

Assignment 5

Dr. Alejandro Molina-Moctezuma

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!



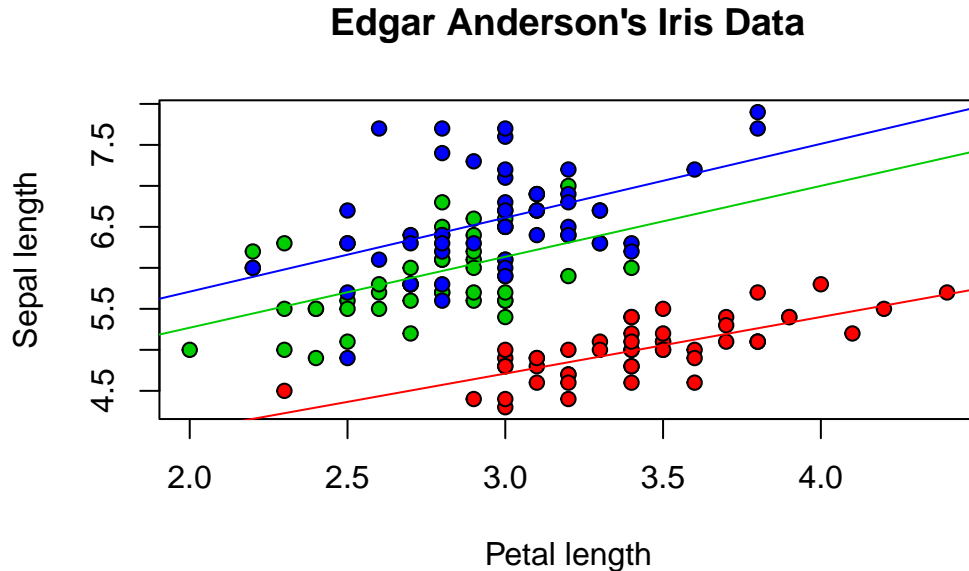
Q1. 3 pts. Run the summary of your model, and interpret it. Look at the F-statistic, and the model p-value. (null hypothesis is, this model explains the data just as well as a null model). Is it significantly better than a null model? Look at each coefficient and whether they are significant. Are they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

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 individually 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$Species)]  
abline(lm(Sepal.Length ~ Sepal.Width, data=iris[which(iris$Species=="setosa"),])$coefficients, lty=2))
```

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



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 last one?

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?

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 4.3.2

```
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 forced 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 is quadratic (more on this later in the course!). They also realized their data needs to be centered. 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

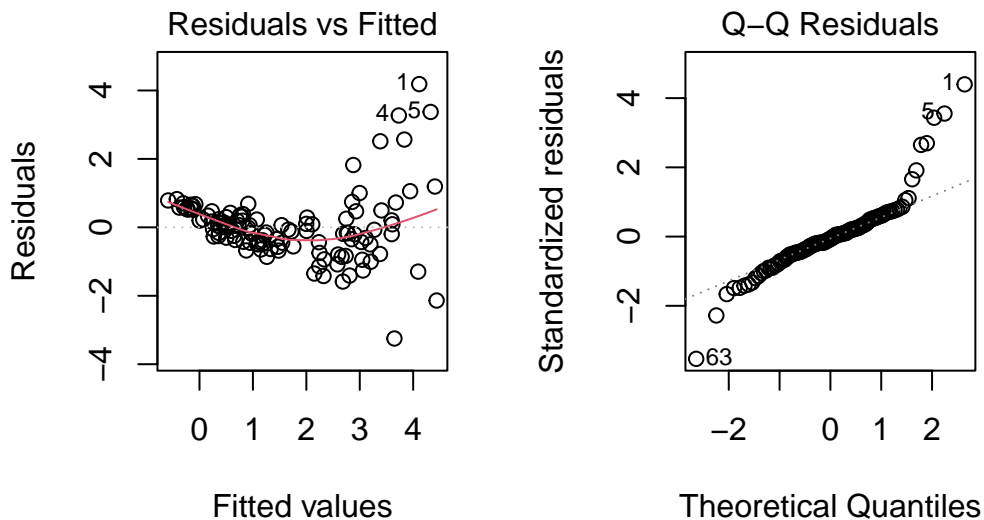
The first step is to run our most complex model. In this case, that's model 8.

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

Please do note how we added: $\text{InitMass_cent} + I(\text{InitMass_cent}^2)$ 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))
```

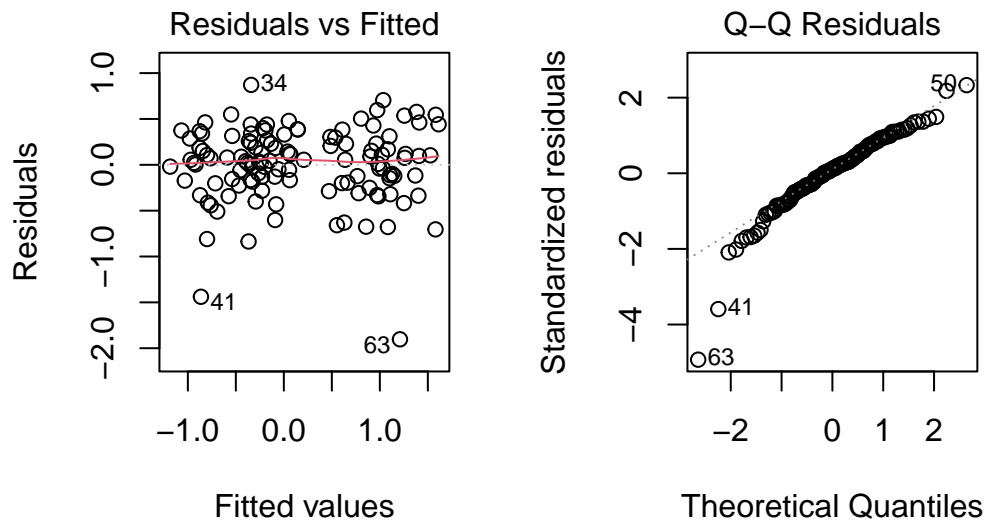
Woah, this looks horrible! What next? You should go back, read the original paper and come up with an appropriate model to run (just kidding!). In this case we will just transform the data, because there is no homoscedasticity.

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?

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



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

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?

Congrats, you are done with this assignment :)

Total points: 21