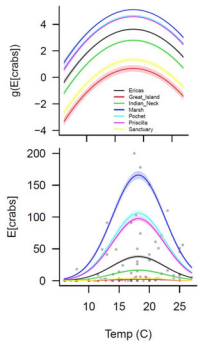
ECO 636
Applied Ecological Statistics

Week 2 – Linear Model Intro – two samples – recorded lecture



Meg Graham MacLean, PhD

Department of Environmental Conservation

mgmaclean@umass.edu

2021 - Spring

Welcome to week 2s recorded lecture!

The Week

Tuesday

- Review of data exploration
- Modeling process!
- Basics of a linear model
  - Null model - tree example

Wednesday (Lab)

- Data exploration

Thursday (asynchronous)

- More linear models (two groups)

Here we will expand on what we learned on Tuesday with our null linear model and move on to talk about linear models when we have two sample groups!

(c) Meg Graham MacLean
mgmaclean@umass.edu

Review!

So far…

Response (Y)	Explanatory (X)	Model	In R
Continuous	None	Intercept-only/null	`lm(y~1)`

Single sample, only a response variable (e.g., mean tree height)

- predicting based on the mean

So, let's quickly review what we talked about on Tuesday – on Tuesday we explored a null model, or a linear model where we don't have an explanatory variable – all we measured was our response variable. For example, we have a stand of trees and we want to make the best possible prediction for the height of a tree within that stand, we would do so based on the mean of the other trees in that stand.

(c) Meg Graham MacLean
mgmaclean@umass.edu

Two samples!

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

Next!

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

Two samples, where data collected is associated with membership in one of two groups (e.g., tall vs. short, stand 1 vs. stand 2)

Compare the population means = *t*-test as a linear model!

- H_0 = no difference between sample means
- H_1 = sample means differ

ECO 636 week 2 - Linear Model Intro - recorded lecture 4

But what if instead we had two sample groups? Like if we extend our earlier example to have two tree stands rather than one. We no long are using the null model and we may find that the means for our two sample groups are different.

Really, this version of a linear model is just a reframing of the t-test – where our null hypothesis is that there is no difference between sample means, and the alternative hypothesis is that there is a difference.

(c) Meg Graham MacLean
mgmaclean@umass.edu

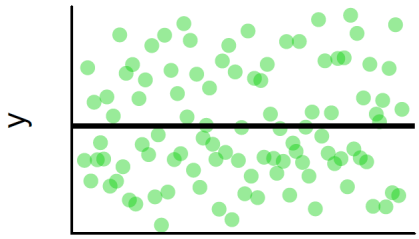
Review!

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

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

$$y_i = \beta_0 + e_i$$

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



5

So here – I am going to have you try to answer each of these questions for yourself before I click next – just to try to get your brain to engage a little in the recorded lecture – I know it's a challenge.

So, from Tuesday – what does the null model look like mathematically?

Great – hopefully you got this – where we have our beta naught as the intercept and e is the stochastic part of our linear model.

Similarly, what does the null model look like graphically?

Again, it's a linear model that just predicts the mean, so it looks like this!

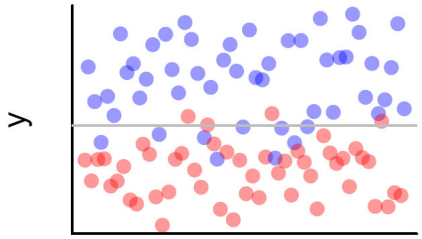
Review!

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

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?



6

Now, let's build on that. What does the t-test or two-level factor explanatory variable version of a linear model look like mathematically?

In this version in addition to beta naught, we also have beta one and an x our explanatory variable.

So, what does this look like graphically?

Well, now instead of having one sample, we have two sample groups, where our x or explanatory variable is either group 1 or group 2 – shown here as either blue or red dots – where red is group 1 and blue is group 2. Since we have no other explanatory variable other than those groups, we can plot these dots however we want, like this as just a random scatter…

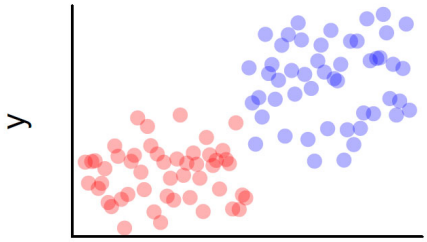
Review!

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

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?



7

Or separated by group. If I separate them by group, you can start to see that it looks like the means between the two groups look different. If we continue to imagine our tree stand example from last class – it looks like the trees in stand two, or the blue dots, are much taller than the trees in stand one, or the red dots. So how would we use our linear model equation to test that?

(c) Meg Graham MacLean
mgmaclean@umass.edu

X is a two-level factor

Describe the model in word form:

- Are the two means different?

Describe the model in mathematical form:

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

- β_0 is the intercept
- β_1 is the 'slope', or in this case the difference between 2 means, aka 'contrast'
- X_i is a two-level factor, either 0 or 1
  - (e.g., in stand 2 = 1, not in stand 2 = 0)
- $e_i \sim N(0, \sigma)$, σ is the estimated standard deviation of the mean

So, just to keep in mind what we are looking at, our model in word form is asking the question: are the two means different, and like we just saw – here is our model in mathematical form. So what do each of the different pieces of this model represent to help us answer that question?

Well, as we learned in our null model – beta naught is the intercept, just like before.

Beta one in any linear model is considered the 'slope', but in this case it is actually the difference between the two means, or also known as the contrast between means. I'll show you how that works on the next slide.

Next, X is the explanatory variable, or in this case the two-level factor based on the group. For our example, zero represents stand one, or basically that it isn't in stand two, and one represents stand two. R and most other statistical software packages will always assign zero to the first alpha-numeric group, so in this case the number 1, or if I had named the stands it would assign the zero to the first name alphabetically. Then the one represents the other group.

Finally, the e is the estimated standard deviation of the mean. So, let's work through why beta one is the contrast between group means when x is a two-level factor!

X is a two-level factor

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

When X_i is 0, what does the deterministic model look like?

$$y_i = \beta_0 + \beta_1 * 0$$
$$y_i = \beta_0$$

So, β_0 is the mean of the first group!

When X_i is 1, what does the deterministic model look like?

$$y_i = \beta_0 + \beta_1 * 1$$
$$y_i = \beta_0 + \beta_1$$

So, $\beta_0 + \beta_1$ is the mean of the second group and β_1 is the difference (aka contrast)!

Let's start with our mathematical model.

If our sample is from stand one, x should be zero. If x is zero, what should our deterministic model look like?

Let's plug in zero for x and solve, and we should end up with just beta naught as our deterministic model for all of our sample group one – neat this is the null model! That means that beta naught is the mean of the first group or stand.

So, let's try that again with stand 2. That means X is one – so what does our deterministic model look like for stand two?

Similarly, we'll plug in one for x and solve, and we end up having beta naught plus beta one equal to the mean for stand two – meaning, the difference between group one and group two will always be beta one! That is why beta one is known as the contrast when x is a factor or categorical variable.

(c) Meg Graham MacLean
mgmaclean@umass.edu

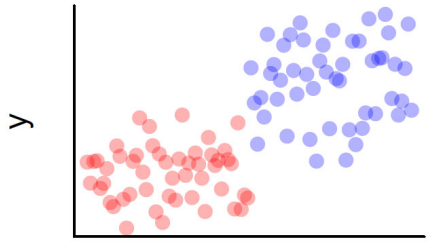
Review!

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

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?



10

Back to what this looks like graphically, if we have our two sample groups plotted like this… lets also plot the means and what they are in our model.

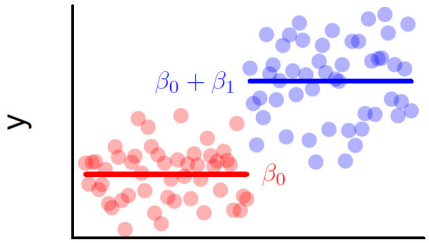
Review!

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

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?



11

So here, the red line is the mean for group one, and is beta naught in our linear model and the blue line is the mean of group two and is beta naught plus beta one in our linear model.

Review!

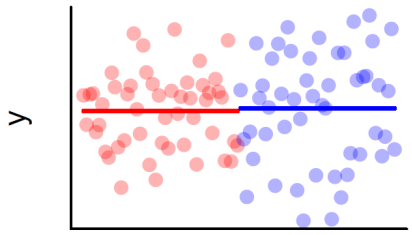
Response (Y)	Explanatory (X)	Model	In R
Continuous	None	Intercept-only/null	`lm(y~1)`
Continuous	Two-level factor	*t*-test	`lm(y~x)`

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?

But what if $\beta_1 \approx 0$?



12

Hm… but let's think about what it would look like if our means were very similar? Basically, if beta one was very close to zero??

This all comes back to our t-test! If beta one is close to zero, there is likely an insignificant difference between our groups - we'll do an example later to see how we determine if there is a significant difference!

Multiple samples!

So far…

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

But first… let's get a little more complicated. If a two-level factor linear model is equivalent to a t-test…

(c) Meg Graham MacLean
mgmaclean@umass.edu

Multiple samples!

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

Next!

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)`

More than 2 samples, where data collected is associated with membership in one of many groups (e.g., treatment 1, 2, 3, or 4)

Compare the group means = ANOVA as a linear model!

- H_0 = no difference between group means
- H_1 = group means differ

An explanatory variable that is a multi-level factor with more than two groups is equivalent to an ANOVA.

Just like we saw with the two-level factor, the null hypothesis is that there is no difference between the group means, and the alternative is that the group means differ. So let's look at what ANOVA looks like as a linear model!

(c) Meg Graham MacLean
mgmaclean@umass.edu

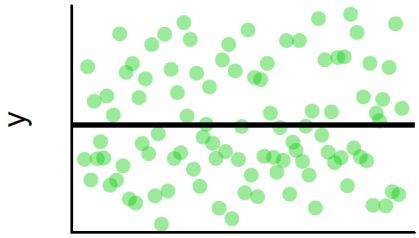
Review!

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?



15

Quick review – here is our null model both mathematically and graphically…

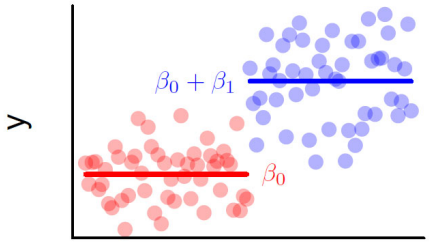
Review!

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?



16

Here is our two-level factor model both mathematically and graphically…

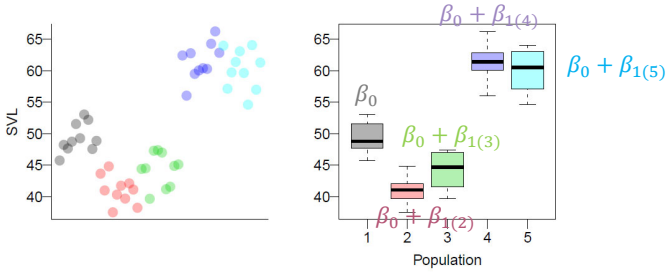
Review!

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?



17

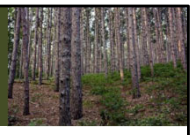
And finally, think for a moment what you think our multiple-level factor model might look like mathematically…

It looks strikingly similar to our two-level factor, but notice that I added a g subscript to the beta one and X, this is because I now have multiple groups, rather than one and I now need to keep track of multiple beta ones and Xs.

Looking at these example data, I have created 5 unique sample groups shown in the different colors… therefore g would go from 1 through 5. So, what would the deterministic representations of the means be for each of these groups? Well, like in our two-factor example, the mean for the first group is going to just be beta naught because X for group 1 is set to 0. This first group is often referred to as our reference group, since all other representations of the mean also have beta naught in them, or the mean from this reference group. Then, for each of the following groups the mean is beta naught plus the unique beta one for that group.

So, to make sure we have down the basics of linear models when our explanatory variable is a factor, lets do an example that has just two levels. We'll try a multiple-level factor example together on Tuesday.

Example



Let's try an example with simulated data (where we know the mean and standard deviation):

- We will simulate 'tree diameter' from two (fake) tree stands (A and B) and we want to know if there is a difference in DBH between the stands.

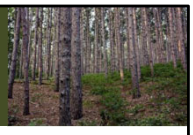
```
> #Simulate DBH data
> set.seed(123)
> mu.dbh <- rep(c(70, 100),each=25)        #deterministic part
> Stand <- gl(n=2,k=25,labels=c("A","B")) #make a two level factor
> DBH <- rnorm(50, mu                      art N(mean,sigma)
> tree <- data.frame(DBH=DBH,Stand=Stand) # make the data frame
```

Please complete the posted exercise in R!

Our example will build on the example we started on Tuesday and simulate two tree stands with different diameters at breast height – which is a more common measure of tree size. I have posted a script for you to run on Moodle to complete this exercise. It is an R markdown file to get you started using this type of R file. If you haven't used this file type before, you use the green arrows at the top of each R chunk to run the script in that chunk. Please reach out if you have questions.

(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

In the R markdown file I will go through all of the modeling process – please walk through it taking notes and making sure you understand each part! I also have a few questions to think about as you go through the code. We'll chat about those on Tuesday.

(c) Meg Graham MacLean
mgmaclean@umass.edu

For next week:

1) Finish reading Ch. 5.1 in the Zuur et al. (2007) book
2) Watch the recorded lecture and do the exercise
3) Finish the posted Week 2 lab
4) Please bring questions to class on Tuesday, as we will recap simple linear regression and move to more complex models!
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