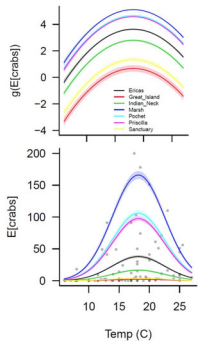
ECO 636
Applied Ecological Statistics

Week 3 – Linear models with multiple categorical predictors – recorded lecture



Meg Graham MacLean, PhD

Department of Environmental Conservation

mgmaclean@umass.edu

2021 - Spring

Welcome to week 3s recorded lecture.

The Week

Tuesday
- Review last week's recorded material
- Linear models with multiple categorical predictors examples

Wednesday (Lab)
- One-way ANOVA

Thursday
- Continue with multiple categorical predictors
- Model selection: AIC

We will finish up this week talking about multiple categorical predictors that interact – as well as model selection using AIC

(c) Meg Graham MacLean
mgmaclean@umass.edu

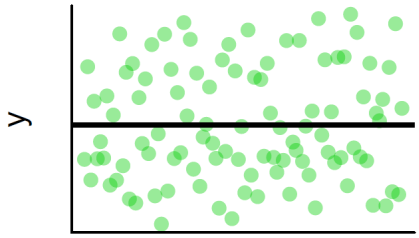
Review!

Response (Y)	Explanatory (X)	Model	In R
Continuous	None	Intercept-only/null	`lm(y~1)`
Continuous	Two-level factor	*t*-test	`lm(y~x)`
Continuous	Multi-level factor	ANOVA	`lm(y~x)`

What does the first (null model) look like mathematically?

$$y_i = \beta_0 + e_i$$

What does the first (null model) look like graphically?



3

I always like to start with a quick review so – here is the null model both mathematically and graphically – remember in this one we just have our intercept or beta naught for the deterministic part of our model.

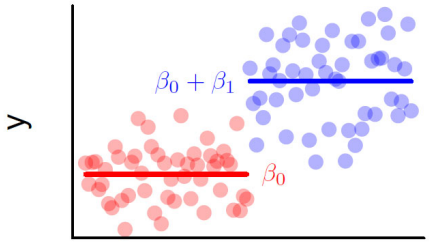
Review!

Response (Y)	Explanatory (X)	Model	In R
Continuous	None	Intercept-only/null	`lm(y~1)`
Continuous	Two-level factor	*t*-test	`lm(y~x)`
Continuous	Multi-level factor	ANOVA	`lm(y~x)`

What does the two-level factor (t-test) look like mathematically?

$$y_i = \beta_0 + \beta_1 X_i + e_i$$

What does the two-level factor (t-test) look like graphically?



4

Next we have our t-test, or when we have a two-level factor for an explanatory variable. Here we have beta-naught plus beta-one x for our deterministic part of our model, where beta-naught is the man for group one and beta-naught plus beta-one is the mean for group two.

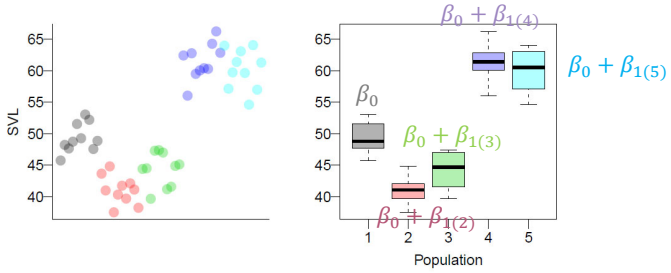
Review!

Response (Y)	Explanatory (X)	Model	In R
Continuous	None	Intercept-only/null	`lm(y~1)`
Continuous	Two-level factor	*t*-test	`lm(y~x)`
Continuous	Multi-level factor	ANOVA	`lm(y~x)`

What does the multiple-level factor (ANOVA) look like mathematically?

$$y_i = \beta_0 + \beta_{1(g)} X_{i(g)} + e_i$$

What does the multiple-level factor (ANOVA) look like graphically?



5

Next we have our one-way anova, where we have a multiple-level factor for x, so now we have many different beta-ones. One for each group other than the first group or our reference group.

(c) Meg Graham MacLean
mgmaclean@umass.edu

>1 explanatory variables multiple samples!

Let's try this again, but with more explanatory variables

Next!

Response (Y)	Explanatory (X)	Model	In R
Continuous	None	Intercept-only/null	`lm(y~1)`
Continuous	Single two-level factor	*t*-test	`lm(y~x)`
Continuous	Single multi-level factor	One-way ANOVA	`lm(y~x)`
Continuous	>1 multi-level factor (+)	Two-way ANOVA	`lm(y~x1+x2)`

Then, on Tuesday we started to introduce a two-way anova, where we have more than one explanatory variable and they are all factors. So far we have stuck with an additive model and today we will expand on that and go through an interaction model. But first, lets go over an additive model again to be sure we have it down.

(c) Meg Graham MacLean
mgmaclean@umass.edu

Two-way ANOVA as a linear model

$$y_i = \beta_0 + \beta_{1(g)}X_{1i(g)} + \beta_{2(g)}X_{2i(g)} + e_i \text{ (additive model)}$$

- β_0 is the mean of the first combination of factors
- $\beta_{1(g)}$ is the group 1 contrasts
  - The difference between the reference level and the other groups in X_1
- $\beta_{2(g)}$ is the group 2 contrasts
  - The difference between the reference level and the other groups in X_2
- $e_i \sim N(0, \sigma)$

So, let's recap the parts of a two-way additive ANOVA. In this model, we still have our reference level, and beta-naught is still the mean of that reference group or level. Usually this reference level is the first combination of factors that appear alpha-numerically - however, you can set your reference level using the relevel() function if you have something like a control group.

Next, beta-one is the set of group one contrast, or the difference between the reference level and the other groups of the first explanatory variable.

Second, we have beta-two, which is the same thing, but for group two.

(c) Meg Graham MacLean
mgmaclean@umass.edu

Example

Example of water vole weights

- 4 water vole sub-populations ("Networks")
- 2 sexes of interest (males and females)
- 100+ voles from each area
- Question: Does weight vary by:
  - Sex? (2-level factor)
  - Population/Network? (4-level factor)
  - Both?
- H_0: there are no differences in weight between:
  - Sexes
  - Network

So, lets jump right into an example! This time we will use a new set of water vole data, posted on Moodle. The picture right here is of a water vole – pretty cute! There are a few different pieces to these data – but we are going to focus on the sub-populations, or networks, as well as the sex of the individuals. What is nice about these data is that there are a lot of vole observations from each area.

So, the question we are interested in exploring today is: does the weight of the voles differ by sex (which is a 2-level factor), or by population (which is a 4-level factor), or finally by both?

Well, we haven't talked about interacting factors yet, so let's start by testing if there are differences in weight by sex or by network. That would be our two-way additive ANOVA.

Example

Example of water vole weights

- 4 water vole sub-populations ("Networks")
- 2 sexes of interest (males and females)
- 100+ voles from each area
- Question: Does weight vary by:
  - Sex? (2-level factor)
  - Population/Network? (4-level factor)
  - Both?
- H_0: there are no differences in weight between:
  - Sexes
  - Network

Follow along in R!

So, for this recorded lecture, I am going to have you follow along in R.

Depending on your comfort level with these commands, you can start a brand new script for the water voles, or feel free to copy the script you used for the salamanders on Tuesday and modify it to work with the water voles. This will be the exercise for today, so the lecture recording is a bit on the longer side – so, please feel free to pause and back up as you are working if you need more time with each step.

(c) Meg Graham MacLean
mgmaclean@umass.edu

Example

Comparing sex-specific vole weights across multiple populations:

```
> vole <- read.table(file = "voleWt.txt", h=T)
> head(vole)
```

	Year	Network	Species	Sex	Weight
1	2010	SGI	water.vole	female	25
2	2010	SGI	water.vole	female	25
3	2011	CRO	water.vole	female	25
4	2011	CRO	water.vole	female	25
5	2011	SGI	water.vole	female	25
6	2004	SGI	water.vole	female	30

Let's first take a look at the data – we have Network (or the site the voles were observed), sex, and weight in grams. We also have year recorded, but we won't use that today.

(c) Meg Graham MacLean
mgmaclean@umass.edu

Example

Comparing sex-specific vole weights across multiple populations:

```
> vole <- read.table(file = "voleWt.txt", h=T)
> head(vole)
> xtabs(~Network+Sex, data=vole)
```

Network	Sex female	male
CRO	508	555
LED	139	152
LEI	140	177
SGI	503	549

Like I pointed out before, we have a lot of observations within each network and by sex, but notice that there is some disparity across network sites. Two of the sites have over 1000 observations (total), but the other two have closer to 300.

Example

Modeling process:

1. *State the question/hypothesis*
   - *What is the question?*
   - *What are the variables (response and explanatory)?*
2. Data exploration
3. Describe the model
   - In word form (should come from your question)
   - In mathematical form
   - Identify the assumptions of the model
4. Fit the model! (In R, of course ☺)
5. Evaluate the output
   - Model validation
   - Model selection
6. Interpret the results

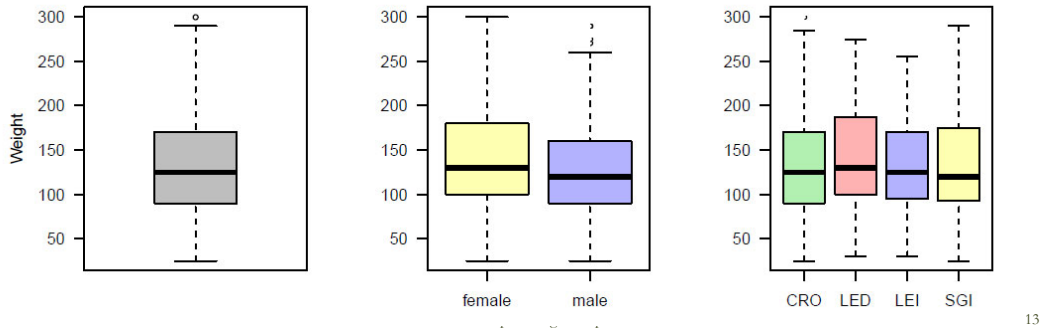
So, let's start the modeling process! We already set up the question for ourselves, and we did a bit of data exploration – but lets do a bit more!

(c) Meg Graham MacLean
mgmaclean@umass.edu

Example

2. Data exploration - boxplots

```
> boxplot(vole$Weight, ylab="Weight")   #all
> boxplot(vole$Weight ~ vole$Sex)       #by sex
> boxplot(vole$Weight ~ vole$Network)   #by network
```



13

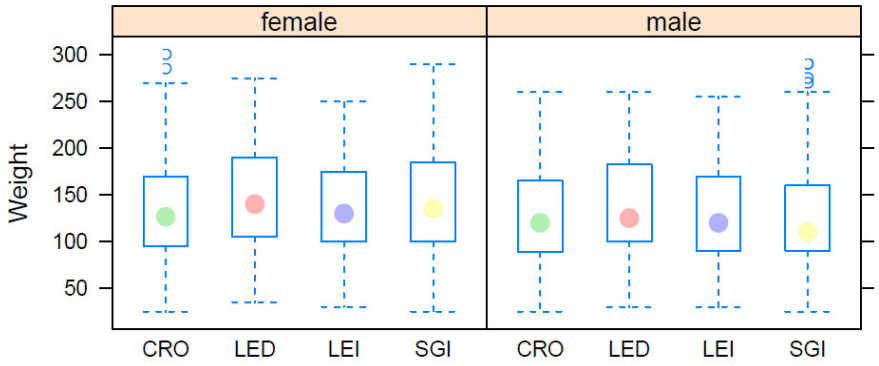
Let's check out the distribution of the data using some box plots. If we look at all the weights, the data look a little skewed, but not too bad, but if we start splitting it up by the explanatory variables we are interested in, we start to see slightly different patterns.

If you had to guess if there were differences in weight between sexes or networks, what would you guess?

Example

2. Data exploration - conditional box plots

```
> #condition on network AND sex
> bwplot(vole$Weight ~ vole$Network | vole$Sex)
```



14

We can also look at the interaction of sex and network using this conditional boxplot as well to start to see differences!

Example

Modeling process:

1. *State the question/hypothesis*
   - *What is the question?*
   - *What are the variables (response and explanatory)?*
2. *Data exploration*
3. Describe the model
   - In word form (should come from your question)
   - In mathematical form
   - Identify the assumptions of the model
4. Fit the model! (In R, of course ☺)
5. Evaluate the output
   - Model validation
   - Model selection
6. Interpret the results

Great – lets move on to describing the model.

(c) Meg Graham MacLean
mgmaclean@umass.edu

Example

3. Describe the model

- In words:
  - Are there significant differences in weight with sex or network?
- As a mathematical model:
  - $y_i = \beta_0 + \beta_{1(g)}Network_{1i(g)} + \beta_{2(g)}Sex_{2i(g)} + e_i$
  - Are $\beta_{1(g)}$'s different from 0?
  - Are $\beta_{2(g)}$'s different from 0?

So our question is: are there significant differences in weight with sex or network? And mathematically that is written out like this. Can you describe what the betas represent in words?

The for each of the betas, the null hypothesis is that the betas, all together, are no different from zero.

(c) Meg Graham MacLean
mgmaclean@umass.edu

Example

3. Describe the model

- In words:
  - Are there significant differences in weight among sex-network combinations?
- As a mathematical model:
  - $y_i = \beta_0 + \beta_{1(g)}Network_{1i(g)} + \beta_{2(g)}Sex_{2i(g)} + e_i$
  - $H_{0(network)}$: $\beta_{1(2)} = \beta_{1(3)} = \beta_{1(4)} = 0$
  - $H_{0(sex)}$: $\beta_{2(male)} = 0$
- Assumptions?
  - Residuals are normally distributed
  - Constant variance (homogeneity)
  - Observations are independent
  - Predictors measured without error (fixed X)

That looks like this, where the betas for all groups are the same and equal to zero (which would indicate that there is no contrast between these groups and the reference group.

What are the assumptions of the model?

Well.. These are the assumptions…

(c) Meg Graham MacLean
mgmaclean@umass.edu

Example

Modeling process:

1. *State the question/hypothesis*
   - *What is the question?*
   - *What are the variables (response and explanatory)?*
2. *Data exploration*
3. *Describe the model*
   - *In word form (should come from your question)*
   - *In mathematical form*
   - *Identify the assumptions of the model*
4. Fit the model! (In R, of course ☺)
5. Evaluate the output
   - Model validation
   - Model selection
6. Interpret the results

Let's fit the model! Here is where you will need your R script ready to go.

(c) Meg Graham MacLean
mgmaclean@umass.edu

Example

4. Fit the model

- Algebra: $y_i = \beta_0 + \beta_{1(g)}Network_{1i(g)} + \beta_{2(g)}Sex_{2i(g)} + e_i$
- R:

```
> mSexPop <- lm(Weight ~ Network + Sex, data = vole)
```


So algebraically, we have described the model – so next we should fit our linear model in R.

(c) Meg Graham MacLean
mgmaclean@umass.edu

Example

4. Fit the model

```
> mSexPop <- lm(Weight ~ Network + Sex, data = vole)
> coef(mSexPop)
  (Intercept)  NetworkLED  NetworkLEI  NetworkSGI     Sexmale
   135.280799    9.027931    1.803145    2.794129  -10.110791
```

```
> (tapply(vole$Weight,list(vole$Network,vole$Sex),mean,na.rm=T))
       female     male
CRO 132.2579 127.9369
LED 143.9209 134.5526
LEI 136.0143 127.8192
SGI 141.5328 124.7960
```

How do the estimates compare to the means?

Here are our model coefficients along with the computed means for each group… take a moment and try to answer, how do these coefficients line up with our computed means? What do each of the estimates represent? We'll talk about that on Tuesday.

(c) Meg Graham MacLean
mgmaclean@umass.edu

Example

4. Fit the model

$$y_i = \beta_0 + \beta_{1(g)} * 0 + \beta_{2(g)} * 0 + e_i$$

```
> mSexPop <- lm(Weight ~ Network + Sex, data = vole)
> coef(mSexPop)
  (Intercept)  NetworkLED  NetworkLEI  NetworkSGI     Sexmale
   135.280799   9.027931    1.803145    2.794129  -10.110791
```

```
> (tapply(vole$Weight,list(vole$Network,vole$Sex),mean,na.rm=T))
       female     male
CRO 132.2579 127.9369
LED 143.9209 134.5526
LEI 136.0143 127.8192
SGI 141.5328 124.7960
```

ECO 636 week 3 - Multiple Categorical predictors 21

So, our first estimate of the intercept, or beta-naught, is the estimate of the mean weight for our reference level, which is females in the CRO network, just based on alphabetic order. You'll notice that up in the upper right of the slide I put the equation for this estimate, and I multiplied both beta-one and beta-two by zero, since the factor should be zero for the reference level. So why do you think the estimate and the mean are different??

Well, you may notice that there are differences between the weights of the males and the females, but that they aren't the same across all network sites. When we use an additive model, we assume that the difference between one set of factors is the same across all sets of other factors. To look at this in numbers, lets calculate the estimated mean for another level.

Example

4. Fit the model

$$y_i = \beta_0 + \beta_{1(2)} * 1 + \beta_{2(g)} * 0 + e_i$$

```
> mSexPop <- lm(Weight ~ Network + Sex, data = vole)
> coef(mSexPop)
  (Intercept)  NetworkLED  NetworkLEI  NetworkSGI     Sexmale
   135.280799   9.027931    1.803145    2.794129  -10.110791
```

```
> (tapply(vole$Weight,list(vole$Network,vole$Sex),mean,na.rm=T))
        female     male
CRO 132.2579 127.9369
LED 143.9209 134.5526
LEI 136.0143 127.8192
SGI 141.5328 124.7960
```

ECO 636 week 3 - Multiple Categorical predictors 22

Let's try females at LED. Since we are still using the reference level for sex, beta-two is multiplied by zero, but now we are not using the reference group, so beta-one should now be the estimate for group 2 or LED.

So, to compute the estimate for females at LED, we would take the estimate of beta-naught, or 135.28, and add the contrast or beta-one for LED, or 9.03, which would give us 144.31, which is a bit closer to the actual mean for that level.

Let's try one more that includes males, let's compute the estimate of the mean for males at the SGI network site.

Example

4. Fit the model

$$y_i = \beta_0 + \beta_{1(4)} * 1 + \beta_{2(2)} * 1 + e_i$$

```
> mSexPop <- lm(Weight ~ Network + Sex, data = vole)
> coef(mSexPop)
  (Intercept)  NetworkLED  NetworkLEI  NetworkSGI     Sexmale
   135.280799    9.027931    1.803145    2.794129  -10.110791
```

```
> (tapply(vole$Weight,list(vole$Network,vole$Sex),mean,na.rm=T))
        female     male
  CRO 132.2579 127.9369
  LED 143.9209 134.5526
  LEI 136.0143 127.8192
  SGI 141.5328 124.7960
```

ECO 636 week 3 - Multiple Categorical predictors 23

Like in the last slide, we are not using the reference level for network site, so beta-one is estimate for group four multiplied by one, and beta-two is the estimate for group two, or males, multiplied by one.

So, for males in SGI, that means beta-naught or 135.28, plus 2.79, plus a negative 10.11, equaling 127.96. This estimate is a little off from our computed mean, but remember, in this additive model we are assuming things like for both sexes, water voles at SGI are around 3 grams heavier, and for all network sites, males are always around 10 grams lighter.

Example

Modeling process:

1. *State the question/hypothesis*
   - *What is the question?*
   - *What are the variables (response and explanatory)?*
2. *Data exploration*
3. *Describe the model*
   - *In word form (should come from your question)*
   - *In mathematical form*
   - *Identify the assumptions of the model*
4. *Fit the model! (In R, of course ☺)*
5. Evaluate the output
   - Model validation
   - Model selection
6. Interpret the results

Given these assumptions, lets have a look at our model fit.

(c) Meg Graham MacLean
mgmaclean@umass.edu

Example

5. Evaluate the output

- Model validation - check assumptions!
  - Residuals are normally distributed
  - Constant variance (homogeneity)
  - Observations are independent
  - Predictors measured without error (fixed X)

```
> par(mfrow=c(2,2), oma=c(0,0,0,0))
> plot(mSexPop)
```

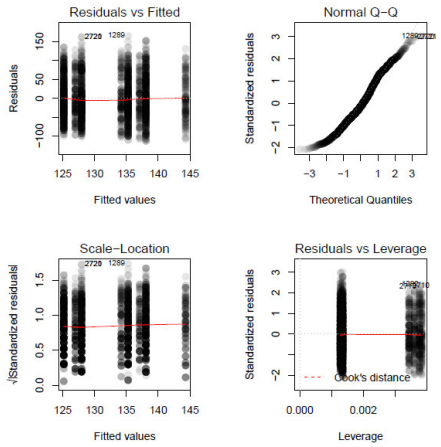
What do you see?

First, lets check our assumptions. We can use the plot() function to do that. Give that a try and see if what you can deduce from the plots.

(c) Meg Graham MacLean
mgmaclean@umass.edu

Example

```
> par(mfrow=c(2,2), oma=c(0,0,0,0))
> plot(mSexPop)
```



Do we meet our assumptions?

26

Do you think we meet our assumptions? To me each of these plots look pretty good, so lets keep going.

Example

5. Evaluate the output

- Model validation - check fit

```
> summary(mSexPop)

Call:
lm(formula = Weight ~ Network + Sex, data = vole)

Residuals:
     Min       1Q   Median       3Q      Max
-113.075  -38.075   -7.964   40.691  164.719

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  135.281      2.000  67.624  < 2e-16 ***
NetworkLED     9.028      3.612   2.499   0.0125 *
NetworkLEI     1.803      3.495   0.516   0.6059
NetworkSGI     2.794      2.374   1.177   0.2394
Sexmale      -10.111      2.096  -4.824 1.49e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 54.6 on 2718 degrees of freedom
Multiple R-squared:  0.01082, Adjusted R-squared:  0.009364
F-statistic: 7.432 on 4 and 2718 DF,  p-value: 5.962e-06
```

Should we include Network?

27

Hmm, in looking at our summary, it doesn't seem like the differences among network sites are very significant, at least between each site and the reference site of CRO.

So, we may ask ourselves… is it worth keeping in Network as an explanatory variable and how might we know that?

Example

5. Evaluate the output

- *Model validation*
- Model selection – what other models could explain water vole weight?
  - Full model: $y_i = \beta_0 + \beta_{1(g)}Network_{1i(g)} + \beta_{2(g)}Sex_{2i(g)} + e_i$
  - Network only: $y_i = \beta_0 + \beta_{1(g)}Network_{1i(g)} + e_i$
  - Sex only: $y_i = \beta_0 + \beta_{2(g)}Sex_{2i(g)} + e_i$
  - Null model! $y_i = \beta_0 + e_i$

Well, here is the model we were just working with, which is often called our full or our global model. So, what are some of the other models we could use to predict water vole weight?

Well, we could use network only, though from our last results it seems like maybe that might not be the best one, we could also just use sex, which might make for a decent model based on our previous results, and finally we should always check the null model! But what should we use to see which is the best one?

(c) Meg Graham MacLean
mgmaclean@umass.edu

Example

5. Evaluate the output

- *Model validation*
- Model selection – what other models could there be to explain water vole weight?
  - Use an information criterion (e.g., AIC) to compare models

```
> modList <- list()
> modList[["mSexPop"]] <- lm(Weight ~ Network + Sex, data = vole)
> modList[["mSex"]]    <- lm(Weight ~ Sex,           data = vole)
> modList[["mPop"]]    <- lm(Weight ~ Network,       data = vole)
> modList[["m0"]]      <- lm(Weight ~ 1,             data = vole)
```


Well, the most recommended way to do model selection is to use an information criterion to determine which model explains the most variation in your data. An example of this is the information criterion or AIC, which I'll explain on the next slide. But first, I wanted to show you one of the ways I set up a model list, or modList, to make it easy to do model selection. By setting up a model list, with model names, it makes it pretty simple and easy to do model selection. I'll often use model lists throughout the semester to do model selection.

(c) Meg Graham MacLean
mgmaclean@umass.edu

AIC

AIC – Akaike Information Criteria

$$AIC = -2\log(likelihood) + 2K$$

- $-2\log(likelihood)$ is a term for goodness of fit
- $2K$ is a penalty for the number of parameters
  - K is the number of parameters

- AIC is a tradeoff between model fit and complexity
  - Pick the most parsimonious model!
  - **Model with the *lowest* AIC is best!**
  - $\Delta AIC = AIC_i - AIC_{min}$

ECO 636 week 3 - Multiple Categ

ΔAIC	Is the model better?
0-2 units	Little difference between models
3-10 units	Some support for the model with a lower AIC
>10 units	Very likely the model with the lower AIC is better

So, back to AIC. AIC is just one version of an information criterion, but it is one of the most common and performs fairly well for most linear models. AIC has two parts, both a measure of goodness of fit, as well as a penalty for the number of parameters, or explanatory variables, in your model. The benefit of having this penalty is to help you not overfit your data. Basically, you don't want to create a model just by throwing variables at some data, because although you may find that you explain the variation in the data, that model won't actually be informative.

So, AIC is a tradeoff between model fit and complexity and it helps you pick the most parsimonious model. In order to do so, you want to pick the model with the lowest AIC. But how do you knowhow much lower and AIC should be for a model to be considered better than another one?

Well, usually you use the delta AIC to determine if one model is better than another, where if there is less than 2 AIC units between two models, there is little difference between the models, somewhere between 3-10 units there is some support for the model with the lower AIC, and finally if there is greater than 10 units difference, it is very likely that the model with the lower AIC is better than the other. Ok, so lets try this with our water voles data.

Example

5. Evaluate the output

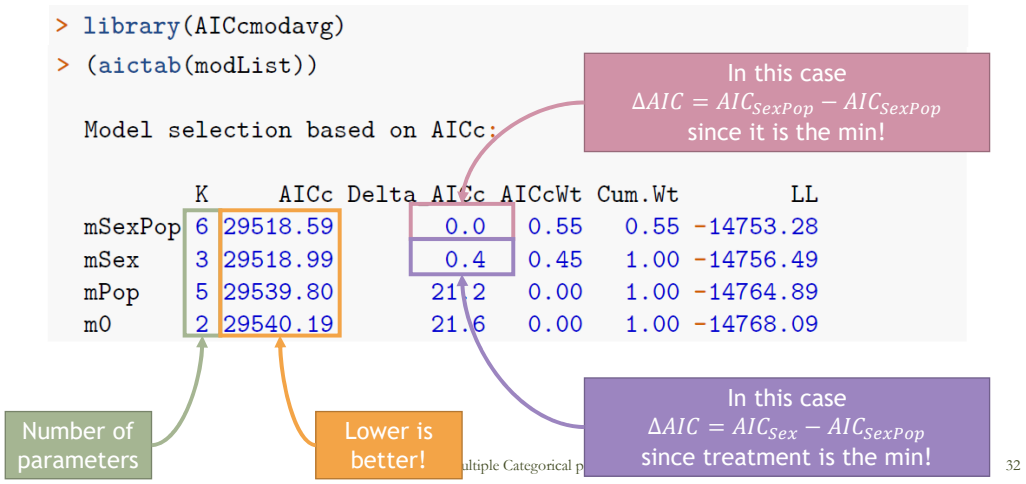
- *Model validation*
- Model selection – what other models could there be to explain water vole weight?
  - Use AIC to compare models

```
> library(AICcmodavg)
> (aictab(modList))

Model selection based on AICc:

         K     AICc Delta_AICc AICcWt Cum.Wt        LL
mSexPop  6 29518.59        0.0   0.55   0.55 -14753.28
mSex     3 29518.99        0.4   0.45   1.00 -14756.49
mPop     5 29539.80       21.2   0.00   1.00 -14764.89
m0       2 29540.19       21.6   0.00   1.00 -14768.09
```

31

Remember we already made our model list – so now we can use the aictab() function from the AICcmodavg package to compute the AICc, or a AIC that has a correction for small sample sizes, and delta AICcs.

AICc



Let's take a look at what each of these mean for our models.

First, K is the number of parameters estimated in each model – for example in our null model we have a K of two, where we are estimating both the beta-naught as well as our e, or stochastic part of our model.

Next is the AICc – where lower is better – so for right now our best model is the full or global model with both sex and network as explanatory variables. However, next up is the delta AICc, where we can see the difference between that model and the AIC of the model with the lowest AIC.

So, in this column, the first Delta_AICc is always zero, because the first model listed is the one with the lowest AICc

However, if we look at the next Delta_AICc, it is the difference between the AIC of the model with just sex as the explanatory variable and the full model. The difference here is quite small! Which means that the difference between these two models is likely not much when it comes to explaining the variation of the data. Looking quickly at the next two models though, these have a delta AICc of more than 10, so we can pretty quickly say that these two models aren't as good as the first two. Given what we saw with our global

mode, this isn't too surprising. If I were analyzing these data, I'd have pretty good reason to use either of the first two models. If I was only interested in the simplest model for explaining water vole weight, I could easily justify using the model with just sex as an explanatory variable. However, let's pretend that I am interested in the differences, or lack thereof, in the network sites – I could easily choose to go with the full model.

Example

Modeling process:

1. *State the question/hypothesis*
   - *What is the question?*
   - *What are the variables (response and explanatory)?*
2. *Data exploration*
3. *Describe the model*
   - *In word form (should come from your question)*
   - *In mathematical form*
   - *Identify the assumptions of the model*
4. *Fit the model! (In R, of course ☺)*
5. *Evaluate the output*
   - *Model validation*
   - *Model selection*
6. Interpret the results

So let's use the full model to interpret the results.

(c) Meg Graham MacLean
mgmaclean@umass.edu

Example

6. Interpret the results

```
> summary(mSexPop)

Call:
lm(formula = Weight ~ Network + Sex, data = vole)

Residuals:
     Min       1Q   Median       3Q      Max
-113.075  -38.075   -7.964   40.691  164.719

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  135.281      2.000  67.624  < 2e-16 ***
NetworkLED     9.028      3.612   2.499   0.0125 *
NetworkLEI     1.803      3.495   0.516   0.6059
NetworkSGI     2.794      2.374   1.177   0.2394
Sexmale      -10.111      2.096  -4.824 1.49e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 54.6 on 2718 degrees of freedom
Multiple R-squared:  0.01082, Adjusted R-squared:  0.009364
F-statistic: 7.432 on 4 and 2718 DF,  p-value: 5.962e-06
```

What about pairwise comparisons?

34

Here is the summary again like we saw before, where we do see a significant difference by sex, but less so between the network sites and the reference network site CRO.

But what about pairwise comparisons across all of the sites? Are they all still insignificantly different?

Example

6. Interpret the results

```
> aov.mSexPop <- aov(mSexPop)           # anova table
> tuk.mSexPop <- TukeyHSD(aov.mSexPop) # pairwise comparisons
```

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = mSexPop)

$Network
              diff         lwr       upr     p adj
LED-CRO  9.025610  -0.2600859 18.311306 0.0603323
LEI-CRO  1.436604  -7.5451574 10.418366 0.9765527
SGI-CRO  2.796598  -3.3071833  8.900378 0.6409641
LEI-LED -7.589006 -18.9834700  3.805459 0.3174740
SGI-LED -6.229012 -15.5251360  3.067111 0.3120836
SGI-LEI  1.359993  -7.6325487 10.352535 0.9800469

$Sex
                 diff      lwr       upr   p adj
male-female -10.10528 -14.2142 -5.996365 1.5e-06
```

35

Here is a Tukey HSD to give us pairwise comparisons, and we find that yes, there aren't any significant differences between network sites, but there is between sexes.

Example

6. Interpret the results

Confidence intervals:

```
> df.F <- data.frame(Network = sort(factor(unique(vole$Network))),
+                    Sex = "female")
> df.M <- data.frame(Network = sort(factor(unique(vole$Network))),
+                    Sex = "male")
> conf.F <- predict(modList[[1]], df.F, interval = "confidence")
> conf.M <- predict(modList[[1]], df.M, interval = "confidence")
> conf.F                          > conf.M
        fit      lwr      upr            fit      lwr      upr
1 135.2808 131.3582 139.2034     1 125.1700 121.3438 128.9963
2 144.3087 137.6758 150.9416     2 134.1979 127.6222 140.7737
3 137.0839 130.6479 143.5200     3 126.9732 120.6921 133.2542
4 138.0749 134.1385 142.0114     4 127.9641 124.1227 131.8056
```

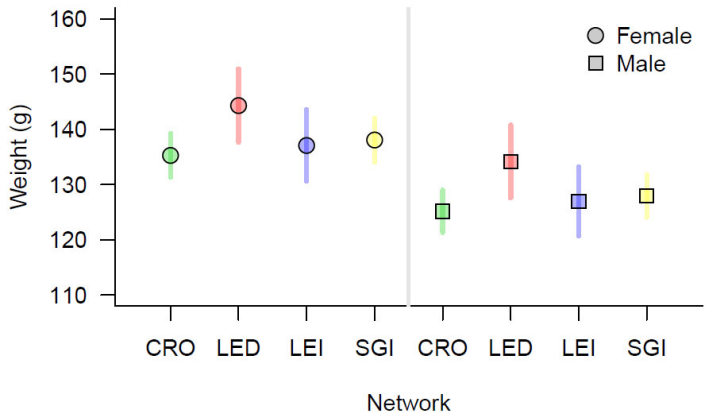

We can then plot the confidence intervals

(c) Meg Graham MacLean
mgmaclean@umass.edu

Example

6. Interpret the results

Confidence intervals:



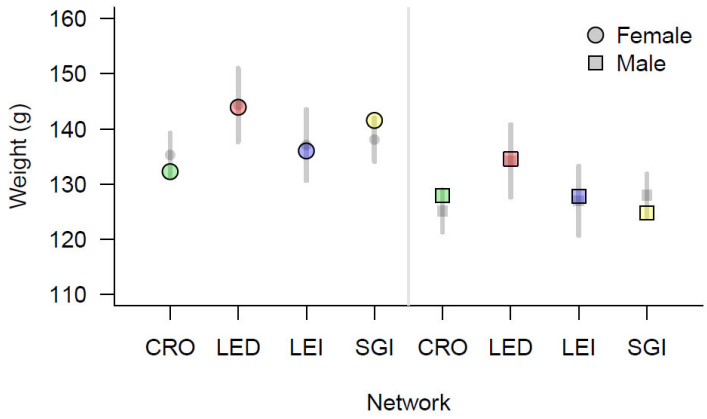
37

Here is one plot of those confidence intervals – feel free to try something similar or simpler in your own code.

Example

6. Interpret the results

Estimated means and CI vs empirical means:



38

Similarly we can also layer the computed means on top – now in color with the estimates in grey. Note that greys follow the same patterns across sites, but the computed means do not, so – here is where an interaction model might be an even better model of water vole weights!

Multiple samples!

Let's try this again, but with two samples (not the null model)

So far…

Response (Y)	Explanatory (X)	Model	In R
Continuous	None	Intercept-only/null	`lm(y~1)`
Continuous	Single two-level factor	*t*-test	`lm(y~x)`
Continuous	Single multi-level factor	One-way ANOVA	`lm(y~x)`
Continuous	>1 multi-level factor (+)	Two-way ANOVA	$lm(y \sim x_1 + x_2)$

So, we just finished an additive two-way ANOVA

(c) Meg Graham MacLean
mgmaclean@umass.edu

Multiple samples!

Let's try this again, but with two samples (not the null model)

Next, with interactions!

Response (Y)	Explanatory (X)	Model	In R
Continuous	None	Intercept-only/null	`lm(y~1)`
Continuous	Single two-level factor	*t*-test	`lm(y~x)`
Continuous	Single multi-level factor	One-way ANOVA	`lm(y~x)`
Continuous	>1 multi-level factor (*)	Two-way ANOVA	`lm(y~x1*x2)`

So now let's do a two-way ANOVA with interactions!

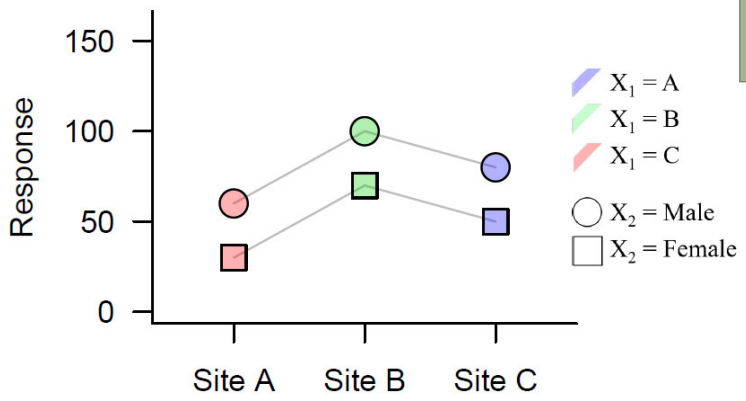
(c) Meg Graham MacLean
mgmaclean@umass.edu

Additive model

What does the additive model look like mathematically?

$$y_i = \beta_0 + \beta_{1(g)}X_{1i(g)} + \beta_{2(g)}X_{2i(g)} + e_i$$

Sex effects identical across sites,
Site effects identical between sexes



41

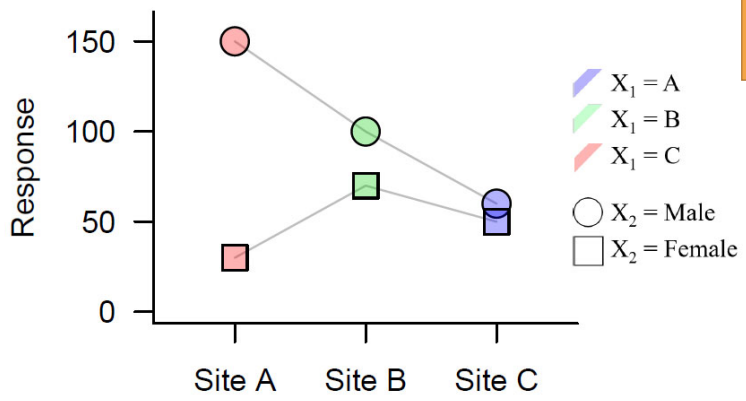
So with our additive model, sex effects stayed identical across sites, and site effects stayed identical between sexes.

Interaction model

What does the interaction model look like mathematically?

$$y_i = \beta_0 + \beta_{1(g)}X_{1i(g)} + \beta_{2(g)}X_{2i(g)} + \beta_{3(g)}X_{1i(g)}X_{2i(g)} + e_i$$

Sex effects *differ* across sites, Site effects *differ* between sexes



42

But with a model with interactions, sex effects can differ across sites, and site effects can differ between sexes!

So what do you think this model might look like mathematically?

It is very similar to our additive mode, but we add a beta-three for the interaction.

Interaction model

What does the interaction model look like mathematically?

$$y_i = \beta_0 + \beta_{1(g)}X_{1i(g)} + \beta_{2(g)}X_{2i(g)} + \beta_{3(g)}X_{1i(g)}X_{2i(g)} + e_i$$

Example:

- X_1: sex (female or male)
- X_2: population (A, B, C, or D)
- β_0 is the mean for population A and females
- $\beta_{1(g)}$: the group 1 contrasts/effects
- $\beta_{2(g)}$: the group 2 contrasts/effects
- $\beta_{3(g)}$: the interaction effects
  - The effect of the 2nd factor on each additional level of the first factor
  - This term only matters when BOTH X_1 and X_2 are 1!
  - Will have the same number of columns as $g2 - 1$

Let's go through each piece. First we have our two explanatory variables, and for this let's pretend we have similar explanatory variables as we had in our water vole example. So we have X1 as sex and X2 as our population with four levels. Then beta-naught is our mean for population A and females (based on the alphabetical order). Then we have beta-one as the group 1 or contrast between the sexes, and beta-two is the contract between the population groups. Finally, we have our new beta-three! Beta-three is the effect of the second factor, in this case population, on each additional level of the first factor, or the sexes – or how the contrast of each specific interaction of sex and population from the reference level. Note that this term only matters when both X1 and X2 are one, or different from the reference level. For example, when sex is male and the group is b. Let's look at what I mean by this in a dummy example.

Interaction model

What does the model matrix look like?

Let's start simple:

```
> sex <- c("F","F","F","F","M","M","M","M")
> pop <- c("A","B","C","D","A","B","C","D")
> model.matrix(~sex) # sex only
```

```
  (Intercept) sexM
1           1    0
2           1    0
3           1    0
4           1    0
5           1    1
6           1    1
7           1    1
8           1    1
```

44

So if I make a dummy set with sex as either male or female, and populations a through d and one observation in each, we can look at when each of our betas matter.

So, let's pretend for just a minute that only sex is in our model, well we have out intercept or beta-naught that is estimated for each group, and then our sexM is our X1, where it only matters if the sex is male. Therefore, our estimate of the mean for the females is just beta-naught and the estimate for the males is beta-naught plus beta-one.

Let's make it a little more complicated.

Interaction model

What does the model matrix look like?

Slightly more complex:

```
> sex <- c("F","F","F","F","M","M","M","M")
> pop <- c("A","B","C","D","A","B","C","D")
> model.matrix(~pop) # pop only
```

```
  (Intercept) popB popC popD
1           1    0    0    0
2           1    1    0    0
3           1    0    1    0
4           1    0    0    1
5           1    0    0    0
6           1    1    0    0
7           1    0    1    0
8           1    0    0    1
```

45

So, what if population was our only explanatory variable? Again, to estimate the mean for population A, we use beta-naught, which is estimated for each group, but for population A, the X1 for all of the other groups is zero. However, if we want to estimate the mean for pop B, X1 for pop b is one, but zero for pop C and pop D. Therefore, we estimate beta-naught and beta-one for popB. Lets expand on this for an additive model.

Interaction model

What does the model matrix look like?

Additive model:

```
> sex <- c("F","F","F","F","M","M","M","M")
> pop <- c("A","B","C","D","A","B","C","D")
> model.matrix(~sex+pop) # both sex and pop
```

	(Intercept)	sexM	popB	popC	popD
1	1	0	0	0	0
2	1	0	1	0	0
3	1	0	0	1	0
4	1	0	0	0	1
5	1	1	0	0	0
6	1	1	1	0	0
7	1	1	0	1	0
8	1	1	0	0	1

46

So now we have both sex and population as explanatory variables. Remember from earlier that if we want to estimate the parameters for females in population b, we set X1 for sex equal to zero since we are using the reference sex, and then the X2 for population b to 1, and estimate beta-naught and beta-two for population b. So, we can use these tables to think through how R codes up and then estimates each parameter. So let's look at this for an interaction model.

Interaction model

What does the model matrix look like?

Interaction model:

```
> sex <- c("F","F","F","F","M","M","M","M")
> pop <- c("A","B","C","D","A","B","C","D")
> model.matrix(~sex*pop) # use '*' for interaction
```

	(Intercept)	sexM	popB	popC	popD	sexM:popB	sexM:popC	sexM:popD
1	1	0	0	0	0	0	0	0
2	1	0	1	0	0	0	0	0
3	1	0	0	1	0	0	0	0
4	1	0	0	0	1	0	0	0
5	1	1	0	0	0	0	0	0
6	1	1	1	0	0	1	0	0
7	1	1	0	1	0	0	1	0
8	1	1	0	0	1	0	0	1

47

Neat! So notice that in this matrix, ones for the interaction are only toward the end of the table. That's because in order for their to be an interaction, both the X1 for sex and X2 for population must be equal to one. That only happens for the male observations in populations other than A, or the reference level.

Example

Example of water vole weights

- 4 water vole sub-populations ("Networks")
- 2 sexes of interest (males and females)
- 100+ voles from each area
- Question: Does weight vary by:
  - Sex? (2-level factor)
  - Population/Network? (4-level factor)
  - Both?
- H_0: there are no differences in weight between:
  - Sexes
  - Network

Earlier we asked the question if weight varies by both sex and network site, but were unable to answer that given an additive model so, let's try an interaction model!

(c) Meg Graham MacLean
mgmaclean@umass.edu

Example

Modeling process:

1. *State the question/hypothesis*
   - *What is the question?*
   - *What are the variables (response and explanatory)?*
2. Data exploration
3. Describe the model
   - In word form (should come from your question)
   - In mathematical form
   - Identify the assumptions of the model
4. Fit the model! (In R, of course ☺)
5. Evaluate the output
   - Model validation
   - Model selection
6. Interpret the results

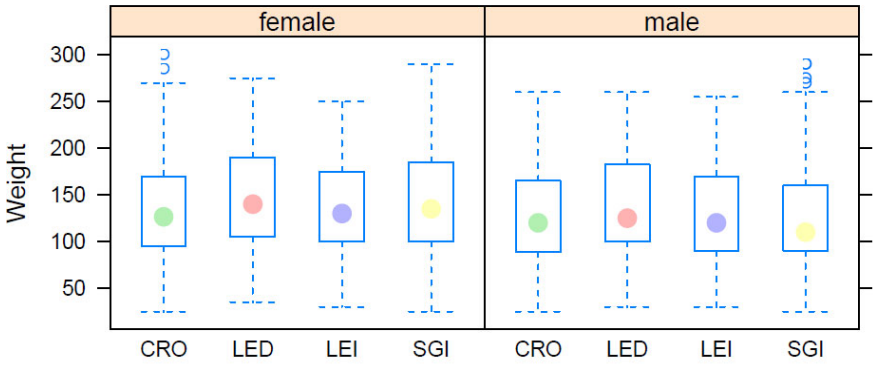
So we already have our question – and we have started some data exploration.

(c) Meg Graham MacLean
mgmaclean@umass.edu

Example

2. Data exploration - conditional box plots

```
> #condition on network AND sex
> bwplot(vole$Weight ~ vole$Network | vole$Sex)
```



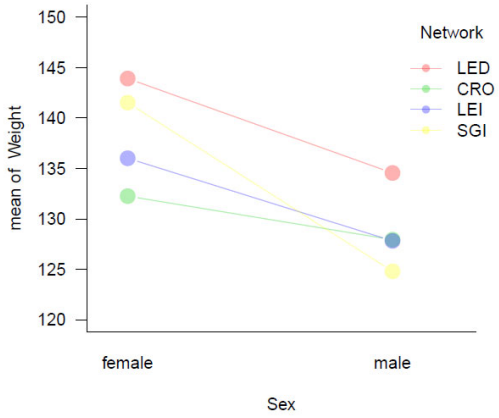
50

Earlier we used this conditional boxplot to look at the different interactions

Example

2. Data exploration – diagnose interactions

```
> interaction.plot(vole$Sex, vole$Network, vole$Weight)
```



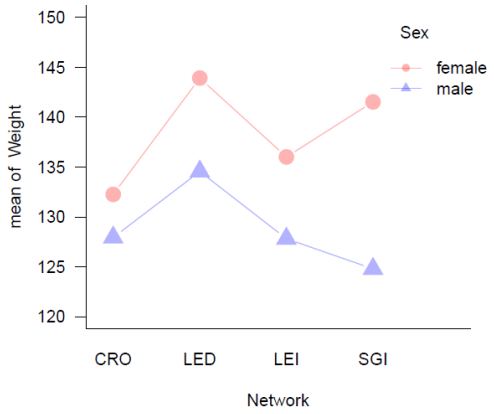
51

But we can also use just the means to see how they differ by sex and site - here it looks like there might be differences, but what if we plotted it with site on the x instead?

Example

2. Data exploration – diagnose interactions

```
> interaction.plot(vole$Network, vole$Sex, vole$Weight)
```



52

It does seem to look like there might be an interaction, especially what we are seeing at SGI

Example

Modeling process:

1. *State the question/hypothesis*
   - *What is the question?*
   - *What are the variables (response and explanatory)?*
2. *Data exploration*
3. Describe the model
   - In word form (should come from your question)
   - In mathematical form
   - Identify the assumptions of the model
4. Fit the model! (In R, of course ☺)
5. Evaluate the output
   - Model validation
   - Model selection
6. Interpret the results

So, let's describe the model and then have you create the code for an interaction model!

(c) Meg Graham MacLean
mgmaclean@umass.edu

Example

3. Describe the model

- In words:
  - Are there significant differences in weight among sex-network combinations?
- As a mathematical model:
  - $y_i = \beta_0 + \beta_{1(g)}Sex_{1i(g)} + \beta_{2(g)}Network_{2i(g)} + \beta_{3(g)}Sex_{1i(g)}Network_{2i(g)} + e_i$
  - Are $\beta_{1(g)}$'s different from 0?
  - Are $\beta_{2(g)}$'s different from 0?
- Assumptions?
  - Residuals are normally distributed
  - Constant variance (homogeneity)
  - Observations are independent
  - Predictors measured without error (fixed X)

Our question is: are there significant differences in water vole weights among sex-network combinations? And we try to figure that out using an interaction model, as described below. Our assumptions for this model are just like any other linear model.

(c) Meg Graham MacLean
mgmaclean@umass.edu

Example

Modeling process:

1. *State the question/hypothesis*
   - *What is the question?*
   - *What are the variables (response and explanatory)?*
2. *Data exploration*
3. *Describe the model*
   - *In word form (should come from your question)*
   - *In mathematical form*
   - *Identify the assumptions of the model*
4. Fit the model! (In R, of course ☺)
5. Evaluate the output
   - Model validation
   - Model selection
6. Interpret the results

Modify the code from the additive model to create an interaction model and do the rest of the modeling process!

So, to round out today, try to modify your code from earlier today to create and interaction model and go through the rest of the modeling process! We will talk about this more on Tuesday.

(c) Meg Graham MacLean
mgmaclean@umass.edu

For next week:

1) Read section 5.2 in the Zuur book (2007).
2) Watch the recorded lecture and complete the exercise
3) OPTIONAL: Read sections 7.1-7.7 in AndyFieldsRBook posted in Moodle R help files.
4) Finish the lab in week 3.
5) Complete the individual assessment on Moodle by 11:55pm Monday night.

Thanks and see you on Tuesday!

(c) Meg Graham MacLean
mgmaclean@umass.edu