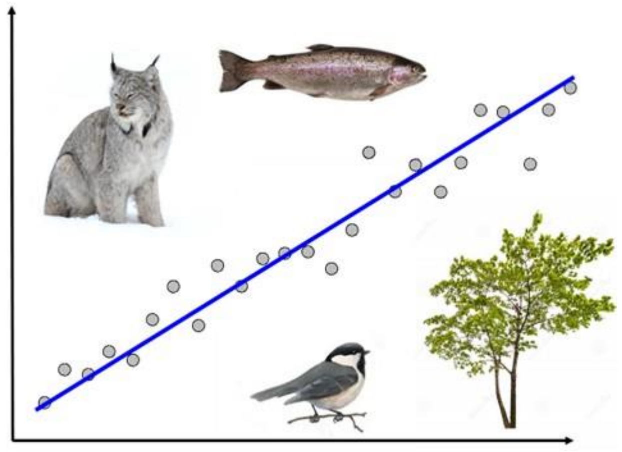
NRC 290b Introduction to Quantitative Ecology

Week 10 – Regression



Meg Graham MacLean, PhD

Department of Environmental Conservation

mgmaclean@umass.edu

2019 - Fall

This week

Monday

- Regression
  - Multiple regression
    - How does this relate to other tests?

Wednesday

- Group exercise
  - Voles! Multiple regression

Statistical testing

When doing a regression – what is the H_0 the p-value is testing?

a) The slope is no different from 0
b) The explanatory variable is no different than the response variable
c) The explanatory variable is significantly correlated with the response variable
d) There is no significant difference between groups

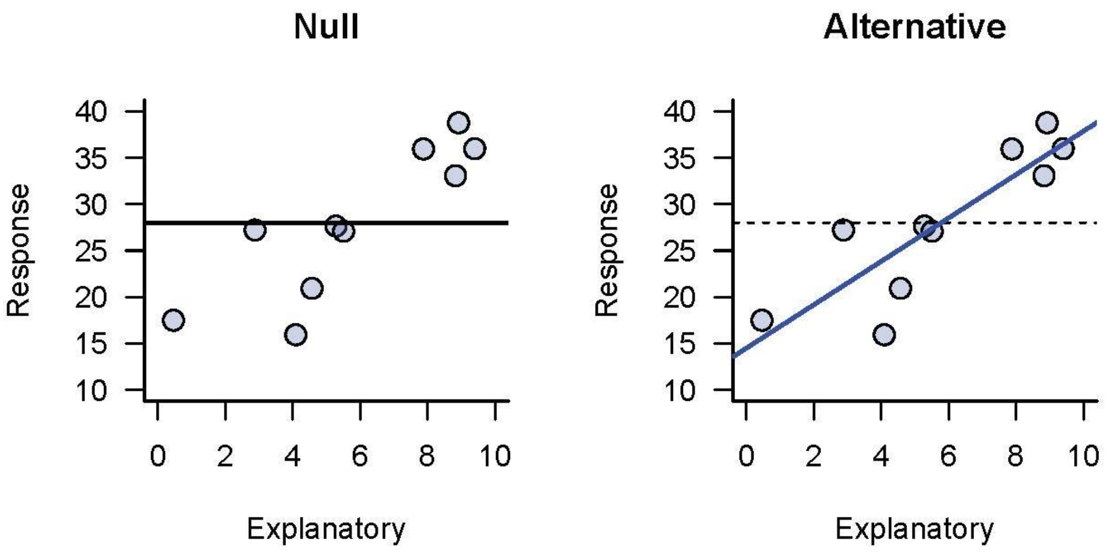
Regression vs. Correlation

Simple linear regression at its core is no different than a simple correlation!

$$y = mx + c$$

Except:

- H_0: slope is no different from 0
  - So, the p-value tells you something about the slope, rather than the strength of the correlation



Regression

You have already done a simple linear regression model in R!

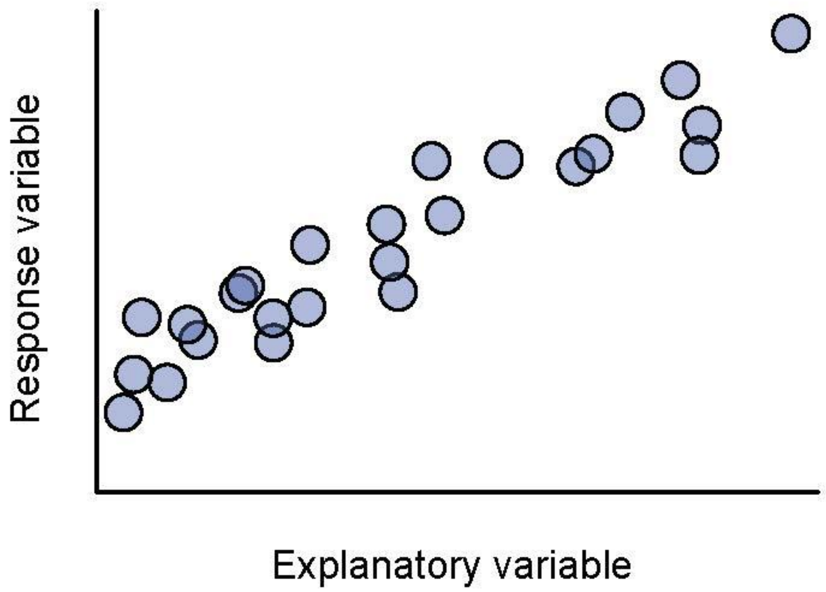
```
mod <- lm(Response ~ Explanatory, data = df)
```

lm(Response ~ Explanatory) uses the equation:

$$y = mx + c$$

And calculates the slope, intercept,

and if the slope is significantly different from 0



```r
mod <- lm(Response ~ Explanatory, data = df)
summary(mod)

Call:
lm(formula = Response ~ Explanatory, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-8.1126 -1.6674  0.2598  2.7585  5.9932

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  14.5011     3.1439   4.612  0.00173 **
Explanatory   2.3353     0.4896   4.770  0.00141 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.325 on 8 degrees of freedom
Multiple R-squared:  0.7399,    Adjusted R-squared:  0.7073
F-statistic: 22.75 on 1 and 8 DF,  p-value: 0.001409
```

Slope

Slope $\neq 0$

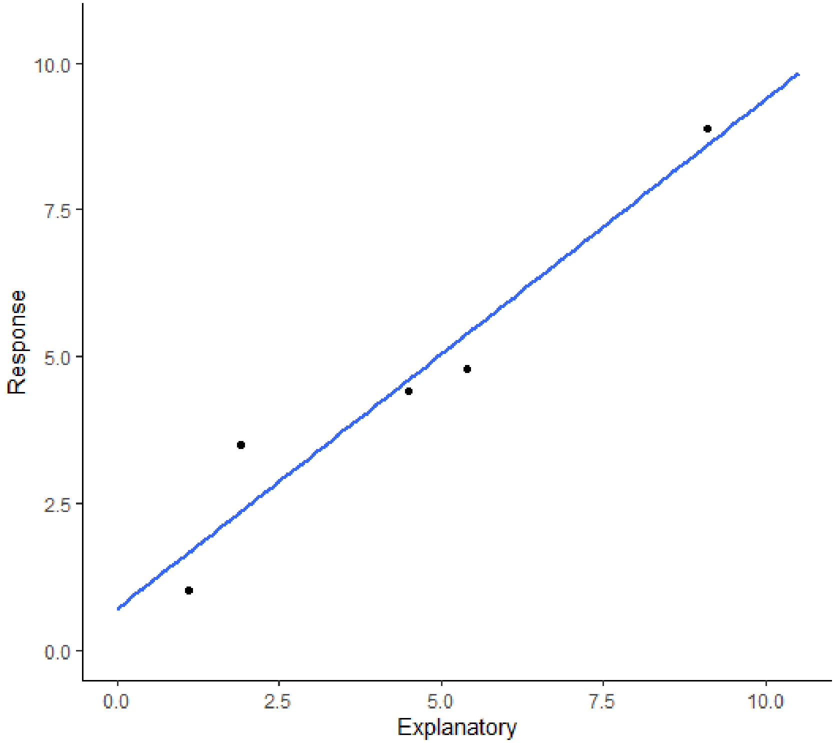
```r
mod <- lm(Response ~ Explanatory, data = df)
summary(mod)

Call:
lm(formula = Response ~ Explanatory, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-8.1126 -1.6674  0.2598  2.7585  5.9932

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  14.5011     3.1439   4.612  0.00173 **
Explanatory   2.3353     0.4896   4.770  0.00141 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.325 on 8 degrees of freedom
Multiple R-squared:  0.7399,    Adjusted R-squared:  0.7073
F-statistic: 22.75 on 1 and 8 DF,  p-value: 0.001409
```

?

Correlation = simple Regression

R-squared (R^2) = (Pearson's correlation coefficient)^2

- R^2 = how much of the variation in the response variable is explained by the explanatory variable

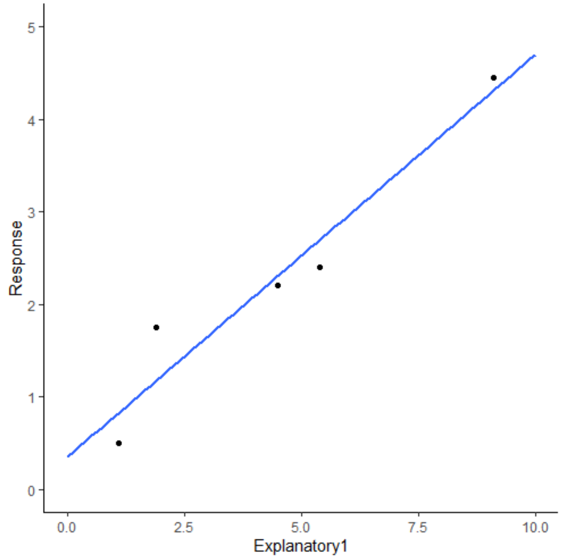


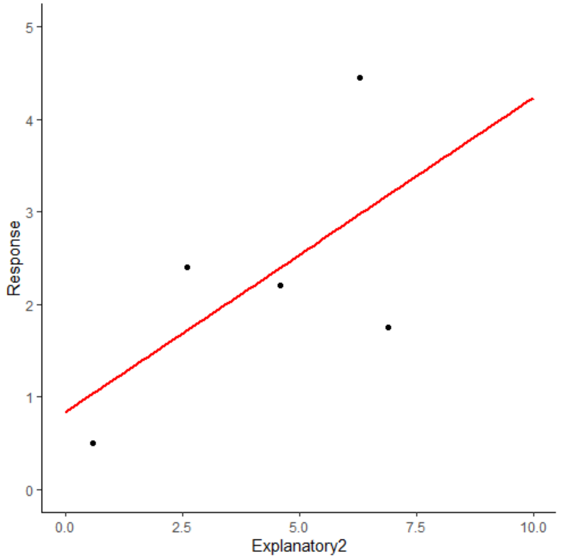
Multiple Regression

What happens when we have more than one explanatory variable?

$$y = m_1x_1 + m_2x_2 + \cdots + c$$

- H_0: there is *no significant* difference between the slopes and 0
- Assumptions:
  - Data are normally distributed
  - Relationship is linear between the explanatory and response factors (in this case)





$$b'_{y12} = \frac{(r_{1y} - r_{2y}r_{12})}{(1 - r_{12}^2)}$$

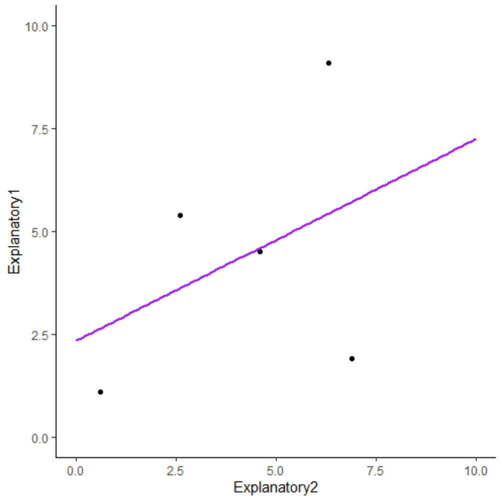
Beta coefficient is a standardized way of comparing the effect of each individual explanatory variable on the response variable

Multiple Regression

What happens when we have more than one explanatory variable?

$$y = m_1x_1 + m_2x_2 + \cdots + c$$

- H_0: there is *no significant* difference between the slopes and 0
- Assumptions:
  - Data are normally distributed
  - Relationship is linear between the explanatory and response factors (in this case)
  - Little to no multicollinearity

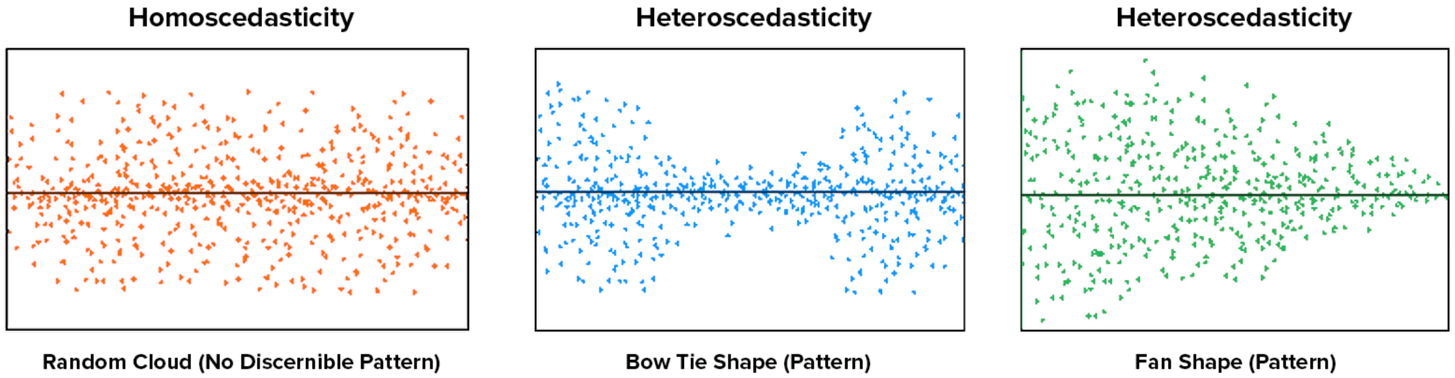


Multiple Regression

What happens when we have more than one explanatory variable?

$$y = m_1x_1 + m_2x_2 + \cdots + c$$

- H_0: there is *no significant* difference between the slopes and 0
- Assumptions:
  - Data are normally distributed
  - Relationship is linear between the explanatory and response factors (in this case)
  - Little to no multicollinearity
  - Low heteroscedasticity or "uneven error"



Multiple Regression

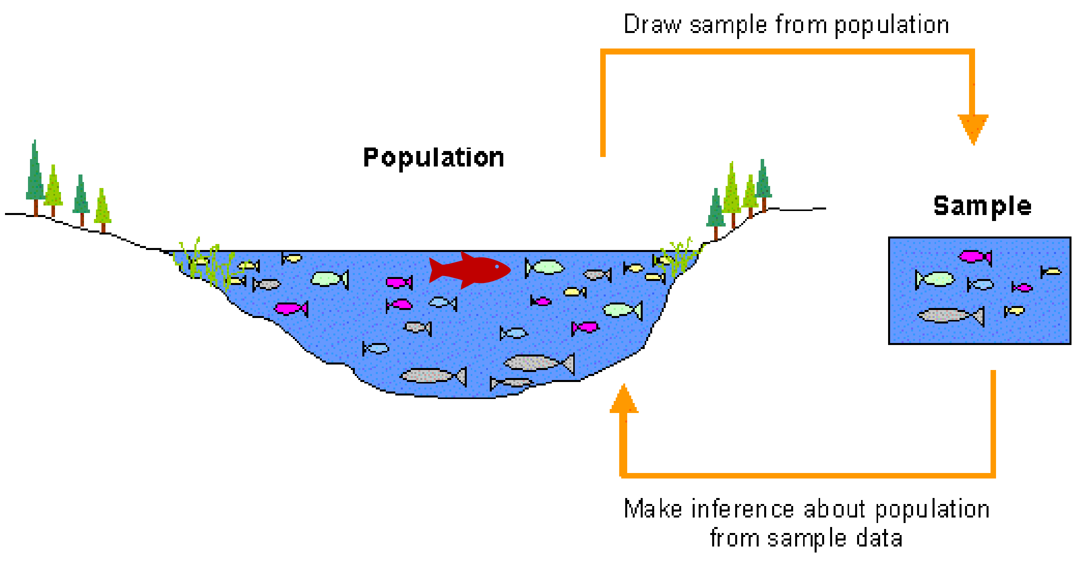
What happens when we have more than one explanatory variable?

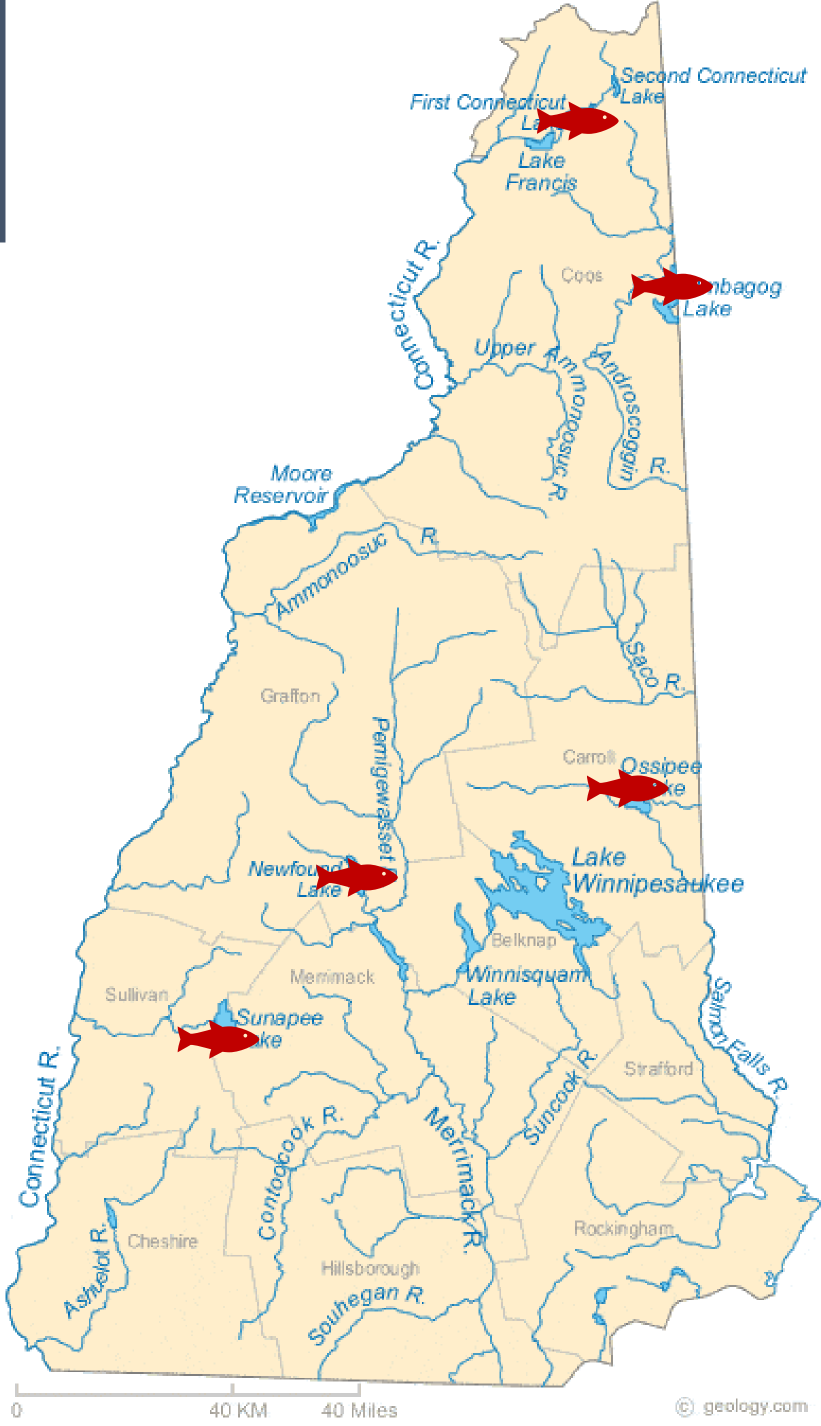
$$y = m_1 x_1 + m_2 x_2 + \cdots + c$$

- H_0: there is *no significant* difference between the slopes and 0
- Assumptions:
  - Data are normally distributed
  - Relationship is linear between the explanatory and response factors (in this case)
  - Little to no multicollinearity
  - Low heteroscedasticity or "uneven error"
  - You want the most "parsimonious" model (best fit and *simplest*!)
    - Adjusted R^2 and AIC penalize "fit" with each additional explanatory variable in the model

GraphSketch.com

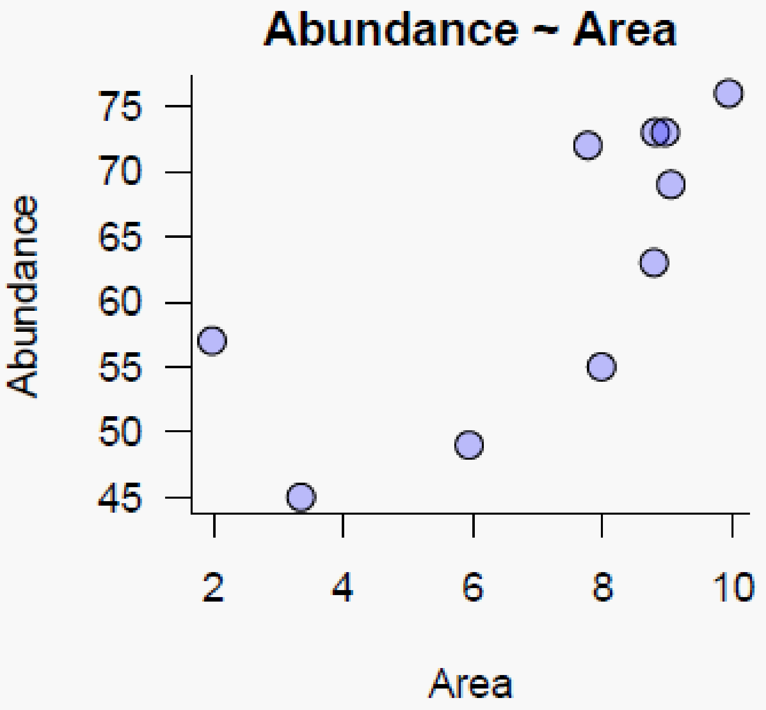
Multiple Regression - example

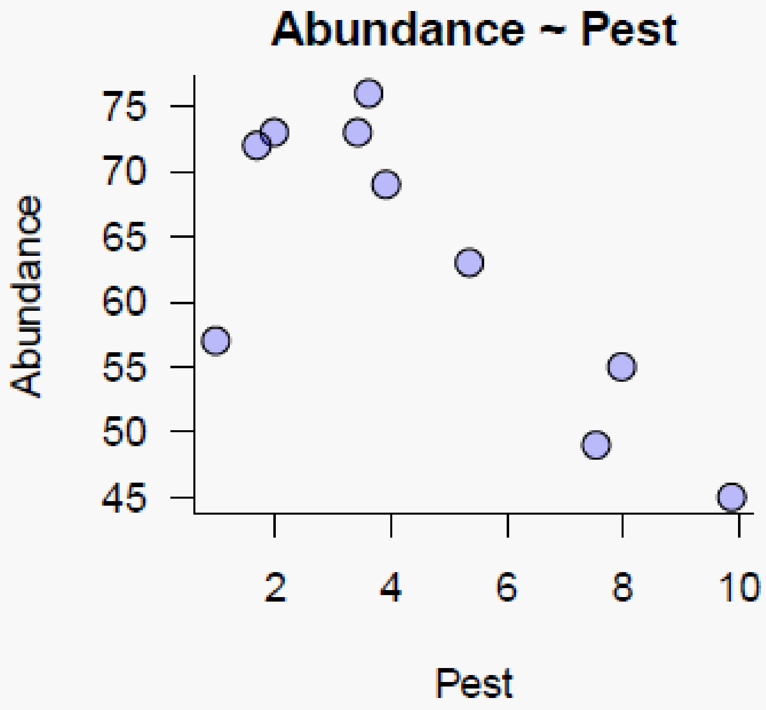




Multiple Regression - example

```
fish
   Abund Area Pest
1     55 7.98 7.98
2     49 5.94 7.54
3     69 9.05 3.91
4     73 8.81 3.42
5     76 9.94 3.61
6     73 8.96 1.98
7     63 8.79 5.35
8     57 1.98 0.97
9     45 3.35 9.88
10    72 7.77 1.68
```





```r
mod.joint <- lm(Abund ~ Area + Pest, data = fish)
summary(mod.joint)
```

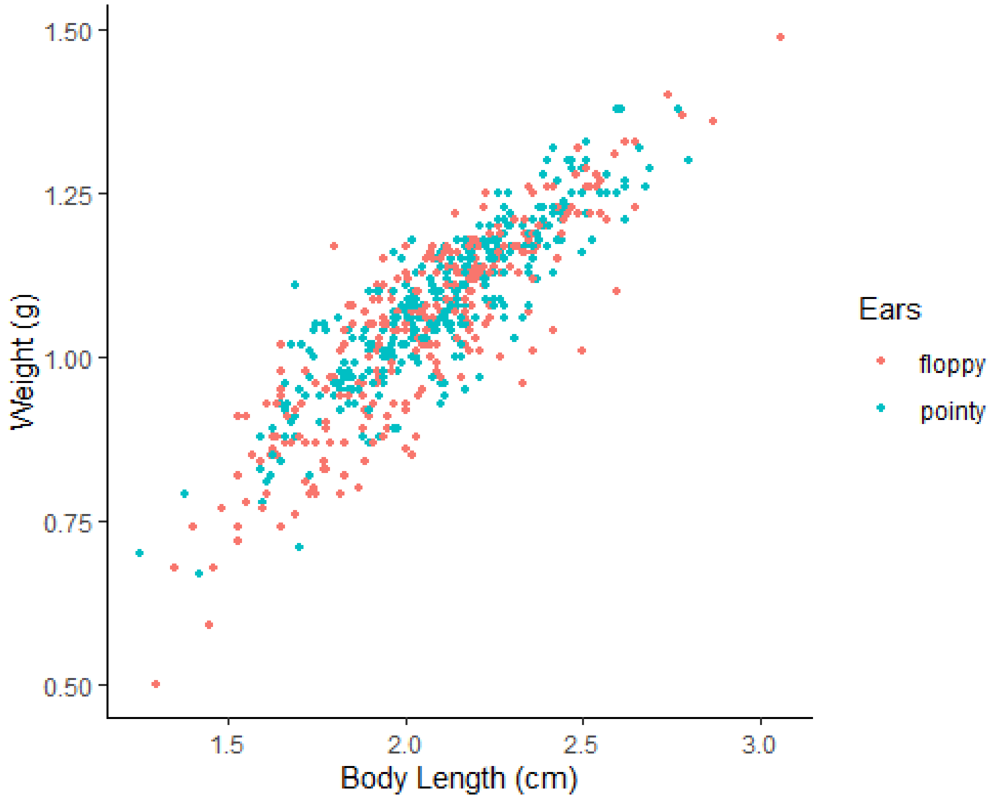
```
Call:
lm(formula = Abund ~ Area + Pest, data = fish)

Residuals:
    Min      1Q  Median      3Q     Max
-3.9362 -1.9270 -0.7162  2.3124  4.3071

Coefficients:
            Estimate Std. Error t value   Pr(>|t|)
(Intercept)  55.1918     3.7078  14.885 0.00000148 ***
Area          2.6037     0.3970   6.558   0.000316 ***
Pest         -2.3503     0.3545  -6.630   0.000296 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.112 on 7 degrees of freedom
Multiple R-squared:  0.9387,    Adjusted R-squared:  0.9212
F-statistic: 53.57 on 2 and 7 DF,  p-value: 0.00005711
```

Picking the best explanatory variables

Akaike Information Criterion (AIC)

- AIC penalizes when you add too many factors!
- Is a measure of how "bad" your model is
  - So the higher the AIC the worse the model

Let's try the bunny data!



Picking the best explanatory variables

```
bunnies.glm_weight

Call:
glm(formula = Floppy ~ Weight, family = "binomial", data = bunnies)

Coefficients:
(Intercept)       Weight
      1.953       -1.915

Degrees of Freedom: 613 Total (i.e. Null); 612 Residual
Null Deviance: 849.9
Residual Deviance: 840    AIC: 844
```

```
bunnies.glm_length

Call:
glm(formula = Floppy ~ BodyLength, family = "binomial", data = bunnies)

Coefficients:
(Intercept)       BodyLength
     1.1057          -0.5736

Degrees of Freedom: 613 Total (i.e. Null); 612 Residual
Null Deviance: 849.9
Residual Deviance: 846.3  AIC: 850.3
```

Picking the best explanatory variables

Weight only model: AIC = 844; Length only model: AIC = 850.3

```
summary(bunnies.glm)

Call:
glm(formula = Floppy ~ Weight + BodyLength, family = "binomial", data = bunnies)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5492  -1.1149  -0.9526   1.2166   1.5181

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.7304     0.6753   2.562  0.01039 *
Weight       -3.7348     1.2501  -2.988  0.00281 **
BodyLength    1.0370     0.6166   1.682  0.09262 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null Deviance: 849.91 on 613 degrees of freedom
Residual Deviance: 837.10 on 611 degrees of freedom
AIC: 843.1
```

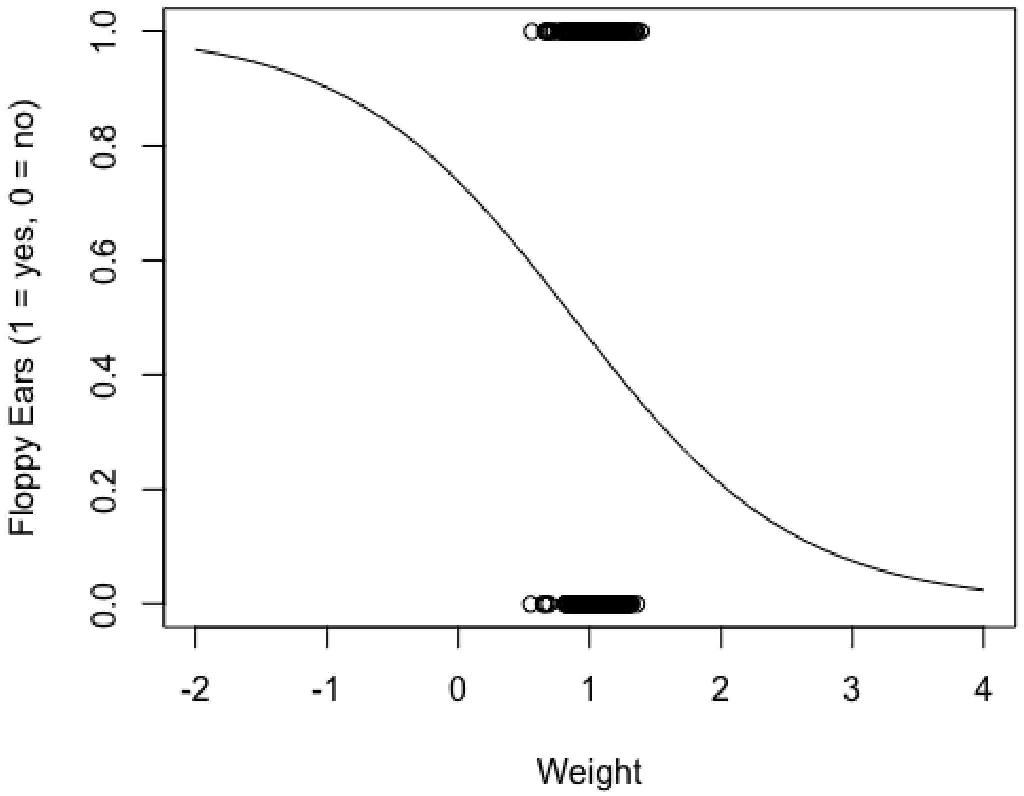
ΔAIC < 2
models aren't
different

How is it all regression?

Statistical test	Question	Regression version
T-test	Are these two samples different?	Response variable is categorical and explanatory variable is continuous (logistic regression)
ANOVA	Are these three (or more) samples different?	response variable is continuous and explanatory variables are categorical
Correlation (Pearson's)	Is there a link between these two continuous factors?	Simple 1 factor with only continuous variables
Association (Chi-square)	Is there a link between the categorical factors?	All categorical variables

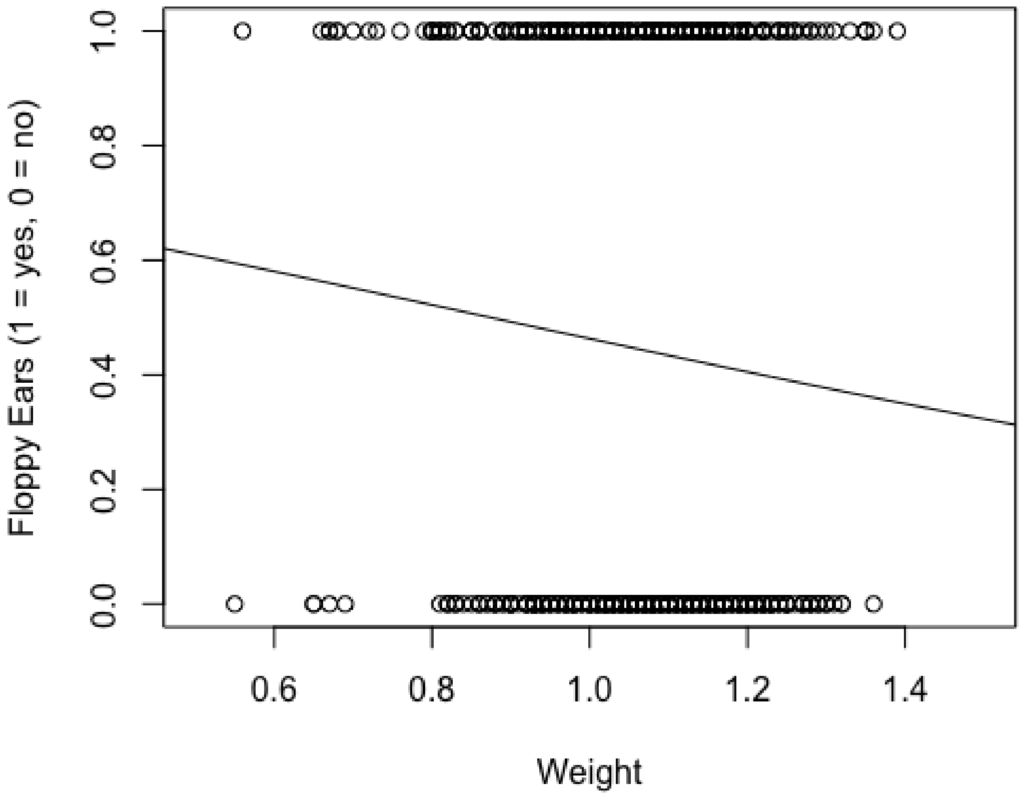
Logistic Regression

```
glm(formula = Floppy ~ Weight, family = "binomial", data =
bunnies)

xweight <- seq(-2, 4, 0.01)
yweight <- predict(bunnies.glm_weight, list(Weight =
xweight), type = "response")

plot(bunnies$Weight, bunnies$Floppy, xlab = "Weight", ylab
= "Floppy Ears (1 = yes, 0 = no)", xlim = c(-2,4))
lines(xweight, yweight)
```



Logistic Regression

```
glm(formula = Floppy ~ Weight, family = "binomial", data =
bunnies)

xweight <- seq(-2, 4, 0.01)
yweight <- predict(bunnies.glm_weight, list(Weight =
xweight), type = "response")

plot(bunnies$Weight, bunnies$Floppy, xlab = "Weight", ylab
= "Floppy Ears (1 = yes, 0 = no)", xlim = c(0.5,1.5))
lines(xweight, yweight)
```



Dangerous to extrapolate beyond sampled range!

Week 10 – Regression

Part II - Wednesday

Today's Exercise

New data! Vole population data where we have counts and some landscape information on where they were found. We're trying to answer the question:

1. Does percent vegetation (PercVeg) or distance to road (Dist2Road) influence vole population locations?

To do so – split yourselves into 3 teams and create **one .R script** with everything below to turn in *before the end of today*:

1. Data exploration team
   - How do each of the explanatory variables influence the response variable
     - Create scatter plots for each explanatory variable vs response variable and guess what you think the relationship is and make notes in your R comments
2. Multiple regression team
   - What hypotheses are you testing?
   - Create a model with both explanatory variables
     - Which variable(s) seems to explain some of the variation of the vole population?
3. Model choice team
   - Use the add1() function and AIC to pick the best model for explaining vole population
     - Which model is the best? Why?

For Monday:

1) Review all of your notes
2) Send me *at least* one question you would like me to try to review next week!

All before 11:55pm on Sunday