**Assignment 2**

Complete the questions and record the answers as you go. When you are confident, input the answers using the quiz in LMS. Remember that plagiarism is considered a serious offence at UWA – you are welcome to discuss and compare your work with other people but you must write your own script and do all the questions yourself.

***Question 1: George’s Cockroaches***

George is obsessed with cockroach racing. He wants to win the annual cockroach races at the Downunder Tavern in Kangaroo Point more than anything else in the whole world. He wonders if giving his roaches coffee or red bull, or training them with loud music might improve their performances. He randomly selects 60 roaches from under the kitchen floorboards, randomly divides them into two groups of 30, and trains one group with loud music and one group without. He then randomly divides each group of 30 into three subgroups of 10, gives each subgroup a different treatment (red bull or coffee, and one group is the control group – they get nothing extra), and then times their performance on the standard Queensland cockroach racing track. But while he has a special connection with his roaches, statistics is not his thing, so he has sent the data to you to analyse.

What are the two factors? How many levels in each factor? What kind of test should we use to analyse this data?

George’s data is in the ‘georges\_data.xls’ excel file downloadable from LMS. It shows the time (in seconds) taken for each cockroach to complete the race track. Read in the data and create a boxplot of the data. Analyse the data for George making sure you do the right analysis. Which effects are significant? If necessary, use a Tukey Honest Significant Difference Test (TukeyHSD in R) in conjunction with boxplots or other plots to determine which levels are different (greater or less) from which other levels within any factor that is significant. What can we conclude?

***Question 2: Understanding the ANOVA tables***

In a simple linear regression on 21 data points I get the following ANOVA table:

 Df Sum Sq Mean Sq F value Pr(>F)

x # 179.72 # # #

Residuals # # 20.531

Fill in all the blanks marked with #.

In a one-way ANOVA with three treatments, and six replicates in each treatment, I get the following ANOVA table:

 Df Sum Sq Mean Sq F value Pr(>F)

tr # # # 5.9615 #

Residuals # 39 #

Fill in all the blanks marked with #.

***Question 3: Gaby’s Seagrass Biomass***

(Note, some parts of this question are meant to be fairly challenging, so don’t panic if you can’t get it all perfectly right! Just do your best with the time you have available.)

Gaby loves seagrass, so she decides to do her PhD on it. She wants to know whether a desalination plant that has been operational for about one year is having an effect on her beloved seagrass. She selects a number of sites around the desalination plant outlet at varying distances as shown in this figure:

 The open black circle is the brine outlet, the 10 sites represented by the black dots are in a ring 50m from the outlet, the 10 sites represented by the red dots are in a ring 500m from the outlet and the 10 sites represented by the blue dots are in a ring 1000m from the outlet. The location of the sites around the ring is chosen at random, while ensuring that the sites are within areas previously reported as being good quality seagrass meadows. She carries out a relative biomass estimate at each site, based on the density of the visible leaves, and also records the depth below sea level of each site. The data she collects is in the ‘gabys data.xls’ file.

1. First, use a one-way ANOVA (not a regression) to look at whether there is a relationship between the distance of the sites from the outlet, and their depth ie. is there a general trend in topology that could confound our results. Is there a significant relationship? Record the p-value you get for this test. If there is a significant relationship, use a boxplot and a TukeyHSD test to decide where the deepest sites are located (at which distance(s)).
2. Use a one-way ANOVA (not a regression) to look at whether there is a relationship between the distance of the sites from the outlet, and their relative biomass. Is there a significant relationship? Record the p-value you get for this test. If there is a significant relationship, use a boxplot and a TukeyHSD test to decide where the sites with greatest biomass are located (at which distance(s)).
3. Use a linear regression analysis to look at whether there is a relationship between the distance of the sites from the outlet, and their relative biomass. Is there a significant relationship? Record the ***F-value*** you get for this test. If there is a significant relationship, does biomass increase or decrease with greater distance?
4. Is there a difference in the conclusions to be drawn from the ANOVA test and the regression test on the effect of distance on biomass?
5. Use a linear regression analysis to look at whether there is a relationship between the depth of the sites, and their relative biomass. Is there a significant relationship? Record the ***F-value*** you get for this test. If there is a significant relationship, does biomass increase or decrease with greater depth?
6. Now, use an ANCOVA analysis with an interaction looking at the effects of both distance (as a factor) and depth (as a covariate). Since depth is the covariate and not the main thing we are interested in, we normally put this as the first explanatory variable in the linear model, so its effect is taken out first. Make sure you do it that way this time. Which of the main effects and interactions are significant (at α=0.05)?
7. Now try fitting the same linear model but with distance as the first explanatory variable. Notice that it makes a difference to the results! Which effect was significant in one of these models but not in the second? And the p-value of which effect did not change at all between the models?
8. Now try fitting a linear model with depth as the first explanatory variable and distance as the second, and an interaction, but with both of the variables as covariates, not factors (a multiple regression). This is possible because both variables are numeric. Notice that it also makes a difference to the results, with all the p-values being lower now, compared to the ANCOVA analysis with depth first! Why is the p-value for depth smaller?\* Why is the p-value for distance smaller?\* Why is the p-value for the interaction smaller?\*

\*You need to explain these changes in p-values in terms of changes in sums of squares and/or degrees of freedom for the factor and/or the residual. The options in the quiz will be:

* higher explained MSS because of higher explained SS
* higher explained MSS because of lower degrees of freedom
* lower residual MSS
* higher explained MSS because of higher explained SS, and also lower residual MSS
* higher explained MSS because of lower degrees of freedom, and also lower residual MSS
* higher explained MSS because of higher explained SS and lower degrees of freedom, and also lower residual MSS

(MSS means mean sum of squares and SS means sum of squares)

1. Does the result we got in part e) for the (non) significance of depth contradict the result we get here? Why/ why not?
2. Since the last model has a smaller residual mean squares error, we can conclude that it is the best model (there are actually many ways to decide which is the best model, but this is one possible way). Given the significant interaction in this model, and carefully considering the data, what conclusion could we draw regarding the effects of salinity and depth on seagrass biomass (assuming that greater distance from the outlet does indeed mean less effect of salinity)?

Hint: Drawing some plots for the model predictions like we did in the guided ANCOVA example for Day 2 would really help for drawing conclusions. Another way to get some helpful plots would be to divide the data into three subsets according to distance, and then fit and plot a simple linear model against depth for each subset. Or you could also create a ‘depth factor’ with two or three levels (based on the depth covariate) and then fit a two-way ANOVA and make boxplots and/or interaction plots as necessary. Only one of these options should be necessary, and there may be other ways to plot or table the data that will be helpful too. Make sure you have a clear plot or table like this with you when you come to answer the assignment questions online.

***Question 4: Stream Water Quality***

This question really depends on carefully investigating and understanding the data, rather than just routinely applying statistical tests. You need to understand the difference between the output from the R ‘summary’ and ‘anova’ functions. Make sure you’ve worked through the icecream/vegan lab script and read and understood the ‘extra explanation: Regression and ANOVA” document provided with the Day 2 lectures.

Consultants have collected water quality monitoring data on a number of the streams that flow into a river. For each stream they have calculated a water quality index, where a higher value means the water in the stream is better quality. They also recorded the average flow rate of the stream (using GIS software) and whether the stream flows into the river from the north side or the south side. The data is available in the 'steamswaterquality.csv' file. They think that flow rate may be related to water quality, and want to test this hypothesis statistically.

Fit a simple linear regression of water quality predicted by flow rate. Look at the anova output for this fitted model.

Does this simple test support their hypothesis?

Now look at the summary output.

Do the anova and summary outputs agree?

Now fit a model with both flow rate and direction as possible predictors. Include an interaction. Make sure the continuous covariate is the first term in the model.

Look at the anova output for this fitted model.

Is there evidence that the effect of flow rate on water quality is different for streams coming from the north versus those flowing from the south? Is there evidence that flow rate affects water quality? Is there evidence that direction affects water quality?

Now look at the summary output.

Is there evidence that the effect of flow rate on water quality is different for streams coming from the north versus those flowing from the south?

Is there evidence that flow rate affects water quality?

Is there evidence that direction affects water quality?

Does the significance of the three model terms (flow rate, direction and the interaction) in the summary output agree with their significance in the anova output?

+++

Now fit another model with both flow rate and direction as possible predictors, but this time in the other order ie, make sure the continuous covariate is the second term in the model. Include an interaction again.

Look at the anova output for this fitted model.

Is there evidence that the effect of flow rate on water quality is different for streams coming from the north versus those flowing from the south?

Is there evidence that flow rate affects water quality?

Is there evidence that direction affects water quality?

Now look at the summary output.

Is there evidence that the effect of flow rate on water quality is different for streams coming from the north versus those flowing from the south?

Is there evidence that flow rate affects water quality?

Is there evidence that direction affects water quality?

Which of the three model terms (flow rate, direction and the interaction) has the same p-value in the summary output as in the anova output?

Why?

Has changing the order of the model terms made a difference to the p-values of the three terms in the anova output?

Has changing the order of the model terms made a difference to the p-values of the three terms in the summary output?

It seems something strange is going on here. P-values and significance seem to depend quite a lot in some cases on whether we look at the summary or anova output, and on the order in which we fit the model terms. To try to get some idea of why, let’s go back to the data itself. Plot the data in a way (or ways) that illustrates the relationships between flow rate, direction of flow, and quality index. Based on the plot(s) and the tests above, and any other tests you’d like to run, which of the following could you confidently conclude from this data (and in general)?

* Stream flow rate significantly affects water quality
* Water quality is significantly higher in streams flowing into this river from the south than in those flowing from the north
* The way that flow rate affects water quality depends on which direction the stream flows
* Flow rate is significantly higher in streams flowing into this river from the south than in those flowing from the north
* Flow rate and direction of stream flow are quite independent in this data set
* The relationship between flow rate and direction of stream flow in this data set makes it hard to be sure what is really affecting water quality
* The p-values obtained for each model term/factor via the ‘anova’ output on a fitted model (ie via f-tests on adding each additional model factor in turn) can be quite different to those obtained via the ‘summary’ output (ie via marginal t-tests) on the same fitted model
* The way that model terms are ordered in a fitted model can make a big difference to the p-values obtained via the ‘summary’ output on the fitted model
* The way that model terms are ordered in a fitted model can make a big difference to the p-values obtained via the ‘anova’ output on the fitted model