Assignment 5

Melissa Pulley

Lab 5. Multiple linear models and model selection

Now that we are done with simple and multiple linear models it's time to analyze data in R.

Update either an R file or an RMarkdown file with your answers

Interactive effects

First off, let's run an interactive effect model. Using the **iris** dataset, run a model in which Sepal Length is the response variable (call it model 1), and the explanatory variables are Sepal Width and Species. Make sure you are including an interaction term!

Q1. 3 pts. Run the summary of your model, and interpret it. Look at the F-statistic, and the model p-value. (null hypothesis is, this model explains the data just as well as a null model). Is it significantly better than a null model? Look at each coefficient and whether they are significan. Are they significant? Would you say this is a good model?

```
head(iris)
```

```
  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1          5.1         3.5          1.4         0.2  setosa
2          4.9         3.0          1.4         0.2  setosa
3          4.7         3.2          1.3         0.2  setosa
4          4.6         3.1          1.5         0.2  setosa
5          5.0         3.6          1.4         0.2  setosa
6          5.4         3.9          1.7         0.4  setosa
```

```r
model1 = lm(Sepal.Length~Sepal.Width + Species+ Sepal.Width:Species, data=iris)
summary(model1)
```

```
Call:
lm(formula = Sepal.Length ~ Sepal.Width + Species + Sepal.Width:Species,
    data = iris)

Residuals:
     Min       1Q   Median       3Q      Max
-1.26067 -0.25861 -0.03305  0.18929  1.44917

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                    2.6390     0.5715   4.618 8.53e-06 ***
Sepal.Width                    0.6905     0.1657   4.166 5.31e-05 ***
Speciesversicolor              0.9007     0.7988   1.128    0.261
Speciesvirginica               1.2678     0.8162   1.553    0.123
Sepal.Width:Speciesversicolor  0.1746     0.2599   0.672    0.503
Sepal.Width:Speciesvirginica   0.2110     0.2558   0.825    0.411
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4397 on 144 degrees of freedom
Multiple R-squared:  0.7274,    Adjusted R-squared:  0.718
F-statistic: 76.87 on 5 and 144 DF,  p-value: < 2.2e-16
```

The null hypothesis is that the null model better represents the data than the proposed model. Since our p-value is very small, so have significant support for our proposed model. The p-values for the intercept and sepal width coefficents are below 0.05, meaning they have support to be non-zero. However, the coefficient for the other coefficients is above 0.05, meaning that we cannot discount that these coefficients may be zero, so this model could be better.

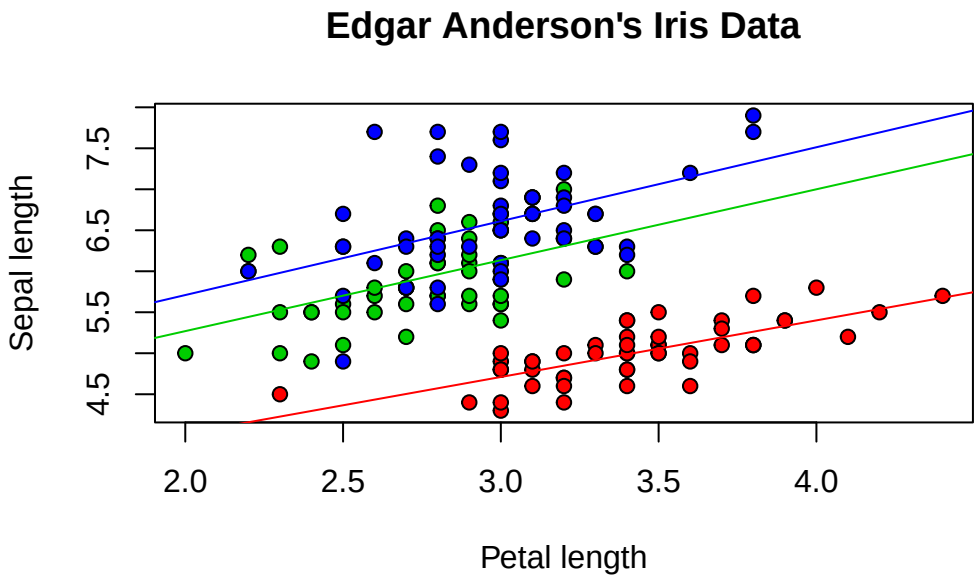
Visualizing data is a great way to understand what's going on. We are going to do the following:

1. Plot all the datapoints for Sepal Width~Petal Length and have each species as a different color.
2. We will estimate the lm for each species indiviually and plot the line

3. Look at the plot! :)

```
plot(iris$Sepal.Width, iris$Sepal.Length, pch=21, bg=c("red","green3","blue")[unclass(iris$Sp

abline(lm(Sepal.Length ~ Sepal.Width, data=iris[which(iris$Species=="setosa"),])$coefficients
abline(lm(Sepal.Length ~ Sepal.Width, data=iris[which(iris$Species=="versicolor"),])$coeffici
abline(lm(Sepal.Length ~ Sepal.Width, data=iris[which(iris$Species=="virginica"),])$coefficie
```



Look at your plot. Does it seem to have an interaction term? Do you think the last model was the most appropriate?

Let's run the same model again, but without the interaction term

Q2. 3 pts. Run the summary of your model, and interpret it. Look at the F-statistic, and the model p-value. (null hypothesis is, this model explains the data just as well as a null model?). Is it significantly better than a null model? Look at each coefficient and whether they are significant. Are they significant? Would you say this is a good model? Would you say it's better than the las one?

A2. In this new model, we also have a p-value significantly below 0.05, which gives significant support for this proposed model over the null model. Furthermore, all of the coefficients have p-values below 0.05, supporting that they are all non-zero. I would say that this is a good model and that it is better than model 1 since all of the coefficients have significant p-values.

```
model2 = lm(Sepal.Length~Sepal.Width + Species, data=iris)
summary(model2)
```

```
Call:
lm(formula = Sepal.Length ~ Sepal.Width + Species, data = iris)

Residuals:
     Min       1Q   Median       3Q      Max
-1.30711 -0.25713 -0.05325  0.19542  1.41253

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)         2.2514     0.3698   6.089 9.57e-09 ***
Sepal.Width         0.8036     0.1063   7.557 4.19e-12 ***
Speciesversicolor   1.4587     0.1121  13.012  < 2e-16 ***
Speciesvirginica    1.9468     0.1000  19.465  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.438 on 146 degrees of freedom
Multiple R-squared:  0.7259,    Adjusted R-squared:  0.7203
F-statistic: 128.9 on 3 and 146 DF,  p-value: < 2.2e-16
```

Now, hopefully you've been on top of your game, you have submitted your lab 4, and the study guide, and you're actually working on this during Friday lab (or during the weekend!). And you are asking yourself, how can we tell if this is a better model? We haven't talked about it in class!

AIC to the rescue!

We will talk about this on Monday (or, already talked about this, depending on when you decided to work on this). We can use AIC (Akaike Information Criterion) to test and select best models. So far, we know (hopefully) that both of the previous models were *significant*, which essentially compares those models to a null model, with H_0 being no difference in the

way those models explain the data (and variance), and H_1 being, our model is significantly better at explaining the data.

You might also be asking yourself, what is a **null** model anyway? Essentially a null model is a no-effect or no-differences model (that's why it represents the null hypothesis!) A null model only has one coefficients (and intercept) and you can think of that intercept as a mean for the whole data. Essentially it means there is no effect of Sepal Width or Species on Sepal Length. And you can run it using the following:

```
model3<-lm(Sepal.Length ~ 1, data=iris)
summary(model3)
```

```
Call:
lm(formula = Sepal.Length ~ 1, data = iris)

Residuals:
     Min       1Q   Median       3Q      Max
-1.54333 -0.74333 -0.04333  0.55667  2.05667

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.84333    0.06761   86.42   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8281 on 149 degrees of freedom
```

As you can see, it gives you a P-value for the one coefficient. But it gives you no overall F statistic or P-value for the model overall (this makes sense!). Essentially, it's telling you the mean length is 5.84 and that is independent of Sepal Width or species.

Back to AIC, we can use it to select models. We will talk more about it during class. For now, just trust me, that a **lower AIC represents a better model**. So, to test models, you should run AIC:

```
AIC(model1,model2,model3)
```

```
       df      AIC
model1  7 187.0922
model2  5 183.9366
model3  2 372.0795
```

Q3. 1 pts. Based on the AIC, which is the best model?

A3. Because it has the lowest AIC, model 2 is the best model from these options.

Green Frogs

We will use data from the following publication by M. Mazerolle (2006): *Improving data analysis in herpetology: using Akaike's Information Criterion (AIC) to assess the strength of biological hypotheses*. The data feature mass lost by green frogs (*Lithobates clamitans*) after spending two hours on one of three substrates that are encountered in some landscape types. The response variable is the mass lost (Mass_lost) and we are interested in testing difference among substrate types.

To make things simpler, we will only use main effects. No interactions (things can get very complicated, very fast).

In order to analyze this dataset, you will need to download and load the "AICcmodavg" package. And you will need to load the package. Once you have loaded the package, you can load the dataset. We will only use the first 7 columns of the dataset.

```
Warning: package 'AICcmodavg' was built under R version 4.3.3
```

```
data(dry.frog)
frog<-dry.frog[,1:7]
```

And let's explore the dataset!

```
head(frog)
```

```
  Individual Species Shade  SVL Substrate Initial_mass Mass_lost
1          1   Racla     0 7.27      SOIL         38.5       8.3
2          2   Racla     0 7.00  SPHAGNUM         31.0       3.6
3          3   Racla     0 6.83      PEAT         23.6       4.7
4          4   Racla     0 7.26      PEAT         37.4       7.0
5          5   Racla     0 7.43      SOIL         44.4       7.7
6          6   Racla     0 5.75  SPHAGNUM         16.4       1.6
```

```
str(frog)
```

```
'data.frame':   121 obs. of  7 variables:
 $ Individual  : int  1 2 3 4 5 6 7 8 9 10 ...
 $ Species     : Factor w/ 1 level "Racla": 1 1 1 1 1 1 1 1 1 1 ...
 $ Shade       : int  0 0 0 0 0 0 0 0 0 0 ...
 $ SVL         : num  7.27 7 6.83 7.26 7.43 5.75 7.66 6.42 7.64 6.57 ...
 $ Substrate   : Factor w/ 3 levels "PEAT","SOIL",..: 2 3 1 1 2 3 1 2 3 2 ...
 $ Initial_mass: num  38.5 31 23.6 37.4 44.4 16.4 39.8 25.9 35.6 29 ...
 $ Mass_lost   : num  8.3 3.6 4.7 7 7.7 1.6 6.4 5.9 2.8 3.4 ...
```

```
summary(frog)
```

```
   Individual     Species        Shade             SVL           Substrate
 Min.   : 1.00   Racla:121   Min.   :0.0000   Min.   :3.080   PEAT    :39
 1st Qu.:16.00               1st Qu.:0.0000   1st Qu.:4.060   SOIL    :41
 Median :31.00               Median :1.0000   Median :4.640   SPHAGNUM:41
 Mean   :31.69               Mean   :0.5207   Mean   :5.383
 3rd Qu.:48.00               3rd Qu.:1.0000   3rd Qu.:7.080
 Max.   :63.00               Max.   :1.0000   Max.   :8.810
  Initial_mass     Mass_lost
 Min.   : 2.90   Min.   :0.000
 1st Qu.: 5.90   1st Qu.:0.500
 Median : 8.60   Median :1.000
 Mean   :18.14   Mean   :1.632
 3rd Qu.:30.80   3rd Qu.:2.300
 Max.   :66.30   Max.   :8.300
```

We are interested in running 8 models and compare them. Each model represents a different biological hypothesis, and we can test them all. Isn't this cool? We aren't forces to only test null vs alternative hypothesis!

Look at the data. We are wondering if the loss of mass by frogs is different depending on the substrate. Furthermore, we believe that the shade might have an effect. NORMALLY, WE WOULD ALSO TEST FOR INTERACTIONS, BUT WE WON'T IN THIS CASE (at least at first). Finally, we aren't interested in exploring the effect of initial mass on mass loss, but we think it might be an important factor to add to the models.

The 8 models we want to test are:

1. **Null model**. Biological hypothesis: Mass lost by frogs is constant.

2. **Null modelwith mass.** Biological Hypothesis: Mass lost by frogs is a result of frog size. But there is **NO EFFECT** of shade or substrate (the 2 variables you're interested in).
3. **Shade model.** Biological Hypothesis: Mass lost by frogs varies with shade
4. **Shade model with mass.** Biological Hypothesis: Mass lost by frogs varies with shade and frog size
5. **Substrate model.** Biological Hypothesis: Mass lost by frogs varies with substrate type
6. **Substrate model with mass.** Biological Hypothesis: Mass lost by frogs varies with substrate type and frog size
7. **Shade and Substrate model.** Biological Hypothesis: Mass lost by frogs varies with shade and substrate type
8. **Shade and Substrate model with mass.** Biological Hypothesis: Mass lost by frogs varies with shade, substrate type, and frog size.

Pff... that was a lot of typing. Now imagine, some studies run > 100 models each representing a different hypothesis. Also, we skipped the interactive models that we would usually run!

I recommend reading Mazerolle's paper. Unfortunately, we haven't had time in class to explore the complexities of real biological data. From previous research, we know that the effect of mass isquadratic (more on this later in the course!). They also realized their data needs to be cenetered. We will center initial mass by subtracting the mean of the variable from each value:

```
frog$InitMass_cent <- frog$Initial_mass - mean(frog$Initial_mass)
```

This might seem super confusing and complicated. No worries! All you need to know for now is that:

1. This is a specific situation with this dataset. Don't focus too much on this
2. We will explain quadratic models after spring break
3. When you include the effects of mass in a model, in order to make it quadratic you need to include the following:

   ```
   InitMass_cent + I(InitMass_cent^2)
   ```

So, a quadratic effect actually has 2 parameters.

Again, don't worry too much about that for the time being.

Running the global model

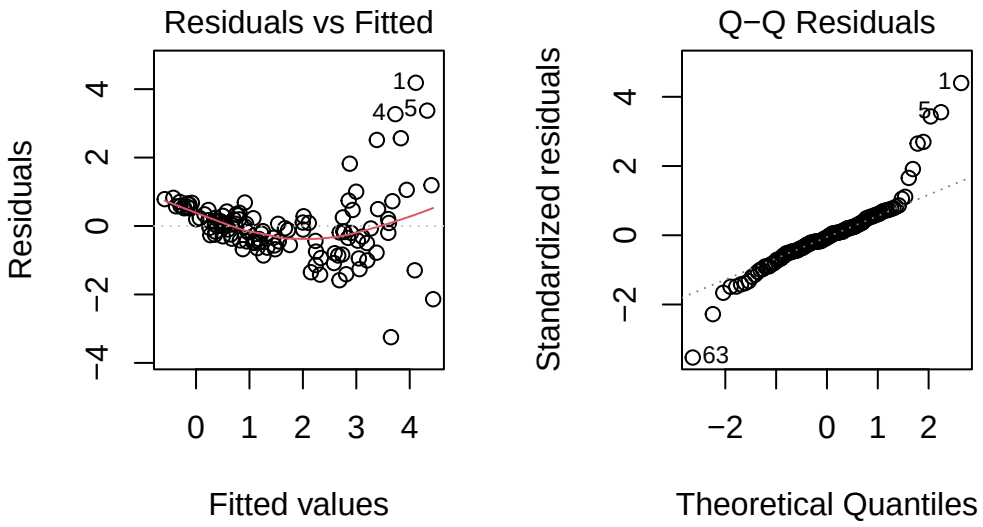
The first step is to run our most complex model. In this case, that's model 8.

```
frogM8<-lm(Mass_lost ~ InitMass_cent + I(InitMass_cent^2) + Substrate + Shade,
data = frog)
```

Please do note how we added: InitMass_cent + I(InitMass_cent^2) to include the effect of mass.

Then we need to check the assumptions. We can test the assumptions by running a residuals vs fitted plot and a QQplot. You already know how to do this. Here is a different and new way to do it:

```
par(mfrow = c(1, 2))
plot(frogM8,which=c(1,2))
```



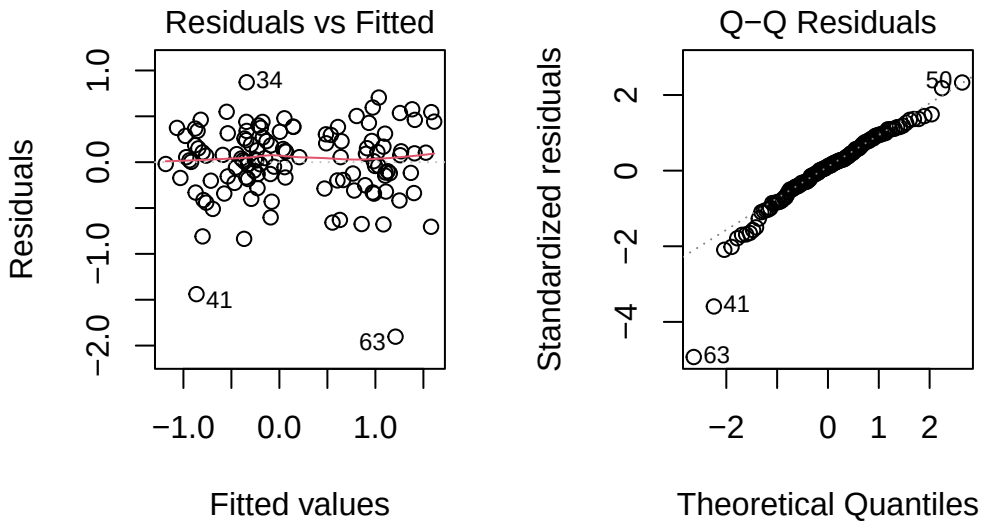

```
par(mfrow = c(1, 1))
```

Woah, this looks horrible! What next? You should go back, read the original paper and come up with an appropriate model to run (just kidding!). In this case we will just transform the data, because there is no homoscedasticity.

We will:

1. Create a new variable which is the log-transformed data for lost mass
2. Add a 0.1 to the data (there are zeroes!)
3. Run the model with the new variable
4. Check the new plots

```
frog$logMass_lost <- log(frog$Mass_lost + 0.1)
frogM8<-lm(logMass_lost ~ InitMass_cent + I(InitMass_cent^2) + Substrate + Shade,
data = frog)
par(mfrow = c(1, 2))
plot(frogM8,which=c(1,2))
```




```
par(mfrow = c(1, 1))
```

Well, not great, but definitely better. And this is usually good enough to work with. It's also how ecological data looks more often than not!

Running the candidate models

Now it's time for you to run all the other 7 candidate models. Please make sure to use the log transformed response variable when running your model. Finally, using AIC, you can select the best model.

Q4. 8 pts. Run the other 7 models. And then compare ALL models using AIC. Which one is the best model?

```
frogM1<-lm(logMass_lost ~ 1, data=frog)
frogM2<-lm(logMass_lost ~ 1+ (InitMass_cent + I(InitMass_cent^2)), data=frog)
frogM3<-lm(logMass_lost~Shade, data=frog)
frogM4<-lm(logMass_lost~Shade+(InitMass_cent + I(InitMass_cent^2)), data=frog)
frogM5<-lm(logMass_lost~Substrate, data=frog)
frogM6<-lm(logMass_lost~Substrate + (InitMass_cent + I(InitMass_cent^2)), data=frog)
frogM7<-lm(logMass_lost~Shade+Substrate,data=frog)
frogM8<-lm(logMass_lost ~ InitMass_cent + I(InitMass_cent^2) + Substrate + Shade,
data = frog)
```

Now, we can check the AIC

```
AIC(frogM1,frogM2,frogM3,frogM4,frogM5,frogM6,frogM7,frogM8)
```

```
       df      AIC
frogM1  2 314.9176
frogM2  4 185.0897
frogM3  3 313.5351
frogM4  5 175.2572
frogM5  4 303.0660
frogM6  6 150.3579
frogM7  5 301.0611
frogM8  7 135.3607
```

Now that we know the best model, you can focus on that one model. Look at the best model summary again.

A4.Model 8 is the best because it has the lowest AIC.

Q5. 2 pts. Describe the best model. What factors affect the mass loss in frogs?

A5. 2pts. Model 8 models the frog's lost mass with explanatory variables including the initial mass, substrate (including soil and Sphagnum), and shade. This model has an overall significant p-value, and significant p-values for all individual coefficients as well.

```
summary(frogM8)
```

```
Call:
lm(formula = logMass_lost ~ InitMass_cent + I(InitMass_cent^2) +
    Substrate + Shade, data = frog)

Residuals:
     Min       1Q   Median       3Q      Max
-1.90283 -0.18367  0.04429  0.26902  0.87140

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         0.6912268  0.0851158   8.121 5.81e-13 ***
InitMass_cent       0.0576657  0.0033549  17.189  < 2e-16 ***
I(InitMass_cent^2) -0.0011286  0.0001736  -6.502 2.13e-09 ***
SubstrateSOIL       0.1849743  0.0926116   1.997   0.0482 *
SubstrateSPHAGNUM  -0.4414430  0.0923340  -4.781 5.21e-06 ***
Shade              -0.3114384  0.0747821  -4.165 6.06e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4098 on 115 degrees of freedom
Multiple R-squared:  0.7912,    Adjusted R-squared:  0.7822
F-statistic: 87.18 on 5 and 115 DF,  p-value: < 2.2e-16
```

Interactive models

Finally, and just for fun, try running the global model again (frogM8). Just this time, there is an interactive effect between shade and substrate (but not with mass). Call it frogM9.

Q6. 4 pts. Run frogM9. Is it better than your best model from Questions 4 and 5? How would you interpret this model?

A6. Model 9 models the frog's lost mass with explanatory variables including the initial mass, substrate (including soil and Sphagnum), and shade, and also incorporates interaction terms between the substrate types and shade. According to AIC, Model 9 now has the lowest AIC. However, this value is only about 2 below the AIC of Model 8. Since the AIC are so close, the two models are considered largely the same. So, I would say that model 8 is the best since the AIC analysis indicates the models are not substantially different, but this model has less complexity. We can also affirm this decision by examining the model summary for FrogM9, where we see that the p-values for the coefficients on the interaction terms are larger than 0.05 (that is, not signicant), meaning we do not have sufficient support to conclude that their coefficients are nonzero.

```
frogM9<-lm(logMass_lost ~ InitMass_cent + I(InitMass_cent^2) + Substrate + Shade + Substrate
data = frog)
AIC(frogM1,frogM2,frogM3,frogM4,frogM5,frogM6,frogM7,frogM8,frogM9)
```

```
       df      AIC
frogM1  2 314.9176
frogM2  4 185.0897
frogM3  3 313.5351
frogM4  5 175.2572
frogM5  4 303.0660
frogM6  6 150.3579
frogM7  5 301.0611
frogM8  7 135.3607
frogM9  9 133.0256
```

```
summary(frogM9)
```

```
Call:
lm(formula = logMass_lost ~ InitMass_cent + I(InitMass_cent^2) +
    Substrate + Shade + Substrate:Shade, data = frog)

Residuals:
     Min       1Q   Median       3Q      Max
-1.76711 -0.19315  0.00278  0.24119  0.89008

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              0.6862475  0.1014936   6.761 6.30e-10 ***
```

```
InitMass_cent            0.0580825  0.0033087  17.554  < 2e-16 ***
I(InitMass_cent^2)      -0.0011685  0.0001713  -6.820 4.73e-10 ***
SubstrateSOIL            0.3197970  0.1323167   2.417   0.0173 *
SubstrateSPHAGNUM       -0.5322374  0.1310136  -4.062 8.99e-05 ***
Shade                   -0.2877646  0.1297333  -2.218   0.0285 *
SubstrateSOIL:Shade     -0.2560159  0.1813199  -1.412   0.1607
SubstrateSPHAGNUM:Shade  0.1826643  0.1811127   1.009   0.3153
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4028 on 113 degrees of freedom
Multiple R-squared:  0.8019,    Adjusted R-squared:  0.7896
F-statistic: 65.34 on 7 and 113 DF,  p-value: < 2.2e-16
```

Congrats, you are done with this assignment :)

Total points: 21