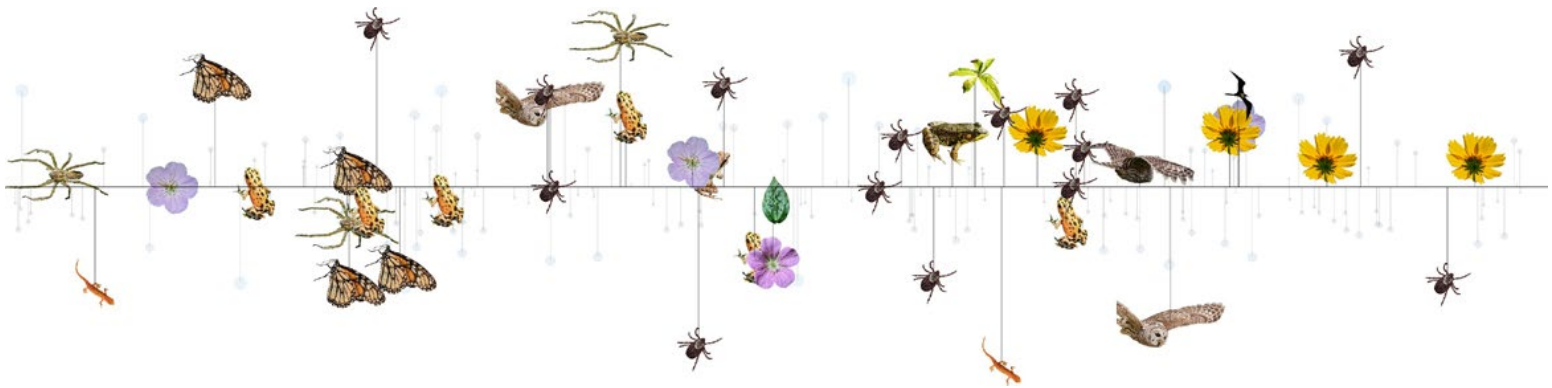
Intro to Quantitative Ecology

Deck 10B - Regression 4

Michael France Nelson

Spring 2023



Overview

For Today

- Regression with continuous and categorical predictors
- Questions and Answers
- Group time for water vole assignment

For Next Week

- ANCOVA, Interactions, Intro to Frequentism, Confidence Intervals

This deck's concepts

- Regressions with continuous and categorical predictors
  - Analysis of Covariance: ANCOVA
- Dummy Variables
- Base cases and model coefficients

Regression and ANOVA

Regression with Numerical and Categorical Predictors

What happens if we have data that contain a mix of numerical and categorical predictor variables?

- Take a look at the penguin data:

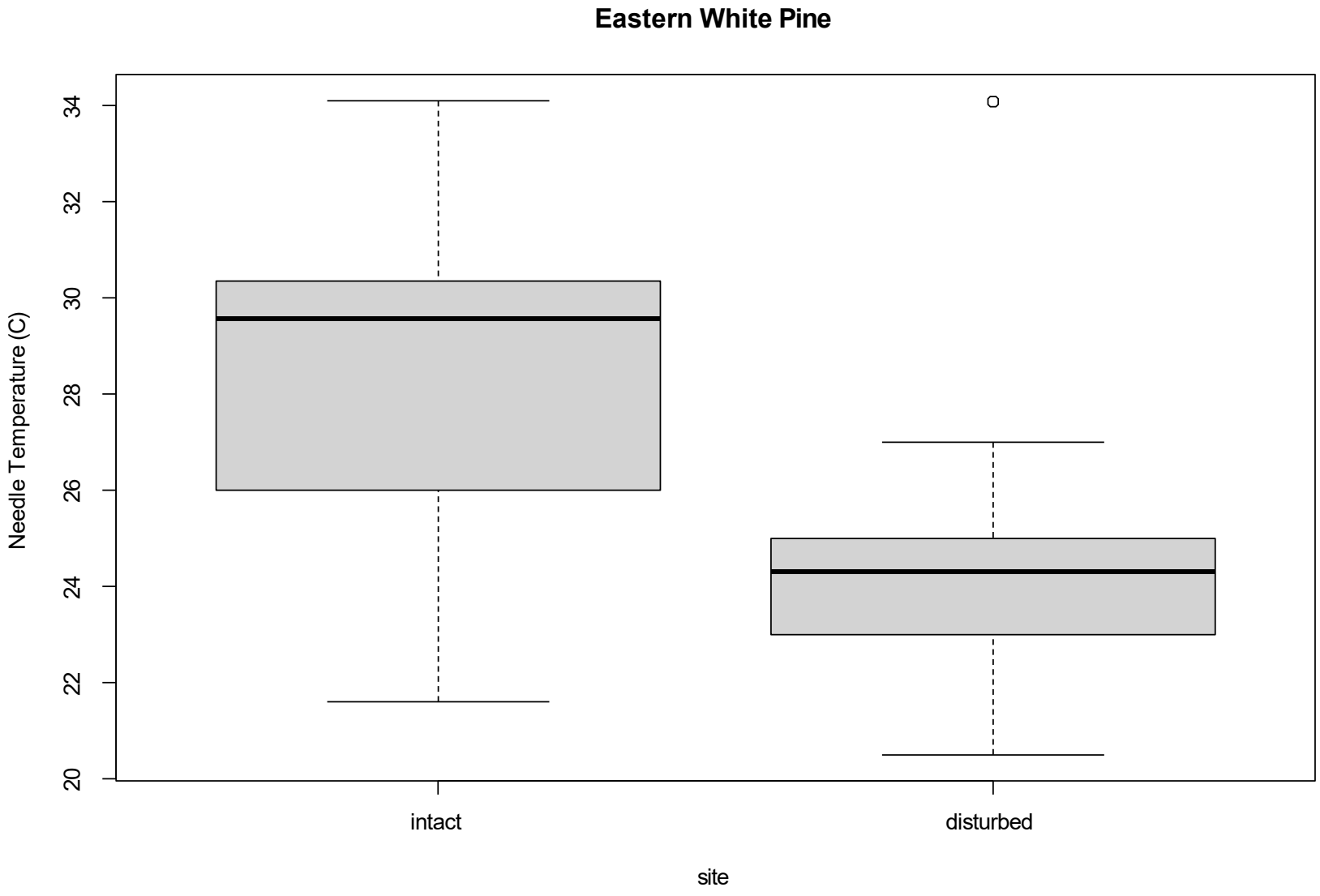
species	island	bill_depth_mm	body_mass_g	sex
Adelie	Torgersen	18.7	3750	male
Adelie	Torgersen	17.4	3800	female
Adelie	Torgersen	18.0	3250	female
Adelie	Torgersen	NA	NA	NA
Adelie	Torgersen	19.3	3450	female
Adelie	Torgersen	20.6	3650	male

Eastern White Pines

We have data from a study on eastern white pines from the University of Michigan biological station.

- The objective was to determine whether the needle temperature varied between intact and disturbed forest sites.
- The temperature of needles is important for numerous ecophysiological processes including transpiration and photosynthesis/photorespiration rates.
  - Photorespiration can occur in overheated or stressed plants causing a loss of energy.
- It is a rich dataset, allowing us to address multiple research questions.

Graphical Exploration



Numerical exploration

Let's take a peek at a summary of the data:

```
summary(dat_white_pine)
```

DBH	HEIGHT	Needle_Temp	TREAT
Min. : 0.800	Min. : 1.400	Min. :20.50	intact :30
1st Qu.: 2.420	1st Qu.: 2.300	1st Qu.:23.60	disturbed:33
Median : 3.700	Median : 3.000	Median :25.20	
Mean : 4.847	Mean : 3.617	Mean :26.18	
3rd Qu.: 6.040	3rd Qu.: 4.750	3rd Qu.:29.56	
Max. :22.450	Max. :10.000	Max. :34.10	

ANOVA

We could use a t-test on the two site types:

```
t.test(Needle_Temp ~ TREAT, data = dat_white_pine)

        Welch Two Sample t-test

data:  Needle_Temp by TREAT
t = 5.7968, df = 54.177, p-value = 0.0000003569
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 2.750419 5.658490
sample estimates:
    mean in group intact mean in group disturbed
                28.38567                 24.18121
```

ANOVA

But… our data set contains other information.

What if tree height is also an important factor?

- If so, it could improve our model.

We want something like a multiple regression using:

- disturbance treatment (a factor)
- tree height (a number)

However we are mixing categorical and numeric variables!

ANCOVA

When we build a regression model using a mix of categorical and numeric data, it is referred to as **Analysis of Covariance**: ANCOVA.

WE can build an *additive* model of needle temperature as predicted by tree height and site disturbance.

The syntax in R is identical to a multiple regression:

```
fit1 = lm(Needle_Temp ~ HEIGHT + TREAT, data = dat_white_pine)
```

- Recall that TREAT is the disturbance factor.

ANCOVA Model Table

Let's interpret the model table:

```
Call:
lm(formula = Needle_Temp ~ HEIGHT + TREAT, data = dat_white_pine)

Residuals:
     Min       1Q   Median       3Q      Max
-6.8457 -1.8109  0.6136  1.6649  9.9555

Coefficients:
                 Estimate Std. Error t value             Pr(>|t|)
(Intercept)       28.7968     0.8644  33.315 < 0.0000000000000002 ***
HEIGHT            -0.1091     0.1831  -0.596                0.553
TREATdisturbed    -4.2357     0.7219  -5.867          0.000000204 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.854 on 60 degrees of freedom
Multiple R-squared:  0.3648,    Adjusted R-squared:  0.3436
F-statistic: 17.23 on 2 and 60 DF,  p-value: 0.000001224
```

Interpreting the Model Coefficients

Let's look at just the coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	28.7967651	0.8643722	33.315238	0.0000000
HEIGHT	-0.1091412	0.1830896	-0.596108	0.5533454
TREATdisturbed	-4.2357086	0.7219018	-5.867431	0.0000002

Can we interpret the *HEIGHT* coefficient?

The Disturbance Coefficient

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	28.7967651	0.8643722	33.315238	0.0000000
HEIGHT	-0.1091412	0.1830896	-0.596108	0.5533454
TREATdisturbed	-4.2357086	0.7219018	-5.867431	0.0000002

Can we interpret the disturbance coefficient?

- First, why does it have a funny name?
  - **TREATdisturbed** looks like a combination of the column name and one of its factor levels!

```
levels(dat_white_pine$TREAT)
[1] "intact"    "disturbed"
```

Announcements

- Default group – We cannot grade
  - Grades will convert to zeroes at end of semester
- We can't grade RMd file or R scripts (unless we specifically ask for them)
  - You should render your RMds or compile things into a document
- Office hours
  - Ana and myself are available if you can't make the posted office hours
  - It's an easy 5% of your grade!

Coding Factors: Dummy Variables and the Design Matrix I

The Data

As a reminder, the White Pine data look like:

	DBH	HEIGHT	Needle_Temp	TREAT
1	12.10	7	29	intact
40	10.42	8	23	disturbed

Note that TREAT has two levels: “intact”, and “disturbed”.

Dummy Variables

We could define a new column named "TREATdisturbed" that contains 0 values if the observation is "intact", or 1 if "disturbed":

```
dat_white_pine$TREATdisturbed  =
  as.numeric(dat_white_pine$TREAT  ==  "disturbed")
```

You can think of the new column as a new numeric predictor for which we only have observations of 1 and 0.

	DBH	HEIGHT	Needle_Temp	TREAT	TREATdisturbed
1	12.10	7	29	intact	0
40	10.42	8	23	disturbed	1

Question: In our new column, what is the **base case**?

Dummy Model Fit

To prove to ourselves that R is doing this behind the scenes, let's fit a model using our new predictor variable:

```
fit2 = lm(
  formula = Needle_Temp ~ HEIGHT + TREATdisturbed,
  data = dat_white_pine)
```

Dummy Model Summary

Here's the model summary for the new fit:

```
Call:
lm(formula = Needle_Temp ~ HEIGHT + TREATdisturbed, data = dat_white_pine)

Residuals:
     Min       1Q   Median       3Q      Max
-6.8457 -1.8109  0.6136  1.6649  9.9555

Coefficients:
                Estimate Std. Error t value             Pr(>|t|)
(Intercept)      28.7968     0.8644  33.315 < 0.0000000000000002 ***
HEIGHT           -0.1091     0.1831  -0.596                0.553
TREATdisturbed   -4.2357     0.7219  -5.867          0.000000204 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.854 on 60 degrees of freedom
Multiple R-squared:  0.3648,    Adjusted R-squared:  0.3436
F-statistic: 17.23 on 2 and 60 DF,  p-value: 0.000001224
```

Comparing the Factor and Dummy Models

We'll check that the coefficient values are the same for the two models:

The factor variable model:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	28.7967651	0.8643722	33.315238	0.0000000
HEIGHT	-0.1091412	0.1830896	-0.596108	0.5533454
TREATdisturbed	-4.2357086	0.7219018	-5.867431	0.0000002

The dummy variable model:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	28.7967651	0.8643722	33.315238	0.0000000
HEIGHT	-0.1091412	0.1830896	-0.596108	0.5533454
TREATdisturbed	-4.2357086	0.7219018	-5.867431	0.0000002

The Design Matrix

When R builds a model, it can only work with numeric data.

Behind the scenes, R makes a new numeric column out of the factor column so that it can perform mathematical operations.

To build the model, R extracts the columns that we specify in the model formula:

	Needle_Temp	TREATdisturbed	HEIGHT
1	29.00	0	7.0
2	29.00	0	2.5
3	29.20	0	3.0
4	30.00	0	5.0
5	30.35	0	2.5
6	31.00	0	1.7

The Design Matrix

Notice the similarity to the regression equation:

- Needle_Temp is the y
- TREATdisturbed is x1
- HEIGHT is x2

$$NeedleTemp = \beta_0 + \beta_1 * TREATdisturbed + \beta_2 * HEIGHT$$

Or more formally:

$$y_i = \beta_0 + \beta_1 * x1_i + \beta_2 * x2_i + E$$

ANCOVA Model

```
Call:
lm(formula = Needle_Temp ~ HEIGHT + TREAT, data = dat_white_pine)

Residuals:
     Min       1Q   Median       3Q      Max
-6.8457 -1.8109  0.6136  1.6649  9.9555

Coefficients:
                Estimate Std. Error t value             Pr(>|t|)
(Intercept)      28.7968     0.8644  33.315 < 0.0000000000000002 ***
HEIGHT           -0.1091     0.1831  -0.596                0.553
TREATdisturbed   -4.2357     0.7219  -5.867          0.000000204 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.854 on 60 degrees of freedom
Multiple R-squared:  0.3648,    Adjusted R-squared:  0.3436
F-statistic: 17.23 on 2 and 60 DF,  p-value: 0.000001224
```

ANOVA and Regression

The concept of **dummy variables** is the key to understanding the connection between linear regression and ANOVA!

The pine data had only two disturbance treatments, and 1 dummy variable.

- How many dummy variables would we need if there were 3 treatments?

Categorical Variables and Model Coefficients

Model Coefficientables for Categorical Predictors

Let's step back and build a linear model of pine needle temperature predicted only by disturbance:

```
Call:
lm(formula = Needle_Temp ~ TREAT, data = dat_white_pine)

Residuals:
     Min       1Q   Median       3Q      Max
-6.7857 -1.7312  0.6143  1.6143  9.8988

Coefficients:
                  Estimate Std. Error t value             Pr(>|t|)
(Intercept)        28.3857     0.5183  54.763 < 0.0000000000000002 ***
TREATdisturbed     -4.2045     0.7162  -5.871          0.000000193 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.839 on 61 degrees of freedom
Multiple R-squared:  0.361,  Adjusted R-squared:  0.3505
F-statistic: 34.46 on 1 and 61 DF,  p-value: 0.0000001926
```

Comparing Model Cofficients and the ANOVA Table I

The ANOVA table for our needle temperature ~ disturbance model:

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
TREAT	1	277.7883	277.78831	34.46467	0.0000002
Residuals	61	491.6655	8.06009	NA	NA

We've worked with ANOVA tables before.

- what are the among- and within-group degrees of freedom?
- Does disturbance status seem important?

Comparing Model Cofficients and the ANOVA Table II

Now the model coefficients

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	28.385667	0.5183335	54.763322	0.0000000
TREATdisturbed	-4.204454	0.7161807	-5.870661	0.0000002

How do we interpret them?

Categorical Predictor Base Cases

Note that the model coefficient for disturbance treatment contains the text disturbed.

- This tells us something important: the *intact* level of disturbance is a **base case**.
- Base cases are represented as *intercept* terms in models with categorical predictors.
  - The base factor level determines how R calculates the intercept and slope:
  - R considers the first factor level to be the *base case*

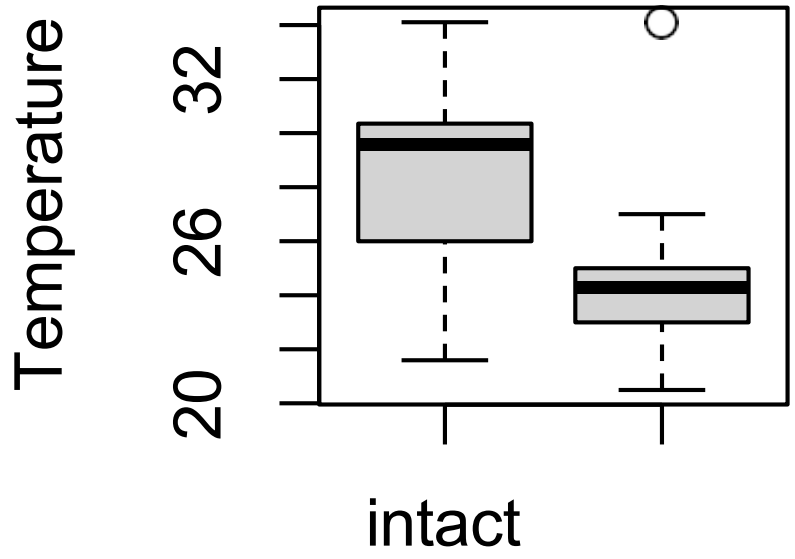
```
levels(factor(dat_white_pine$TREAT))
[1] "intact"    "disturbed"
```

What is the base case for penguin species?

Base Case Graphical Intuition I

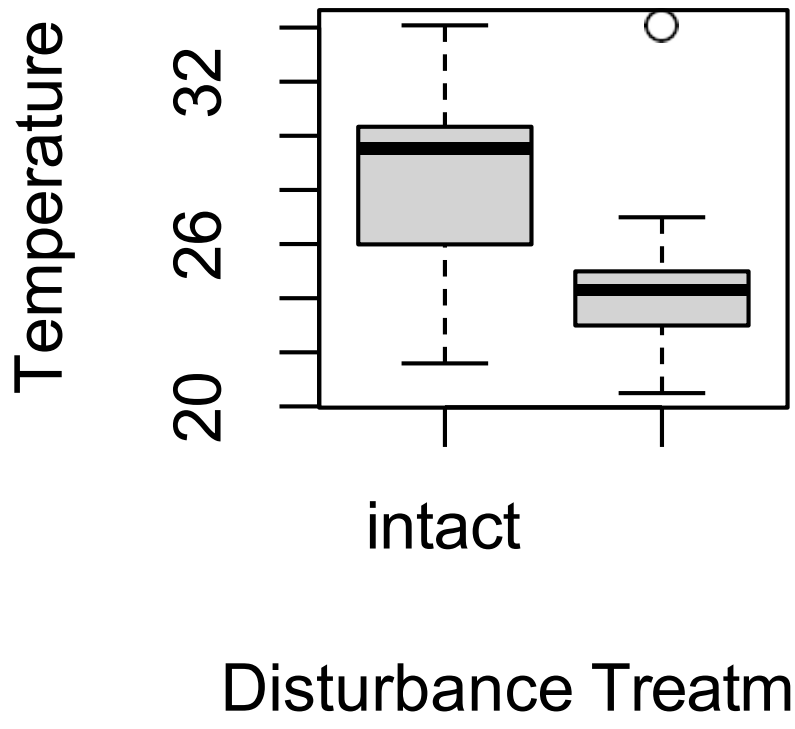
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	28.385667	0.5183335	54.763322	0.0000000
TREATdisturbed	-4.204454	0.7161807	-5.870661	0.0000002

It looks like the intercept corresponds to the mean needle temperature of trees in intact sites:



Base Case Graphical Intuition II

- It looks like the the temperature in disturbed sites is about 4 degrees cooler than in intact sites. This is the value of the *TREATdisturbed* slope coefficient!



Base Case and Slope Interpretation

You can understand the model coefficients for a predictor variable as:

- The base case (the intercept) is the mean value of observations within the *base group*.
- The slope tells you the difference between the base case and the other groups.

This makes sense when we think in terms of dummy variables.

Categorical Variable With 3 Levels

The penguins data set has three species: Adelie, Chinstrap, and Gentoo. What is the base penguin species?

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3700.66225	37.61935	98.3712321	0.0000000
speciesChinstrap	32.42598	67.51168	0.4803018	0.6313226
speciesGentoo	1375.35401	56.14797	24.4951686	0.0000000

How heavy are Gentoo penguins?