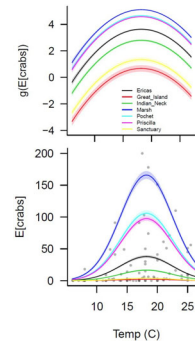# ECO 636
Applied Ecological Statistics

Week 5 – ANCOVA – recorded lecture

Meg Graham MacLean, PhD

Department of Environmental Conservation

mgmaclean@umass.edu

2021 - Spring

Welcome to week 5s recorded lecture!

The Week

Monday
• Extra lab day!

Tuesday
• Collinearity

Wednesday (Lab)
• Two-way ANOVA/regression

Thursday
• ANCOVA

Today we will talk about how to run ANCOVA models, or models with both continuous and categorical variables.

Review!

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~x_1*x_2)
Continuous	Single continuous	Simple linear regression	lm(y~x)
Continuous	Multiple continuous	Multiple linear regression	lm(y~x_1*x_2)

- First section: t-test/ANOVA
  - Response: continuous
  - Explanatory: categorical
  - Predicting group means/contrasts
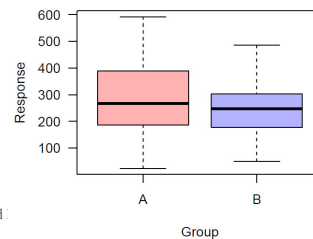
ECO 636 week 5 - ANCOVA - recorded   3

So, for a quick review – we so far have only been dealing with continuous response variables and the first section of the course so far has been to talk about categorical explanatory variables where we are trying to look at group differences or contrasts.
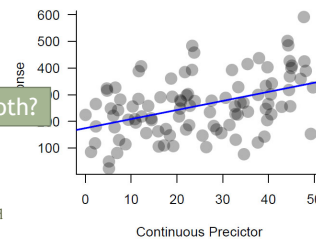
Review!

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~x$_1$*x$_2$)
Continuous	Single continuous	Simple linear regression	lm(y~x)
Continuous	Multiple continuous	Multiple linear regression	lm(y~x$_1$*x$_2$)

- Second section: Regression
  - Response: continuous
  - Explanatory: continuous   But what if we have both?
  - Predicting slopes

ECO 636 week 5 - ANCOVA - recorded   4

Then we went on to talk about regression, where we have continuous explanatory variables and we are predicting the slope or relationship between our explanatory variables and the response. But – what if we have a system where we have both continuous and categorical explanatory variables?

Review!

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)`
Continuous	Single continuous	Simple linear regression	`lm(y~x)`
Continuous	Multiple continuous	Multiple linear regression	`lm(y~x1*x2)`
Continuous	Continuous and factors	ANCOVA	`lm(y~x1*x2)`

- Regression + ANOVA = ANCOVA (analysis of covariance)
  - Estimate intercept and slope(s) *and* compare those parameters between groups
    - Group differences in the intercept (*additive model*) ⬅
   - Group differences in the slope (*interaction model*)

This is the world of ANCOVA or analysis of covariance. In these models we are estimating the intercepts and slopes of our continuous variables, as well as how those parameters may differ between the groups of our categorical variables.

Like in our other models, we can have additive models and interaction models. In the additive model we are only looking at differences in intercepts between groups, while in our interaction model, both the slope and intercept can differ between groups.

Let's start by exploring the additive model and I will have you finish an exercise with an interaction model at the end.

ANCOVA

- Additive ANCOVA
  - Slopes are assumed to be equal (estimates a *single slope* for all groups)
  - Intercepts can vary (*group-specific intercepts*)
    - Tests for differences in the means/intercepts
  - Possible outcomes:
    - No difference between groups
   - Single intercept
      - Single slope
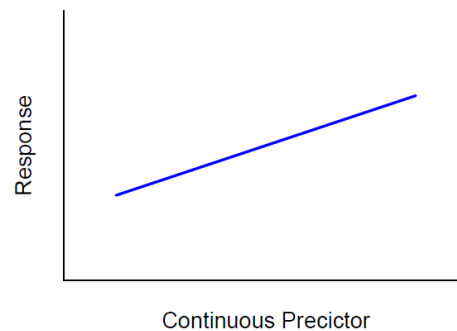
In the additive ANCOVA model, slopes for all groups are assumed to be the same, but the intercepts can be group specific – which means like in our earlier ANOVA models, we can test for differences between group means or intercepts. One of the possible outcomes is this, where we actually don't see a significant difference between groups, meaning there is a single slope and intercept for all groups.

ANCOVA

- Additive ANCOVA
  - Slopes are assumed to be equal (estimates a *single slope* for all groups)
  - Intercepts can vary (*group-specific intercepts*)
    - Tests for differences in the means/intercepts
  - Possible outcomes:
    - Difference between groups
      - g intercepts
      - Single slope
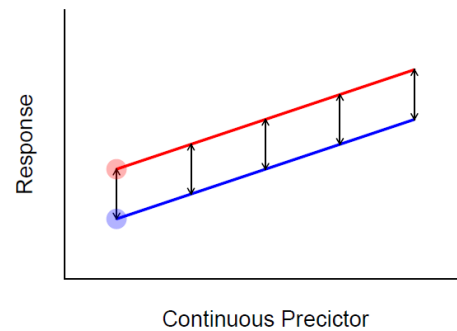
However, we can also find that there is a significant difference between groups, in which case each group may have their own intercept. However, in the additive model, all groups still have the same slope.

$$y_i = \beta_0 + \beta_{cat}X_{cati} + \beta_{cont}X_{conti} + e_i$$

- β_0 is the intercept
  - H_0: $\beta_0 = 0$
- β_{cat} is the difference in intercepts between groups
  - H_0: $\beta_{cat} = 0$, there is no difference in intercepts between groups
- β_{cont} is the slope of the X_{cont} relationship
  - H_0: $\beta_{cont} = 0$, there is no relationship between y_i and X_{cont}

Let's try it with simulated data!

ECO 636 week 5 - ANCOVA - recorded lecture   8

The algebraic model for the additive ANCOVA is vary similar to all of the other models we have looked at so far, but in this simple case we have one categorical explanatory variable and one continuous variable. The null hypothesis that we are testing with the categorical variable is that there is no difference in the intercepts between groups. The null hypothesis for our continuous variable is that there is no relationship between our continuous variable and the response – or that the slope is zero.

So, let's see what this looks like with some simulated data for fun! Feel free to follow along in R if you would like and pause as necessary – or you can just watch this part and then do it yourself later with the interaction model.

ANCOVA

Simulate data:

```
set.seed(123)
covariate <- runif(100,0,50)       #continuous
group <- rep(c("A","B"), each=50)  #categorical
intA <- 100                        #reference intercept
intB <- 150                        #contrast
slopeA <- 6                        #slope
sigma <- 60                        #r.s.e
pars <- c(intA,intB,slopeA)        #parameter vector

dm <- model.matrix(~group+covariate)          #design matrix
response <- rnorm(100,  dm %*% pars, sigma)    #sim data
sim.dat <- data.frame(response,group,covariate)#data frame
plot(response ~ covariate, col=c(1,2))
```

Ok – so if you are following along in R, here is the code to create the data. Things to note are the intA, which is the intercept for our first group, or group A, intB which is actually the contrast, or difference between our reference group A and group B, as well as the slope of our line. These are the parameters we will try to estimate with our ANCOVA model.

ANCOVA

Simulate data:

```
head(sim.dat)
    response group covariate
1   201.4724     A 14.378876
2   334.7787     A 39.415257
3   220.1208     A 20.448846
4   447.0214     A 44.150870
5   368.5939     A 47.023364
6   204.6552     A  2.277825
```

```
tail(sim.dat)
      response group covariate
95    250.0198     B 16.018662
96    274.4529     B  9.384556
97    396.9830     B 39.114715
98    319.3535     B  4.679749
99    516.0402     B 23.338952
100   326.2298     B 25.575273
```
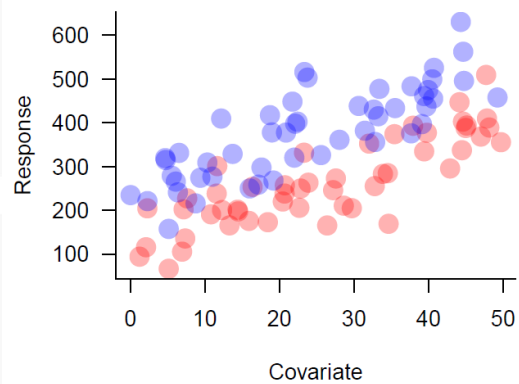
This is what our simulated data look like, with group A in red and group B in blue! Neat!

ANCOVA

Fit the model:

```
Manc <- lm(response ~ group + covariate, data = sim.dat)
```

Now, lets fit our linear model in R with our categorical group variable and continuous covariate variable in order to predict our response.

Fit the model:

```
summary(Manc)

Call:
lm(formula = response ~ group + covariate, data = sim.dat)

Residuals:
     Min       1Q   Median       3Q      Max
-145.840  -36.162   -2.522   32.460  141.152

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 113.2862    13.2572   8.545 1.84e-13 ***
groupB      125.6087    11.4606  10.960  < 2e-16 ***
covariate     5.8269     0.4042  14.417  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 57.14 on 97 degrees of freedom
Multiple R-squared:  0.7592,  Adjusted R-squared:  0.7542
F-statistic: 152.9 on 2 and 97 DF,  p-value: < 2.2e-16
```

How did we do estimating our parameters as compared to what we used to simulate the data?

12

Looking at our summary, how did we do estimating our model parameters??

(c) Meg Graham MacLean
mgmaclean@umass.edu

ANCOVA

Fit the model:

```
intA <- 100   #reference intercept
intB <- 150   #difference betwwen A and B
slopeA <- 6   #effect of covariate on response
sigma <- 60   #r.s.e


coef(Manc)
  (Intercept)      groupB    covariate
   113.286192   125.608734     5.826888


summary(Manc)$sigma
  [1] 57.13765
```

Well, here is a bit of an easier way to look at the results – but to me, it looks like we did ok! I'll point out that the sigma that I used to create the distribution of the points around the mean was 60, so we expect some variation around the model line. So, our intercept, which should be the intercept of group A is estimated at 113, when we used a reference intercept of 100 for our data points. Similarly, our contrast between group B and group A is estimated at 126, rather than our given 150, so a little different, but not outrageous given our sigma. Finally, our slope was estimated pretty well!

ANCOVA

Plot the simulated and estimated slopes/intercepts:

```
(est <- coef(Manc))
  (Intercept)       groupB    covariate
   113.286192   125.608734     5.826888
```

```
plot(response~covariate,col=as.numeric(group))

#real lines
abline(100, 6, lwd=1, lty=2, col=2)
abline(100+150,6,lwd=1,lty=2, col=4)

#fitted lines
abline(est[1], est[3], lwd=1, col=2)
abline(est[1] + est[2], est[3], lwd=2, col=4)
```
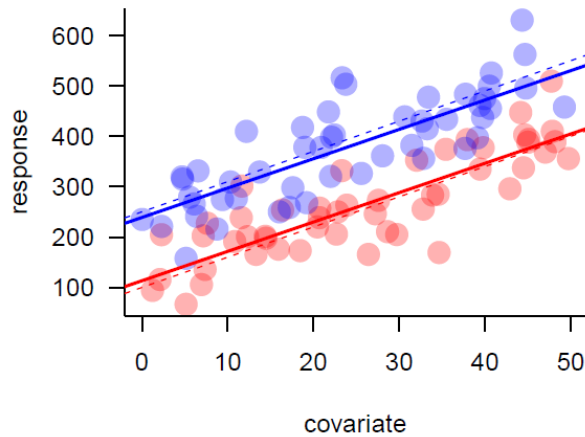
So, let's plot both the input models and the estimated ones. Our "real" lines will be dashed and our fitted or estimated ones will be solid.

ANCOVA

Plot the simulated and estimated slopes/intercepts:

What if we think there are differences in slope between groups?

Neat! Our real versus our modeled lines don't look too bad!

But, let's say we were using real data and we thought there might be some differences in slope between the groups?

ANCOVA

- ANCOVA with interaction!
  - Intercepts can vary (*group-specific intercepts*)
    - Tests for differences in the means/intercepts
  - Slopes also may vary!
    - Tests for differences in the slopes
  - Possible outcomes:
    - No difference between groups
   - Single intercept
      - Single slope (*regression*)
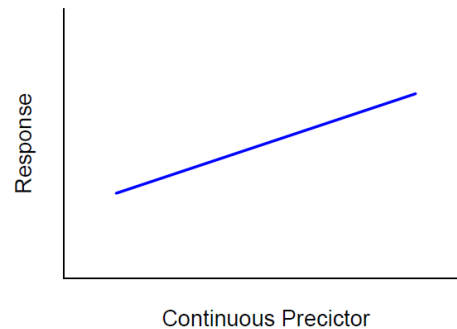
That brings us to ANCOVA with an interaction! So, let's give that a try!

In an ANCOVA with an interaction, not only can our intercepts vary by group – but so can our slopes. Like in the additive model, we may find that there are no differences between groups, in which case we would have just one slope and intercept for all groups, but…

ANCOVA

- ANCOVA with interaction!
  - Intercepts can vary (*group-specific intercepts*)
    - Tests for differences in the means/intercepts
  - Slopes also may vary!
    - Tests for differences in the slopes
  - Possible outcomes:
    - Difference between groups
      - *g* intercepts
   - Single slope (*additive* ANCOVA)
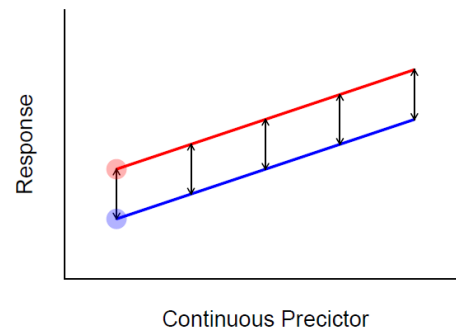
We could also find that we have significantly different intercepts, but not significantly differently slopes, so we would end up with a model that is the same as an additive ANCOVA.

ANCOVA

- ANCOVA with interaction!
  - Intercepts can vary (*group-specific intercepts*)
    - Tests for differences in the means/intercepts
  - Slopes also may vary!
    - Tests for differences in the slopes
  - Possible outcomes:
    - Difference between groups
      - *single* intercept
      - *g* slopes

Response

Continuous Precictor
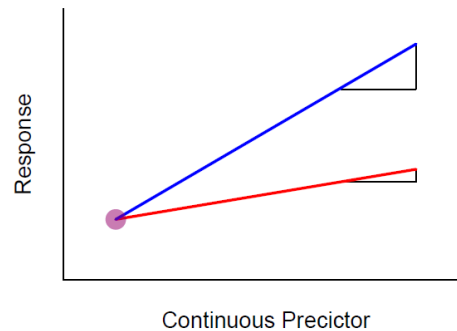
But, we can also find that we have insignificantly different intercepts, while having significantly different slopes – like this!

ANCOVA

- ANCOVA with interaction!
  - Intercepts can vary (*group-specific intercepts*)
    - Tests for differences in the means/intercepts
  - Slopes also may vary!
    - Tests for differences in the slopes
  - Possible outcomes:
    - Difference between groups
      - *g* intercepts
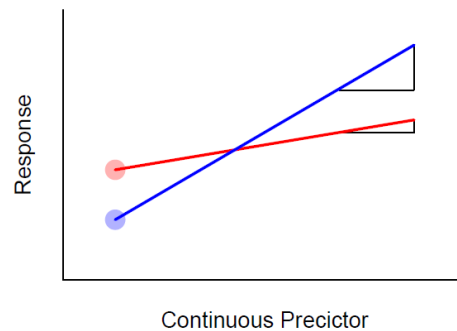   - *g* slopes

Response

Continuous Precictor

Finally, we may find we have significantly different group intercepts and slopes.

ANCOVA

$$y_i = \beta_0 + \beta_{cat}X_{cati} + \beta_{cont}X_{conti} + \beta_{cat:cont}X_{cati}X_{conti} + e_i$$

- β_0 is the intercept
  - H_0: $\beta_0 = 0$
- β_{cat} is the difference in intercepts ($X_{cont} = 0$) between groups
  - H_0: $\beta_{cat} = 0$, there is no difference in intercepts between groups
- β_{cont} is the slope of the X_{cont} relationship
  - H_0: $\beta_{cont} = 0$, there is no relationship between y_i and X_{cont}
- $\beta_{cat:cont}$ is the difference in slopes between groups
  - H_0: $\beta_{cat:cont} = 0$, there is no difference in the slopes

The algebraic model for the interaction ANCOVA is this one – where we add an interaction term. Like in the two-way ANOVA, the interaction beta is the effect of the interaction, so in this case it is the difference in slopes between groups. The null hypothesis is that there is no difference in the slopes.

ANCOVA

$$y_i = \beta_0 + \beta_{cat}X_{cati} + \beta_{cont}X_{conti} + \beta_{cat:cont}X_{cati}X_{conti} + e_i$$

- ANCOVA with interaction!
  - What do the various forms look like in R?

```
lm(y ~ cat + cont + cat:cont) # different intercept and slopes
lm(y ~ cat * cont)            # same as above (shortcut)
lm(y ~ cat:cont)              # same intercept different slopes
lm(y ~ cat + cont)            # same slopes (additive!)
```

Let's try it *together* with simulated data!

ECO 636 week 5 - ANCOVA - recorded lecture   21

In R there are actually a couple of ways to do an interaction ANCOVA , based on whether or not you think there is a difference in intercepts or slopes. If you want to constrain your model to have each group have the same intercept, you can by using the colon instead of the asterisk symbol, and use the additive model to constrain the slopes to be the same.

So, let's try this together, or as an exercise, with simulated data.

ANCOVA

Simulate data:

```r
set.seed(123)
covariate <- runif(100,0,50)       #continuous
group <- rep(c("A","B"), each=50)  #categorical
intA <- 100                        #reference intercept
intB <- 150                        #contrast
slopeA <- 6                        #reference slope
slopeB <- -5                       #dif. in slopes (interaction)
sigma <- 40
pars <- c(intA,intB,slopeA,slopeB) #parameter vector

dm <- model.matrix(~group*covariate)           #design matrix
response <- rnorm(100,  dm %*% pars, sigma)     #simulalte
sim.dat <- data.frame(response,group,covariate) #data frame
plot(response ~ covariate, col=c(1,2))
```

First, like we did before, let's simulate our data! Notice this time I used two slope variables!
One is the slope for group A and the other is the contrast between the slope for group A
and the slope for group B. Try to simulate these data now, and don't forget to set your
seed so we get the same simulated data points!

ANCOVA

Simulate data:

```
head(sim.dat)
    response group covariate
1 196.4060     A 14.378876
2 335.3497     A 39.415257
3 220.9783     A 20.448846
4 419.6493     A 44.150870
5 373.1093     A 47.023364
6 174.3258     A  2.277825


tail(sim.dat)
     response group covariate
95   201.9572     B 16.018662
96   238.1483     B  9.384556
97   230.6445     B 39.114715
98   282.1964     B  4.679749
99   357.3433     B 23.338952
100  224.0941     B 25.575273
```
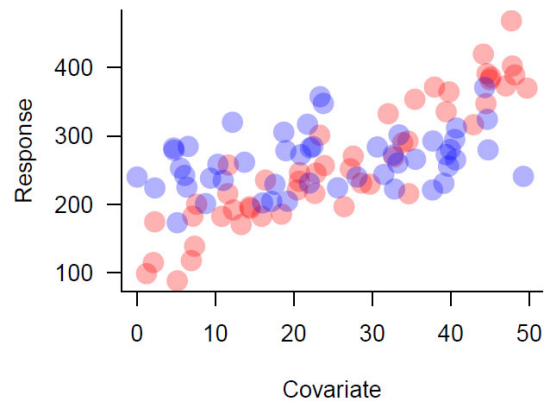

23

Once you have simulated your data, it should look like this!

ANCOVA

Simulate data:

```
head(sim.dat)
    response group covariate
1 196.4060      A 14.378876
2 335.3497      A 39.415257
3 220.9783      A 20.448846
4 419.6493      A 44.150870
5 373.1093      A 47.023364
6 174.3258      A  2.277825
```

What do you think the model matrix looks like?

```
tail(sim.dat)
      response group covariate
95   201.9572      B 16.018662
96   238.1483      B  9.384556
97   230.6445      B 39.114715
98   282.1964      B  4.679749
99   357.3433      B 23.338952
100  224.0941      B 25.575273
```

Given this new data frame, what do you think our model matrix looks like? Remember our model matrix when using categorical data used zeros and ones to keep track of which group an observation was in, but with our continuous data, the model matrix contained the observation data in place of the zeros and ones.

ANCOVA

Model matrix:

$$y_i = \beta_0 + \beta_{cat}X_{cati} + \beta_{cont}X_{conti} + \beta_{cat:cont}X_{cati}X_{conti} + e_i$$

```
head(dm)
    (Intercept) groupB covariate groupB:covariate
  1           1      0 14.378876                0
  2           1      0 39.415257                0
  3           1      0 20.448846                0
  4           1      0 44.150870                0
  5           1      0 47.023364                0
  6           1      0  2.277825                0


tail(dm)
      (Intercept) groupB covariate groupB:covariate
  95            1      1 16.018662        16.018662
  96            1      1  9.384556         9.384556
  97            1      1 39.114715        39.114715
  98            1      1  4.679749         4.679749
  99            1      1 23.338952        23.338952
 100            1      1 25.575273        25.575273
```

So – here is a chunk of our model matrix! Notice that group B is either zero or one based whether or not the observation is in group B – like we had with ANOVA before. But take a look at our interaction term – its full of zeros when the observation is in group A, our reference group, but matches our covariate when the observation is in group B – cool!

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*

> We'll skip this since we simulated our data ☺

4. Fit the model! (In R, of course ☺)
5. Evaluate the output
   - Model validation
   - Model selection
6. Interpret the results

So let's go on to fitting our model. We'll skip the first part of the modeling process since we simulated our data.

ANCOVA

Fit the model:

```
Mint <- lm(response ~ group * covariate, data = sim.dat)
```
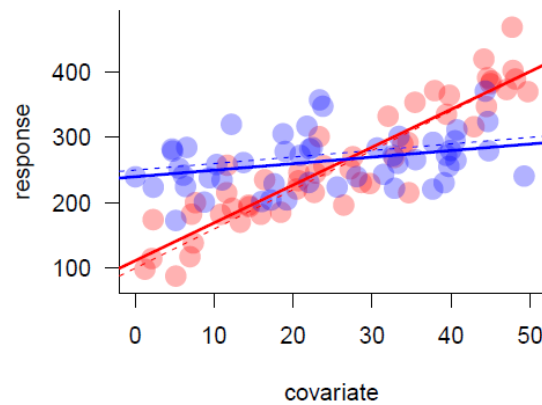
Then take a look and do some model evaluation and interpretation!

Great – so fit the model using a full interaction model

And then try to do some model evaluation and interpretation!

ANCOVA

Try to plot the simulated and estimated slopes/intercepts:

Awesome!
But... what if our y isn't continuous??

While you are at it, try to plot the simulated and estimated slopes and intercepts – it should end up looking something like this. We will talk more about this whole process on Tuesday.

So, this model is pretty awesome and flexible and can cover many different experimental set-ups… but, what if our y or our response isn't continuous? What do we do then?

Review!

So far… y has always been continuous!

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~x`$_1$`*x`$_2$`)`
Continuous	Single continuous	Simple linear regression	`lm(y~x)`
Continuous	Multiple continuous	Multiple linear regression	`lm(y~x`$_1$`+x`$_2$`)`
Continuous	Continuous and factors	ANCOVA	`lm(y~x`$_1$`+x`$_2$`)`

- What if we have:
  - Count data?
 - Binary (yes/no) data?

Generalized linear models!! (GLMs)

What if, for example, we have count data? Or binary, yes/no, data?

Well – here is where GLMs or generalized linear models come in handy!! For much of the rest of the semester we will focus on expanding our models to GLM form, so that we can fit models to a wide variety of data and deal with things like non-normally distributed residuals.
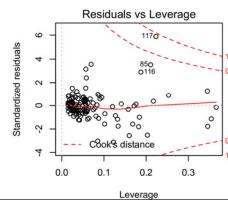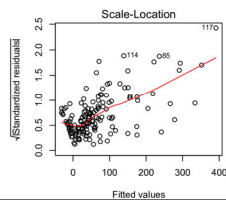
GLMs

What's so nice about GLMs?

- Ecology data often does not fit our assumptions for simple linear models!
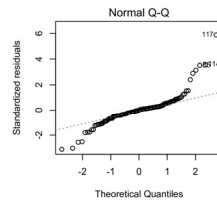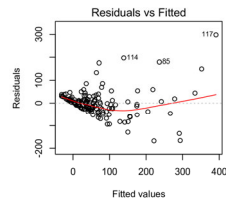
	Generalized linear models
	Includes options for heterogeneous variance!
	Includes options for skewed distributions of residuals!

So, what is so nice about GLMs? Well, ecological datasets often do not fit our assumptions for simple linear models! So, lets look at how GLMs differ from general linear models like we have been doing so far…

…

So, if we plot our diagnostic plots and see all this craziness, we can address some of it by using GLMs rather than general linear models!

GLMs

What's so nice about GLMs?

- Ecology data often does not fit our assumptions for simple linear models!

General linear models	*Generalized* linear models
Assumes homogeneity of variance in residuals	Includes options for heterogeneous variance!
Assumes normally distributed residuals	Includes options for skewed distributions of residuals!
Response is continuous	Response can be binary (or other)
Response can take any values (-∞ to ∞)	Response can be bounded

How does GLM incorporate these options?

Finally, general linear models also assume things like…

So, how does a GLM incorporate these options??

GLMs

What's so nice about GLMs?

- GLMs use a "link" function to connect $\bar{y}$ with its linear predictor (i.e., how the function/model is related to the data)

$$\mu \leftrightarrow \eta$$

- μ is the data mean, η is the linear predictor
- What are some options of linking (these are not all!)?

Link	Distribution describing link	Data types
"identity"	Gaussian (normal!)	Continuous
"log"	Poisson or negative binomial	Counts
"logit"	Binomial	Binary/proportional
"inverse"	Gamma	Counts
"inverse gaussian"	Inverse gaussian	Counts

Well, GLMs use a link function to connect our response with a linear predictor, which is basically saying that we can use a non-linear function to connect a linear predictor – just like the ones we have been using – to non-linear data! Which is super awesome.

So, what are some of the options for a link function? The first is the one we have been using in lab – the identity link – which assumes a gaussian or normal distribution of the residuals and works for a continuous predictor variable. This first link is equivalent to using a simple linear model, like we have been doing in class thus far. We also have many other options for a link – like a log link which works well for count data and assumes a Poisson distribution for the residuals, or a logit link for binary data that assumes a binomial distribution of residuals. We'll go through each of these in more detail later – I just wanted to introduce them here.

GLMs
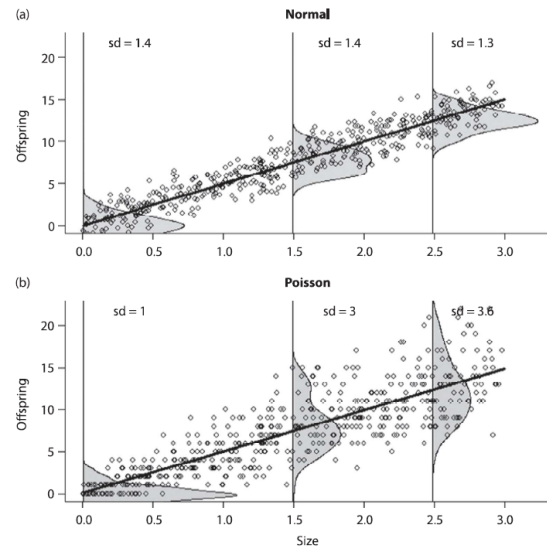
Fox et al., (2015)

For a visual of how these links can differ – on the top is where we assume a normal distribution of the residuals – just like we have in all of our simple linear models. On the bottom though is a GLM with a log link – where we assume a Poisson distribution of the residuals, and what is cool about a Poisson distribution is that the distribution can change with X – so you can see how the shape of the distribution changes from size of zero to size of 3 in this example.

GLMs

What's so nice about GLMs?

- GLMs use a "link" function to connect $\bar{y}$ with its linear predictor (i.e., how the function/model is related to the data)

$$\mu \leftrightarrow \eta$$

- μ is the data mean, η is the linear predictor
- What are some options of linking (these are not all!)?

Link	Distribution	Data types
"identity"	Gaussian (normal!)	Continuous
"log"	Poisson or negative binomial	Counts
"logit"	Binomial	Binary/proportional
"inverse"	Gamma	Counts
"inverse gaussian"	Inverse gaussian	Counts

Ok – great! So we will dive deeper into these link functions and how they work in GLMs next week, with a particular focus on these first three since they are the most often used link functions, but there are a ton of different link functions, since there are a ton of different ways data can be non-normal!

For next week:

1) Read Fox Ch. 6 for an intro to GLMs
2) Watch the recorded lecture and do the exercise
3) Finish the two-part lab exercise for this week on regression and review of two-way ANOVA
4) Complete the individual assessment on Moodle by 11:55pm Monday night.

Thanks and see you on Tuesday!