***Exercise 5: General Linear Models***

***(i.e. linear models with >1 explanatory variable)***

**i)** Let’s start off with a fairly simple model where we have one continuous explanatory variable and one categorical explanatory variable. As we have spent a lot of time looking at soil pH, let’s start with that and how it varies with total base saturation and habitat.

As in our last session, we will use the lm() function to build these models. Note the notation here, we use a ‘+’ sign to separate the different explanatory variables.

**> lm\_pH\_Habitat\_TBS <- lm(Soil\_pH~Habitat + Total\_Base\_Saturation,data=soils)**

You can explore this model using the anova(), summary and plot() function as per the last quantitative skills session. You will see using the plot function that our model does not fit the data that well. In the model, we did not consider interaction effects. We have created a model where soil pH may vary with habitat and with total base saturation, but we are stating that the relationship between soil pH and total base saturation is the same in both habitats. Thus, for every unit increase in total base saturation, there is a similar increase in soil pH, whether or not we are in floodplain or upland environments. However, it may sensible to check if the relationship between soil pH and total base saturation varies in the two habitats. To do so, there are two possible notations.

**> lm\_pH\_Habitat\_TBS\_Interaction <- lm(Soil\_pH~Habitat + Total\_Base\_Saturation + Habitat:Total\_Base\_Saturation,data=soils)**

We have now added an interaction term for the two explanatory variables, by including them and separating them with a ‘:’. Another, quicker way to do this in R is as follows.

**> lm\_pH\_Habitat\_TBS\_Interaction <-lm(Soil\_pH ~ Habitat\*Total\_Base\_Saturation, data=soils)**

**> anova(lm\_pH\_Habitat\_TBS\_Interaction)**

**> summary(lm\_pH\_Habitat\_TBS\_Interaction)**

**> plot(lm\_pH\_Habitat\_TBS\_Interaction)**

In assessing this model, we see that we obtain a slightly better model fit (less heteroscedasticity), but the interaction term is not actually significant. Thus, a more parsimonious model might be to exclude the interaction term.

How do we compare different statistical models, which incorporate different explanatory variables, to decide which model is ‘best’? One sensible approach is to build a model that incorporates all explanatory variables that we think are important to explaining the response variable. If some are insignificant, they will not have much ‘impact’ on the model.

Another common approach is to simply include whichever variables seems to be significant in univariate models using an ANOVA framework or t-tests on the associated parameters (to determine if they are significantly different from zero). While this is not really the approach I recommend, building univariate models and plotting out relationships between pairs of variables is one of the most straightforward ways to come to understand your data. The more time you spend with a given dataset, the better will be your analysis.

Yet another approach is to build the most complex model possible and then to exclude non-significant variables. However, in large datasets sometimes a variable (or parameter) will be ‘significant’, but it actually does not increase the explanatory value of the overall model too much. With linear models or general linear models, the r squared value (or better, the adjusted R squared value) can give you an idea of ‘model fit’, or how well the explanatory variables explain the response variable. One can remove variables from this complex model one at a time to determine which removals cause the r squared value to drop a lot. Those variables should stay in the model. If the r squared does not drop at all when a variable is removed, then there is no reason not to remove it.

One note here: In this dataset, we have lots of variables, but only 25 sample units, or data points (here, rows in the data.frame). If we build models with lots of explanatory variables, we may have problems with ‘overfitting’. This is a common problem in statistics, and one I leave you to read about on your own (Wikipedia has an informative page for example). You may wonder how many explanatory variables is too many? It really depends on what your data look like. Still, there is a useful rule of thumb that you can use as a bit of a guide (i.e. not as a hard and fast rule): you want to have 10 times as many data points as parameters you are trying to estimate (not including the intercept). In this data set, you have 25 data points, so it generally wouldn’t be a good idea to try and fit models with more than 2 or 3 parameters. In the following model comparison exercise, see if you can figure out how many parameters you are estimating.

Here is an example of model comparison for the above question. First we need to create the simpler models.

**> lm\_pH\_Habitat <- lm(Soil\_pH~Habitat,data=soils)**

**> lm\_pH\_TBS <- lm(Soil\_pH~Total\_Base\_Saturation,data=soils)**

Now, let’s compare all 4 models

**> summary(lm\_pH\_Habitat)**

**> summary(lm\_pH\_TBS)**

**> summary(lm\_pH\_Habitat\_TBS)**

**> summary(lm\_pH\_Habitat\_TBS\_Interaction)**

In comparing the r squared values of these models, we see that total base saturation explains more of the variation in soil pH values than habitat. However, a model that incorporates total base saturation and habitat is marginally better. Meanwhile, adding the interaction term doesn’t really increase the explanatory value of our model at all, yet it adds an extra parameter.

**ii)** Let’s now look at a model with two categorical explanatory variables, with and without interactions. Among the categorical variables in our dataset, we have river basin and habitat type. We should first examine the distribution of sample points across these two categorical variables.

**> summary(soils$River\_Basin:soils$Habitat)**

We see that we do not have any upland sites sampled in Madre de Dios, so it may not be useful to include the Madre de Dios basin in any analyses of the interaction between basin and habitat type. Further, we see that there is only 1 data point in floodplain and upland in both the Tambopata and Las Piedras river basins. When one has two categorical variables (and their interaction) in a statistical model, one is essentially estimating a parameter for each group combination. In this case, it means that we are estimating a parameter for each habitat type in each river basin. In order to conduct statistical tests, we should be able to approximate our uncertainty for each parameter estimate, which requires that we have multiple samples per parameter we intend to estimate. E.g., in an anova framework, one cannot compare among to within group variation if one is not able to quantify within group variation. In summary, we would ideally not include groups (or group combinations in this multivariate analysis) with only a single data point. Thus, we should exclude Tambopata and Las Piedras from analyses if we are going to assess the effect of the interaction between habitat type and river basin on soils. Here is a somewhat complicated line of code to do this.

**> soils\_trim <- soils[soils$River\_Basin=="Manu"|soils$River\_Basin=="Los\_Amigos",]**

Here is an alternative:

**> soils\_trim <- soils[soils$River\_Basin %in% c("Manu","Los\_Amigos"),]**

There are many ways to subset data in R. Many R practitioners enter the tidyverse and use add-on packages in R such as dplyr to manage datasets. We will not cover that in this course, but mention it here should you wish to study up on it.

Let’s build some models looking at phosphorous

**> lm\_P\_habitat <- lm(Phosphorus~ Habitat,data= soils\_trim)**

**> lm\_P\_basin <- lm(Phosphorus~ River\_Basin,data= soils\_trim)**

**> lm\_P\_habitat\_basin <- lm(Phosphorus~ Habitat+River\_Basin, data=soils\_trim)**

**> lm\_P\_habitat\_basin\_interaction <- lm(Phosphorus~ Habitat\*River\_Basin, data=soils\_trim)**

And now let’s compare those models

**> summary(lm\_P\_habitat)**

**> summary(lm\_P\_basin)**

**> summary(lm\_P\_habitat\_basin)**

**> summary(lm\_P\_habitat\_basin\_interaction)**

So, it looks like the model with the interaction term is the best model. However, we cannot forget to evaluate the models using the plot function or otherwise. There are some MAJOR non-normality and heteroscedasticity issues here. Thus, we cannot really trust these results. How might you solve these issues? One would have to re-run these various models once a ‘fix’ has been tried (e.g. transforming the response variable). Also, once you come up with your final model, think about what the results actually mean. What does it mean if there is a significant interaction between river basin and habitat type? Also, in terms of our discussion above about over-fitting, we are approaching ‘danger territory’ here in trying to estimate so many parameters with only 19 data points (now that we have excluded some of the river basins).