and age**



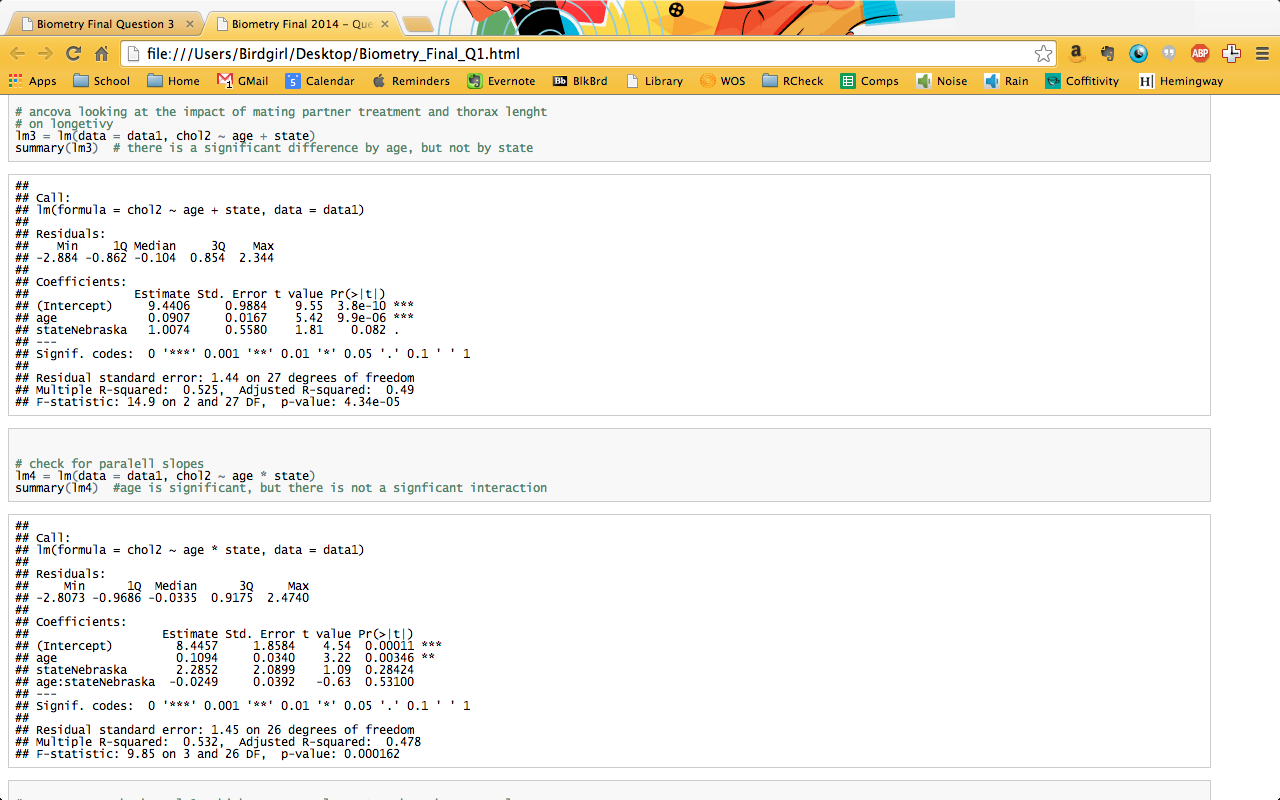
**Table 2 – ANOVA on age and state**

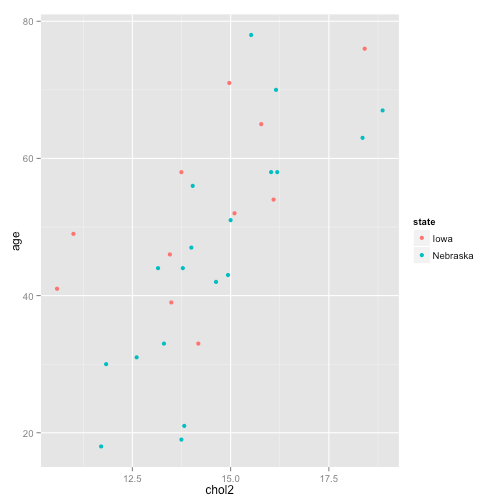


**Table 3 – ANCOVA on cholesterol level by state and age**



**Table 4 – ANCOVA on cholesterol level by state and age with interaction**





**Figure 1 – Scatter Plot of Cholesterol vs. Age with state delineated by color**

**Question 3**

Background

Fish communities are influenced by many factors, including biotic and abiotic factors. Today I will look at the influence of stream and season on fish density and richness in streams in the Boston Mountains.

Statement of Hypothesis

I hypothesize that there will be a significant difference in both density and richness based on stream and season in the Boston Mountains.

Assumptions

Normality

I assume that the response and residuals have a normal distribution for every level of X. I examine this graphically through histograms before running the models, and if necessary transform the data. Normality can also be examined after examining the normal QQ plot and seeing if the points deviate from the line run the model.

Homogeneity of Variance & Covariance

I assume that the response and residuals have an equal variance for each level of X. I can examine this by graphically examining the data through histograms and box plots before running the models and if necessary transform the data. It can also be examined after the model is run by examining the residuals plot and seeing if the residuals are clumped or form wedges.

Outliers

Outliers have a large impact on MANOVA, but so does unbalanced sample design. For the purposes of time for this exam I have decided to leave the outliers in the data set to ensure a balanced design, given more time I would test each outlier’s removal and look at it’s influence on the outcome of the MANOVA

Methods

I choose to perform a MANOVA on the data because we have two predictor variables which are both categorical (stream and season) and two continuous response variables (density and richness).

Results

Verification of Assumptions

Normality

I graphically examined the distribution of each variable using histograms. I found that density had a skewed distribution and that a 4th root transformation helped it meet this assumption.

Homogeneity of Variance and Covariance

I graphically examined the distribution of each variable’s variances through boxplots. I observed that there are a few outliers, but overall the range of variance or spread is similar for all covariates.

Outliers

As mentioned above because of time I have decided to leave the outliers in to ensure a more balanced sampling design. Given more time I would remove each individually and look at their impact on the MANOVA.

Analysis/Results

I first examined the assumptions of the MANOVA (see above).

To begin with I ran a set of four ANOVAs to look at the impact of Stream and Season individually on both richness and density.

Stream and density was significant for Falling Stream and had an R2 of .405. Stream and richness was significant for Falling Stream again and had an R2 of .2. Season and density was significant for August with an R2 of .06, showing that despite this significance season alone does not explain the variation well. Season and richness was significant for August and October, but the R2 was only .1, showing that season alone does not explain the variation in richness well. It is encouraging though that levels in each model are significant, this means that perhaps with a combination of the variables a better fit can be obtained.

Examination of residuals and normal QQ plots from these four ANOVAs showed that the data is not perfectly normally distributed and there is some clumping of residuals, suggesting differences in homogeneity of variance. Since I have already transformed the data I will proceed since I do not believe I can obtain a better transformation to deal with these assumption violations.

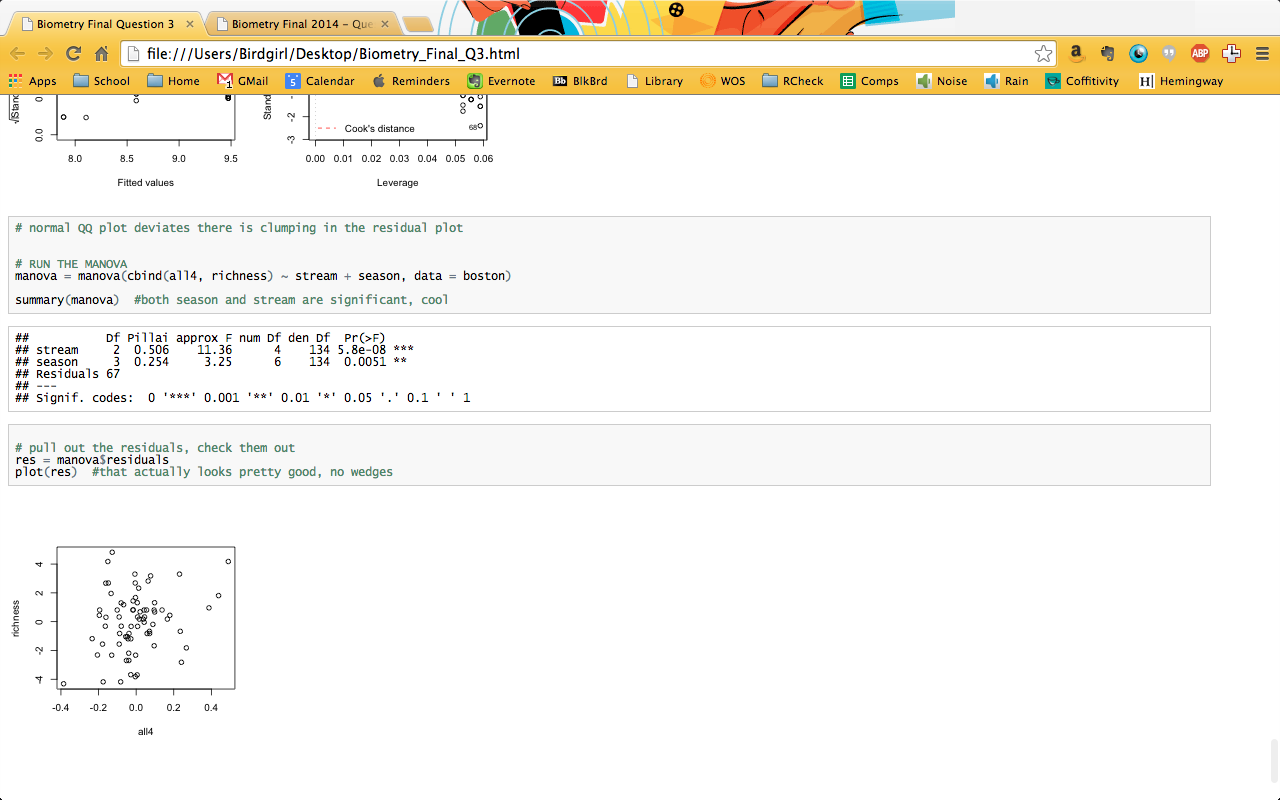
I then ran the MANOVA, looking at both stream and season and their impact the richness and density. Both stream and season were significant, showing that they both having an effect on richness and density (Table 5). The residuals were well distributed, having a shotgun like pattern (Figure 2)

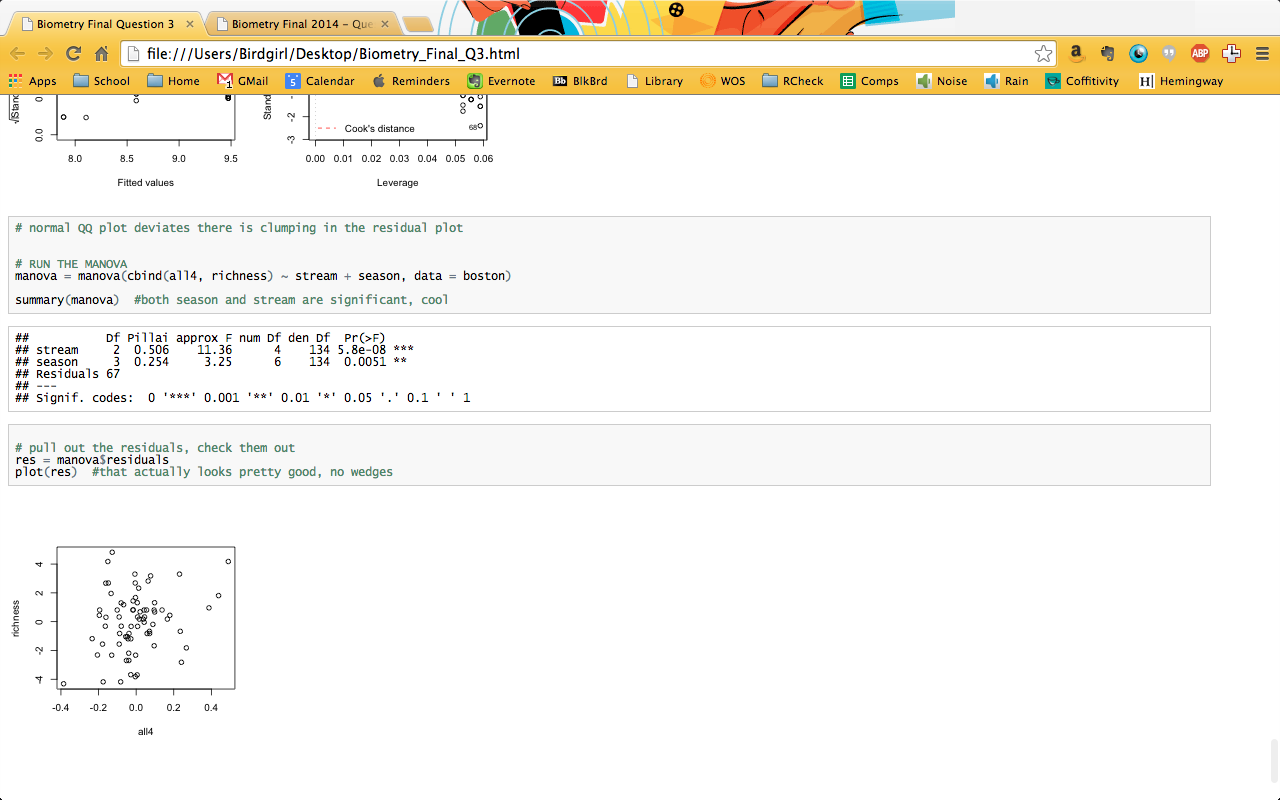
I then wanted to look at the individual predictors on each of the response variables with a univariate F test, but the manova.summary() function in R was not working correctly and per the restrictions of this exam I was not allowed onto the stackoverflow or other online forums to try and determine why this is. So I cannot comment further on the specific influence of different predictors on the responses.

Discussion

I reject the null hypothesis of no difference. The MANOVA did explain the variation of richness and density in the Boston Mountains, but I cannot say with what strength of how those relationships vary. Based on this significance additional work could be done to look at how they are impacting these fish communities.

**Table 5 – results of MANOVA on richness and density of fish communities in the Boston Mountains based on stream and season.**





**Figure 2 – Residuals from the MANOVA on richness and density of fish communities in the Boston Mountains based on stream and season.**

**Question 5**

Randomization is imperative to remove the bias of the scientist(s) performing the experiment from the data. By randomly sampling from a system with sufficient sample size you are able to obtain an accurate picture of the actual processes impact your response of interest, without any influence of the observer’s preconceived notions.

Ideally randomization would also ensure interspersion, which is the mixing together of different treatments on the ‘landscape’ (be that a true landscape in a field experiment, or the landscape of a laboratory). This interspersion prevents the impact of variables of non-interest on our response. For example, the impact of light on fish growth rates in a green house with uneven lighting conditions.

Interspersion can be evaluated quantitatively through various metrics, such as the interspersion equation used in the program FRAGSTATS. (McGarigal et al. 2012). If interspersion and randomization are of concern a certain level of interspersion (say 40% could be determined *a priori*) and random sampling designs could be generated until this metric was met. It is important to have an *a priori* measure of interspersion before hand though, or observers may bias the randomized design.

In field experiments I would argue that randomization is more important then interspersion, since often the environmental gradients that occur across an landscape are variables of interest, and experimental units should be stratified across those gradients, not just interspersed.

In laboratory experiments I think that interspersion is more important then randomization since the environmental impacts of things such as variation in light and temperature are often not variable of interest, and by interspersing the experimental units at some *a priori* level would allow you to ignore those impacts on your results.

Regardless of the experiment you are performing both randomization and interspersion should be considered in the planning stages and included in your experimental design with conscious *a priori* decisions to ensure statistical rigor.

**Citation**

McGarigal, K., Cushman, S. A., & E., E. (2012). FRAGSTATS v4: Spatial Pattern Analysis Program for Categorical and Continuous Maps. Retrieved from http://www.umass.edu/landeco/research/fragstats/fragstats.html