Advanced Statistics In R

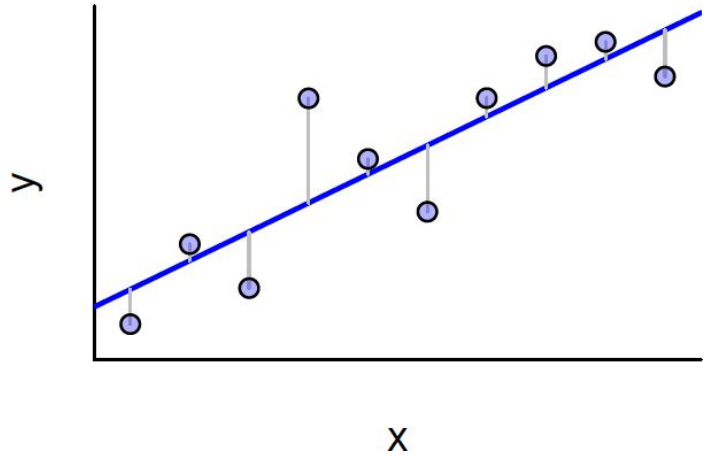
Announcements

- Class Survey!
- Envelope section of lab 6 is optional
- This is our second to last real class, with the rest of the semester being set aside for final projects

Shoutout to Meg Graham MacLean for the slides!

Linear Model Anatomy

- Linear models are the basis for many analytical methods
  - Two fundamental components:
    - Deterministic (signal) – the "expected" value of the response given X
    - Stochastic (noise) – the difference between the "observed" value of the response and the expected

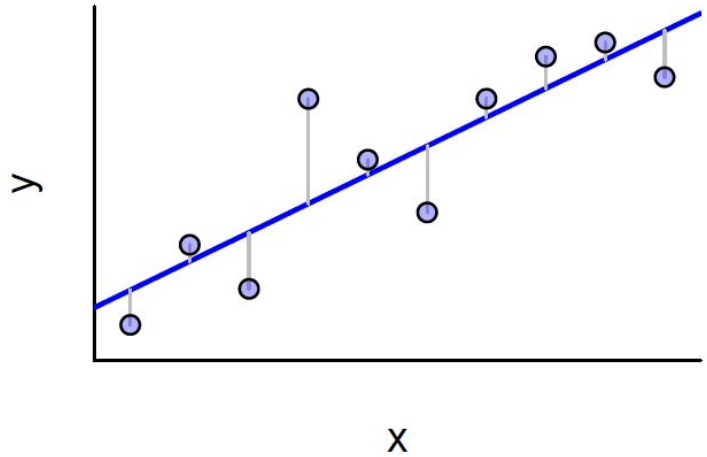


To find the best models we need statistics!

Linear Model Mathematically

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

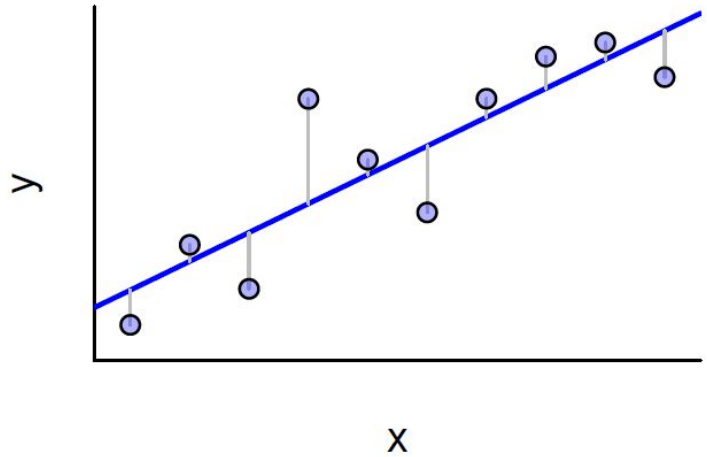
- Two fundamental components:
  - Deterministic (signal) – the "expected" value of the response given X
  - Stochastic (noise) – the difference between the "observed" value of the response and the expected



Simple linear model

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

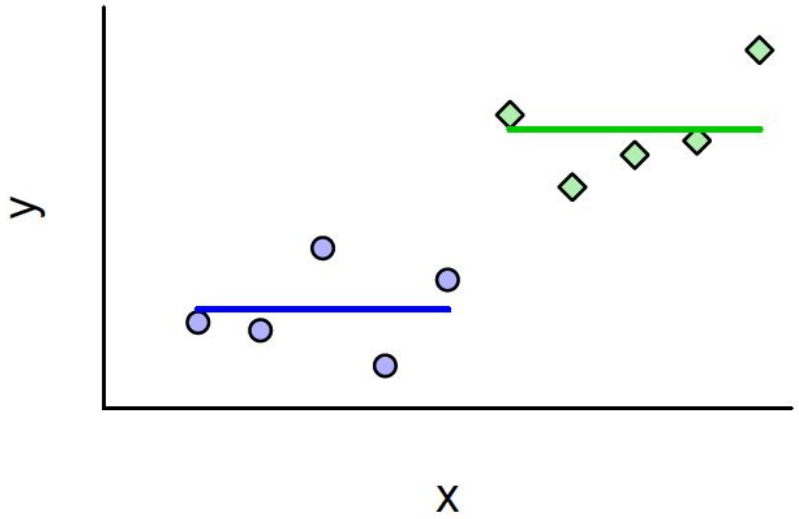
- Deterministic
  - β_0 and β_1 are parameters to be estimated
  - When X is *continuous* (mean = ———)



Simple linear model

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

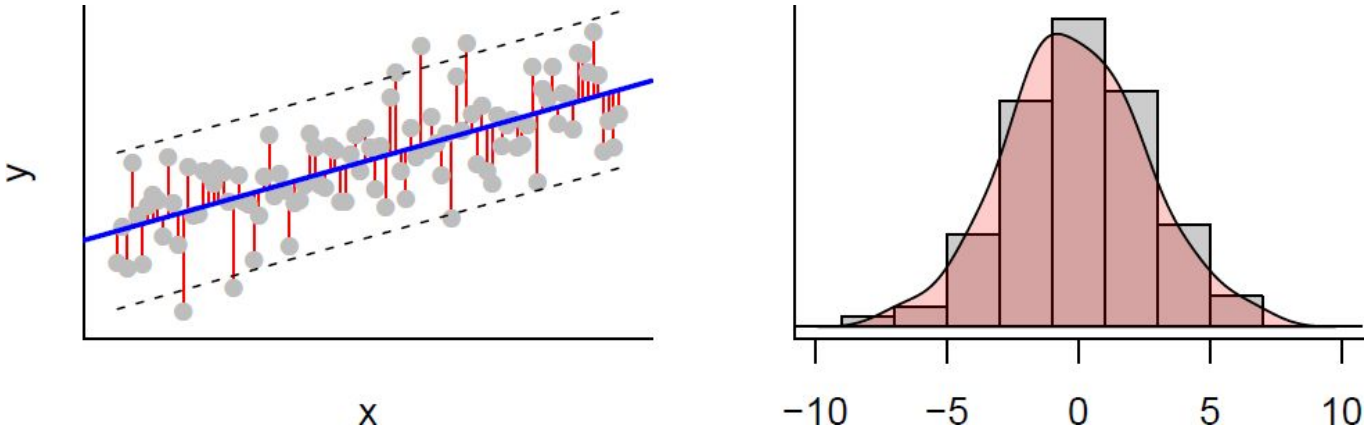
- Deterministic
  - β_0 and β_1 are parameters to be estimated
  - When X is *factor* (means = — —)



Simple linear model

$$y_i = \beta_0 + \beta_1 X_i + e_i$$
$$e_i = y_i - \hat{y}_i$$

- Stochastic – usually called the *residual*
  - Usually assume residuals are normally distributed (i.e., $N(0, \sigma^2)$)

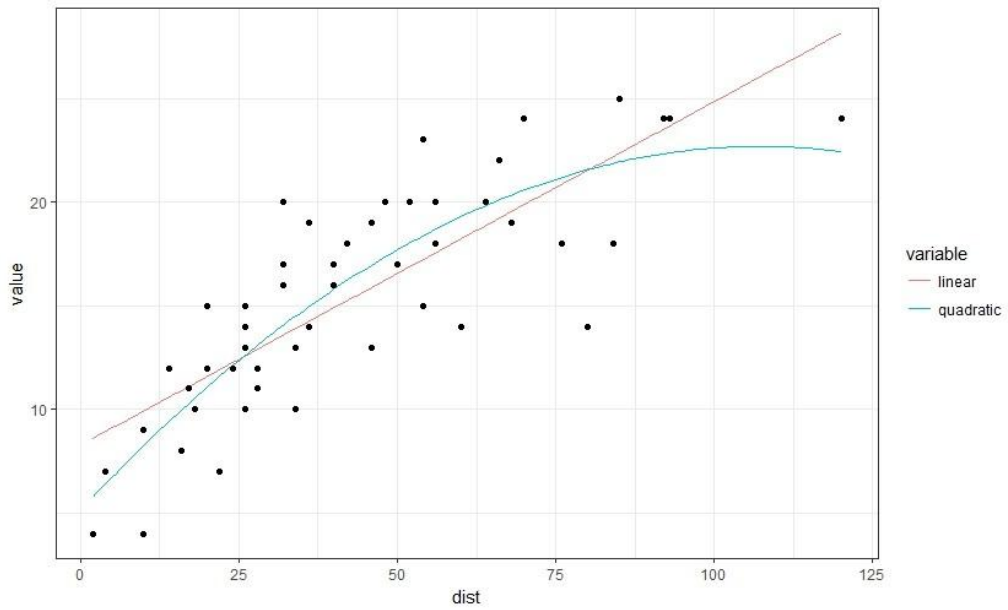


Simple linear model

Does a linear model have to be linear? No ☺

$$y_i = \beta_0 + \beta_1 X_i + \beta_2 X_{i1}^2 + e_i$$

- Y just needs to be expressed as a linear function of X, but that function can be curvy

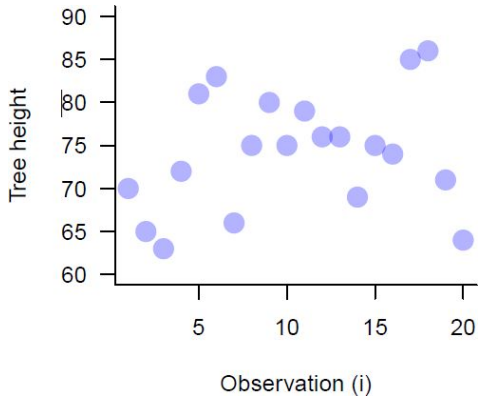


Let's try it – tree heights

Let's say we go into a forest stand that is of interest to us (perhaps we want to harvest some wood). We measure the height of 20 randomly selected trees.

1. State the question/hypothesis
   - What is the expected height of a tree in the stand?
   - Variable: tree height (response)

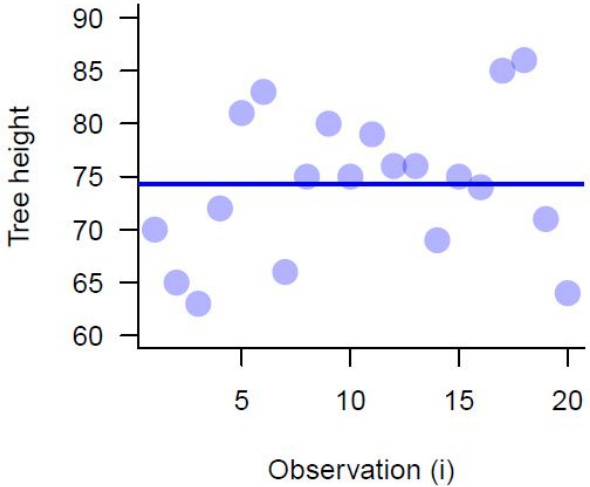
But let's use a linear model!



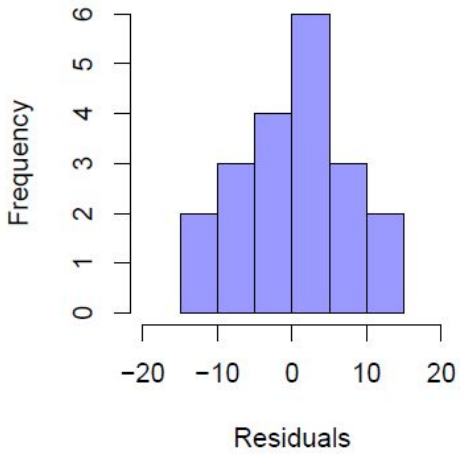
Describe the model

1. Describe the model – in word form:
   - What is the expected height of a randomly selected tree?
2. Describe the model – in mathematical form:

- y_i is height (response)
- β_0 is the intercept
- e_i is the residuals



1
1



Evaluating output and Interpreting results

```
> summary(lm(height~1,data = trees))

Call:
lm(formula = height ~ 1, data = trees)

Residuals:
   Min     1Q Median     3Q    Max
-11.25  -4.50   0.75   5.00  11.75

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   74.250      1.527   48.63   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.828 on 19 degrees of freedom
```

Parameter estimate

Standard errors of the estimate

t-statistic to test if coefficient is significantly different from 0

p-value, or the probability of getting *t* large (or larger) if null hypothesis ($\beta_0 = 0$) is true

1
2

We've done the null model… now what?

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

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

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 = null hypothesis that there is no difference between sample means
- H_1 = alternative hypothesis that the sample means differ

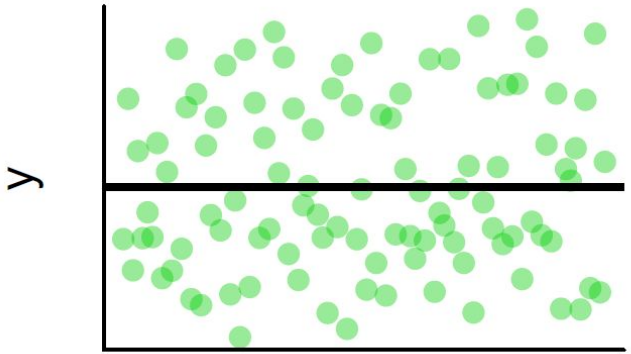
Review!

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?



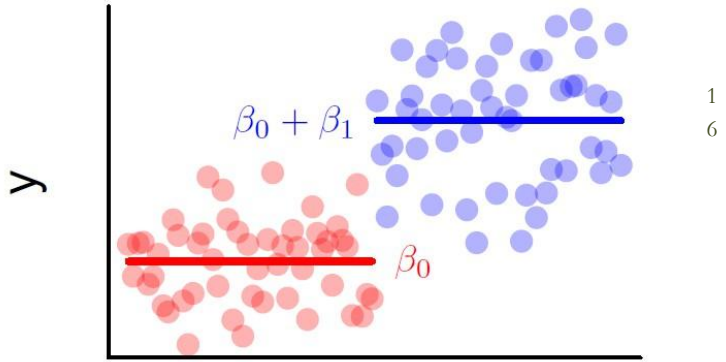
Continuous Two-level Factor Model

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?



Analysis of Variance (ANOVA)

- An **AN**alysis **O**f **VA**riation is used for examining the differences in the mean values of the dependent variable associated with the effect of three or more independent predictor variables.
- Used to determine if at least one predictor variable's mean is significantly different from the others.
- ANOVA Types:
  - **One-way ANOVA** tests differences across multiple groups based on a single factor.
  - **Two-way ANOVA** examines the effect of two different factors on a dependent variable, including interaction effects between factors

Continuous Multi-level Factor Model

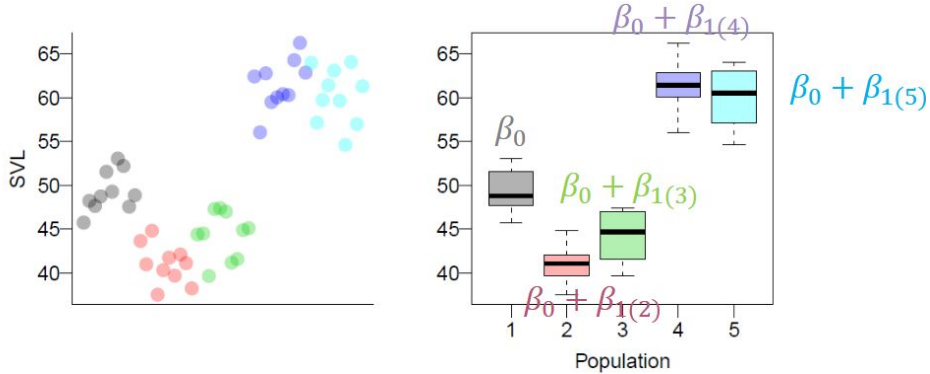
Response (Y)	Explanatory (X)	Model	In R
Continuous	None	Intercept-only/null	`lm(y~1)`
Continuous	Two-level factor	*t*-test	`lm(y~x)`
Continuous	Multi-level factor	ANOVA	`lm(y~x)`

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?

i = x values
g = each factor



Example Results

Interpret results

DBH = diameter at breast height
- Continuous

Stand (study area)
- Categorical

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

```
> summary(lm(DBH~Stand,data = tree))

Call:
lm(formula = DBH ~ Stand, data = tree)

Residuals:
    Min      1Q  Median      3Q     Max
-38.666 -11.168  -3.487  13.100  41.336

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   69.333      3.732  18.580  < 2e-16 ***
StandB        32.709      5.277   6.198 1.25e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 18.66 on 48 degrees of freedom
Multiple R-squared:  0.4446,  Adjusted R-squared:  0.433
F-statistic: 38.42 on 1 and 48 DF,  p-value: 1.248e-07
```

Example

Interpret results

```
> summary(lm(DBH ~ Stand, data=tree))$coefficients
             Estimate Std. Error   t value     Pr(>|t|)
(Intercept) 69.33339   3.731628 18.579934 1.461584e-23
StandB      32.70935   5.277318  6.198101 1.248251e-07
```

Intercept/ mean of Stand A is sig. diff. from 0

Difference between Stand A and Stand B is sig. diff. from 0 *and* Stand B is sig. larger than Stand A!

- Intercept is…
  - The mean of Stand A, or β_0
- StandB is…
  - The difference/contrast between Stand A and Stand B, or β_1

>1 explanatory variables, multiple samples!

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

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)`

Null Model: $y_i = \beta_0 + e_i$

t-Test: $y_i = \beta_0 + \beta_1 X_i + e_i$

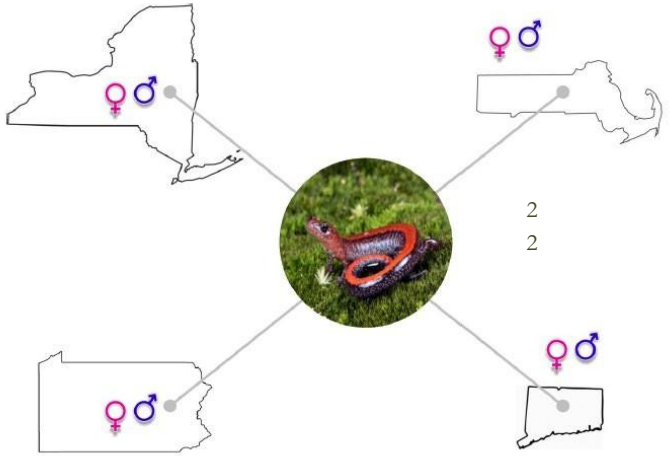
One-way ANOVA: $y_i = \beta_0 + \beta_{1(g)} X_{i(g)} + e_i$

Two-way ANOVA: $y_i = \beta_0 + \beta_{1(g)} X_{1i(g)} + \beta_{2(g)} X_{2i(g)} + e_i$ (additive model)

Two-way ANOVA as a linear model

Example for salamander lengths!

- 4 salamander populations of interest
- 2 sexes of interest
- Question: Does snout-vent-length (SVL) differ among salamander populations and sexes?



Two-way ANOVA as a linear model

Example for salamander lengths!

- 4 salamander populations of interest
- 2 sexes of interest
- Question: Does SVL differ among salamander populations and sexes?

Features of a two-way ANOVA

- Tests for differences between means
  - Means of groups-within-groups
- Tests for differences between factor combinations!

Two-way ANOVA Variables

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

- 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

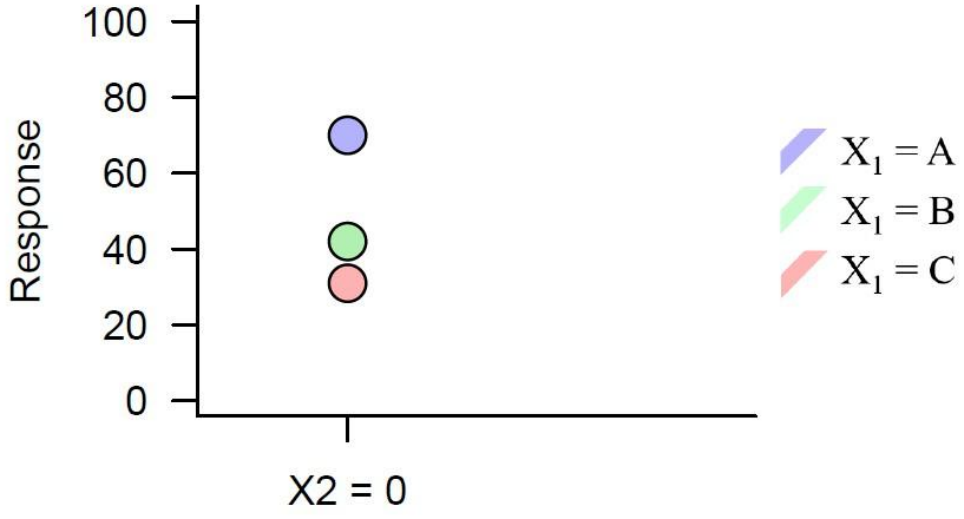
Two-way ANOVA Graphically

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

What if we explore this graphically…

$$X_{2i(g)} = 0$$

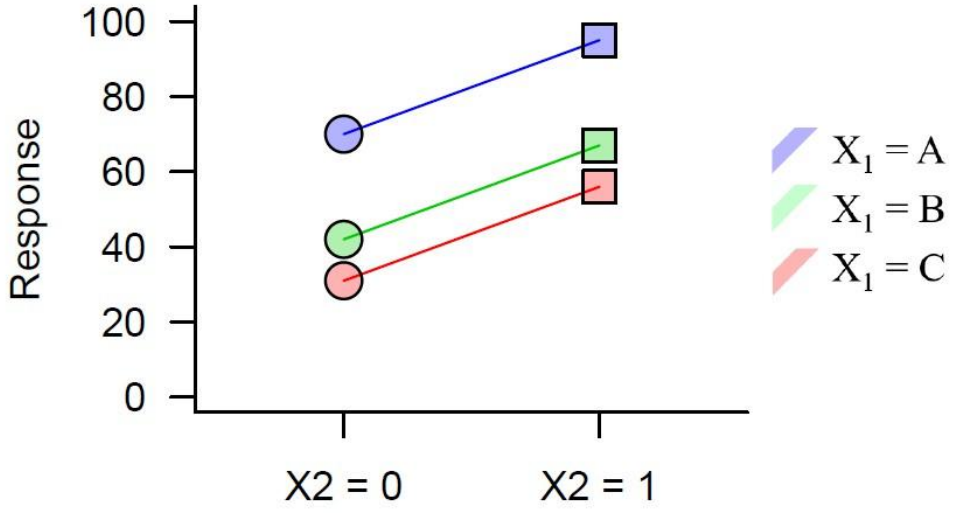
$$y_i = \beta_0 + \beta_{1(g)}X_{1i(g)} + \beta_{2(g)} * 0 + e_i$$



Two-way ANOVA Graphically

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

What if we explore this graphically…



Salamander T-Test

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

Made up data of sexes and populations of salamanders:

```
> sex <- c("F","F","F","F",  "M","M","M","M")
> pop <- c("A","B","C","D",  "A","B","C","D")
```

What if we just had sexes (no population variable)?

- T-test!

```
  (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
```

Salamander One-way ANOVA

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

Made up data of sexes and populations of salamanders:

```
> sex <- c("F","F","F","F",  "M","M","M","M")
> pop <- c("A","B","C","D",  "A","B","C","D")
```

What if we just had population (no sex variable)?

- ANOVA!

```
  (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
```

Salamander Two-way ANOVA

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

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

```
> sex <- c("F","F","F","F",  "M","M","M","M")
> pop <- c("A","B","C","D",  "A","B","C","D")
```

	(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

What is β_0?

Two-way ANOVA as a linear model

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

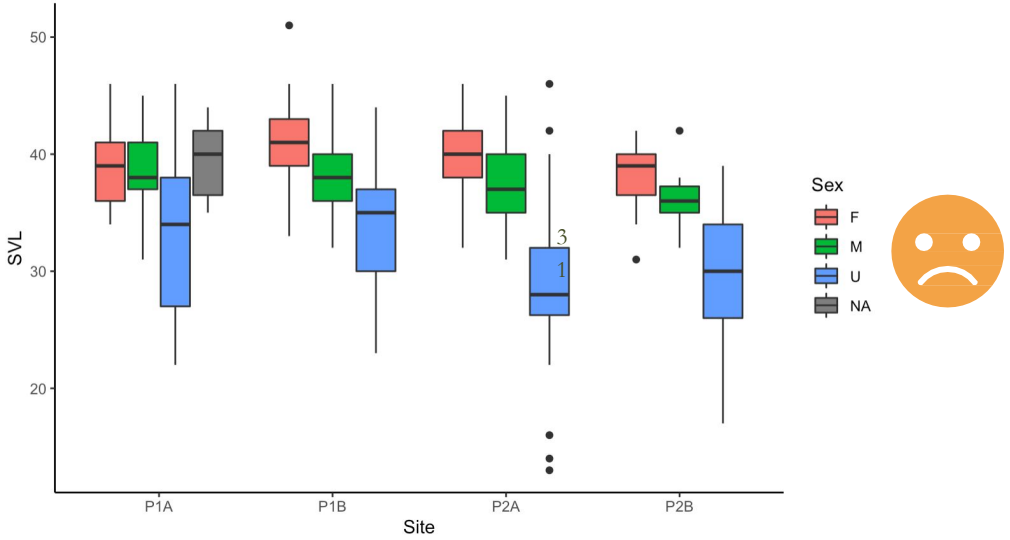
- β_0 is the mean of the first combination of factors – the reference level
  - **Females in Population A** <- super important to know your reference level
- What are the slopes ($\beta_{1(g)}$ and $\beta_{2(g)}$)?
- $\beta_{1(g)}$ is the group 1 contrasts – relate to the sex effect
  - $\beta_{1(g=male)}$ - the difference between males and females *in all populations*
- $\beta_{2(g)}$ is the group 2 contrasts – relate to the population effect
  - $\beta_{2(popB)}$ - the difference between *both sexes* in pop B and pop A
  - $\beta_{2(popC)}$ - the difference between *both sexes* in pop C and pop A
  - $\beta_{2(popD)}$ - the difference between *both sexes* in pop D and pop A

Two-way ANOVA as a linear model

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

Let's go back to our earlier question and modify it a little…

- Is there a significant difference in SVL among salamander populations *OR* sexes?
  - Response:
    - SVL
  - Explanatory:
    - Site (factor)
    - Sex (factor)

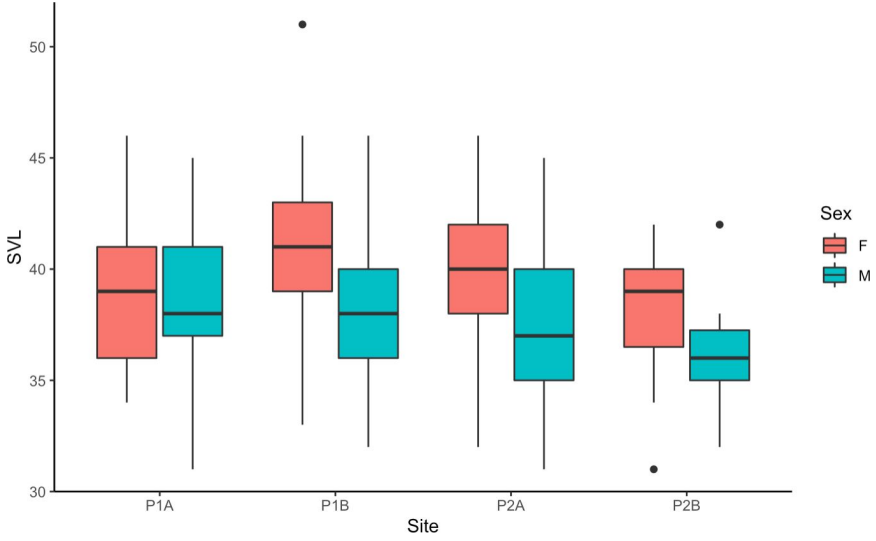


Two-way ANOVA as a linear model

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

Let's go back to our earlier question and modify it a little…

- Is there a significant difference in SVL among salamander populations *OR* sexes?
  - Response:
    - SVL
  - Explanatory:
    - Site (factor)
    - Sex (factor)



Simple Linear Regression

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	$\texttt{lm(y~x}_1\texttt{*x}_2\texttt{)}$
Continuous	Single continuous	Simple linear regression	`lm(y~x)`

Estimating the relationship between variables!

$$y_i = \beta_0 + \beta_1 X_{1i} + e_i$$

Simple linear regression example

- 331 salamanders
- Measured total length (TL) and snout-to-vent length (SVL)
- Tail length (Tail) = TL - SVL

Example

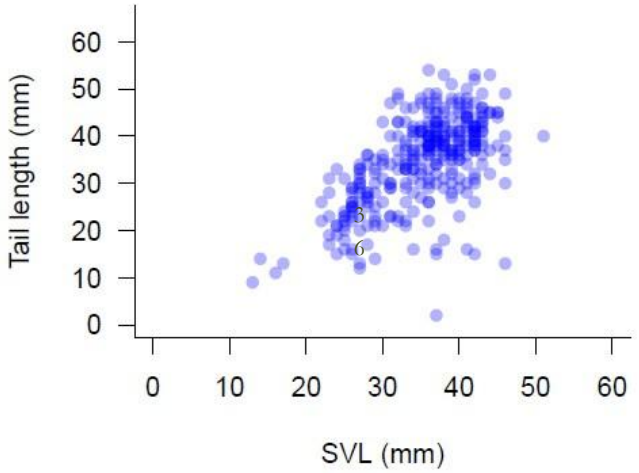
- 331 salamanders
- Measured total length (TL) and snout-to-vent length (SVL)
- Tail length (Tail) = TL - SVL

```
> head(mander)
  Season Site SVL TL Sex Cap     Ind Tail
1 Spring  P1A  43 86   U   N xxBBP1A   43
2 Spring  P1A  33 66   U   N xYxBP1A   33
3 Spring  P1A  42 84   M   N xYBxP1A   42
4 Spring  P1A  36 76   U   N xYYxP1A   40
5 Spring  P1A  44 76   M   N xxBYP1A   32
6 Spring  P1A  42 74   U   N xBxYP1A   32
```

Example

1. State the question:

- Does tail length scale predictably with SVL?
- H_0: there is no relationship between tail length and SVL
  - Response:
    - Tail length
  - Explanatory:
    - SVL



Example

2. Describe the model:

- In words:
  - Is there a significant relationship between tail length and SVL?
- In mathematical form:
  - $y_i = \beta_0 + \beta_1 X_{1i} + e_i$
  - y_i is tail length, X_{1i} is SVL
  - H_0 is $\beta_1 = 0$.
- What are the model assumptions?
  - Residuals are normally distributed
  - Constant variance (homogeneity)
  - Observations are independent
  - Predictors measured without error (fixed X)

Example

3. Fit the model

- Algebraically
  - $y_i = \beta_0 + \beta_1 X_{1i} + e_i$
- In R:

```
> mAllo <- lm(Tail ~ SVL, data = mander)
> coef(mAllo)
  (Intercept)          SVL
    3.1490945    0.8942684
```

$$y = 3.15 + 0.89x$$

Model Choice with Akaike Information Criterion

- AIC is a method used in statistics to measure how well a statistical model fits the data.
- Penalizes models for the number of parameters, helping to avoid overfitting.
- AIC helps in model selection by comparing different models
- The model with the lower AIC value is generally preferred as it indicates a better balance between goodness of fit and complexity.
- AICcmodavg Package in R

Example

4. Evaluate the output

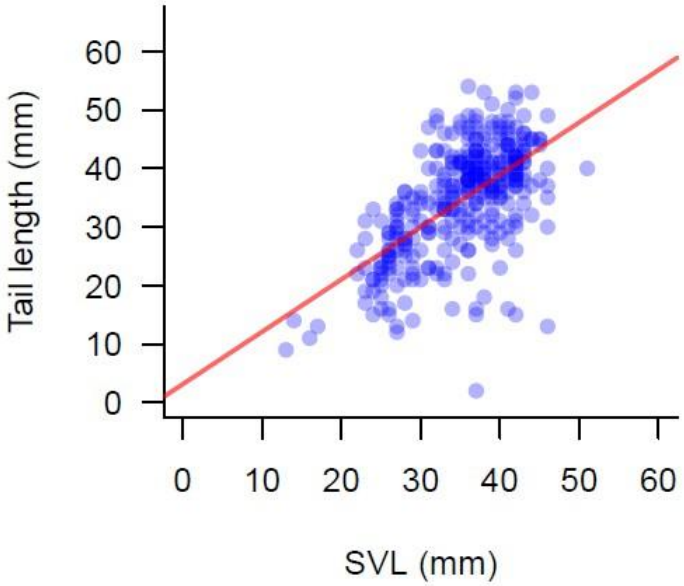
- Model selection
  - Two models: null and linear

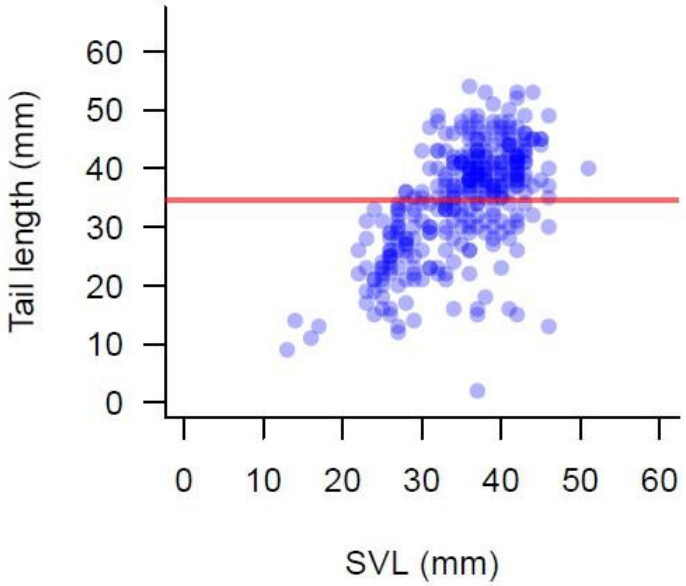
```
> m0    <- lm(Tail ~ 1,   data = mander) # null
> mAllo <- lm(Tail ~ SVL, data = mander) # lin reg
```

```
> aictab(list(m0,mAllo),c("m0","mAllo"))

  Model selection based on AICc:

       K    AICc Delta_AICc AICcWt Cum.Wt       LL
  mAllo 3 2275.21       0.00      1      1 -1134.57
  m0    2 2423.53     148.32      0      1 -1209.75
```

Example

5. Evaluate the output
   - Model selection
     - Two models: null and linear





What about >1 continuous 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)`
Continuous	Single continuous	Simple linear regression	`lm(y~x)`
Continuous	Multiple continuous	Multiple linear regression	`lm(y~x1*x2)`

Estimating the relationship with multiple explanatory variables!

$$y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + e_i$$

Multiple Linear Regression

$$y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + e_i$$

- β_0 is the intercept
- β_1 is the slope of the X_1 relationship
  - The change in $\hat{y}_i$ with one unit change in X_1 at *any* value of X_2 (*additive model!)*
- β_2 is the slope of the X_2 relationship
  - The change in $\hat{y}_i$ with one unit change in X_2 at *any* value of X_1 (*additive model!)*

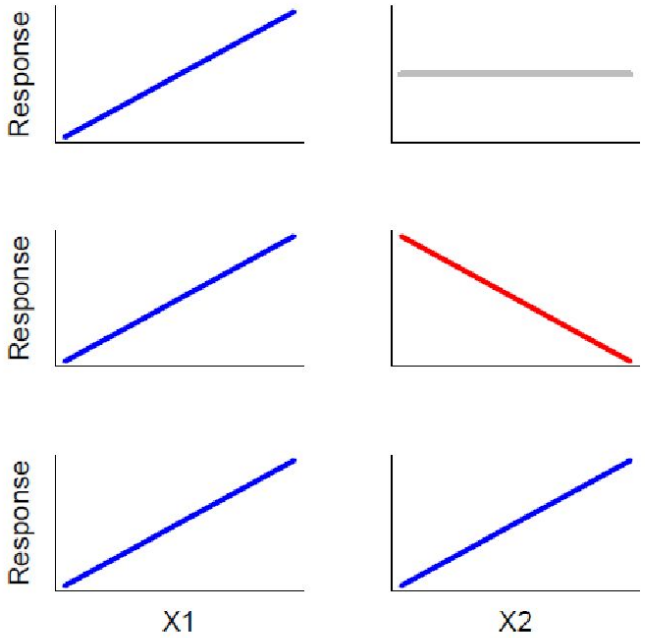
The changes in $\hat{y}_i$ with the change in one explanatory variable while the other(s) held constant are often called: marginal effects

Multiple linear regression

$$y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + e_i$$

What are the possible general outcomes for our parameters?

$\beta_1 > 0$



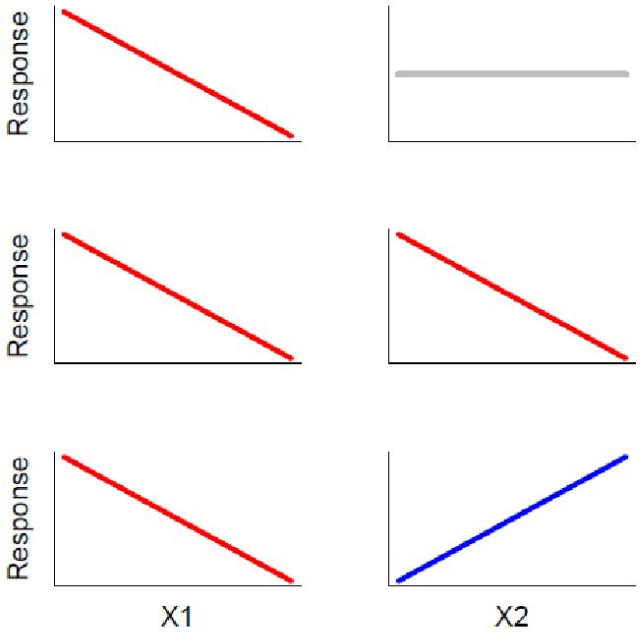
$\beta_2 = 0$
$\beta_2 < 0$
$\beta_2 > 0$

Multiple linear regression

$$y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + e_i$$

What are the possible general outcomes for our parameters?

$\beta_1 < 0$



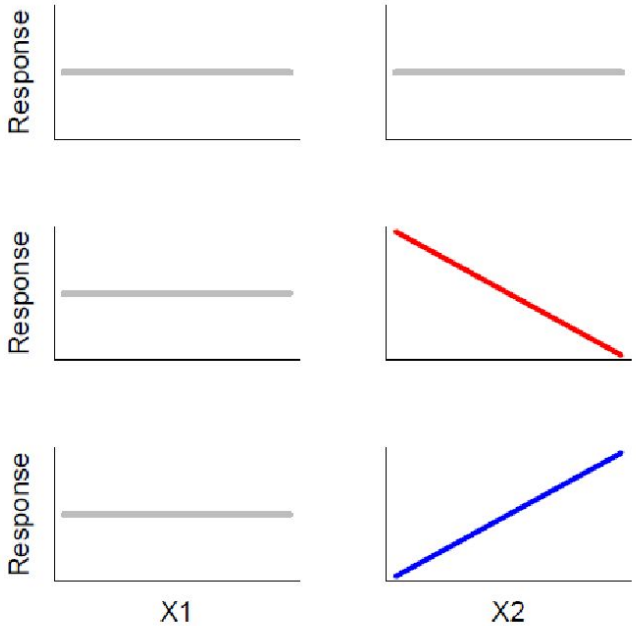
$\beta_2 = 0$
$\beta_2 < 0$
$\beta_2 > 0$

Multiple linear regression

$$y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + e_i$$

What are the possible general outcomes for our parameters?

$\beta_1 = 0$



$\beta_2 = 0$
$\beta_2 < 0$
$\beta_2 > 0$

Example

Indigo snakes

- 92 indigo snakes
- We are interested in the variation in home range sizes of the snakes (hr.size)
  - We log transformed the home range data (log.HR)
- Our proposed covariates are surrounding habitat structure (proportion)
  - Urban1.50
  - Upland1.50
  - Wetland1.50

Example

1. State the question/hypothesis

Question:

Does habitat composition at the home range scale explain variation in home range size?

H_0: There is no relationship between home range size and habitat composition.

Variables:

- Response:
  - Home range
- Explanatory
  - Urban percent
  - Upland percent
  - Wetland percent

Example

3. Describe the model

- In words:
  - Does habitat composition at the home range scale explain variation in home range size?
- As a mathematical model:
  - $y_i = \beta_0 + \beta_1 X_{urbi} + \beta_2 X_{upi} + \beta_3 X_{weti} + e_i$
  - H_0: $\beta_1 = 0, \beta_2 = 0, \beta_3 = 0$
- Assumptions?
  - Residuals are normally distributed
  - Constant variance (homogeneity)
  - Observations are independent
  - Predictors measured without error (fixed X)

Example

4. Fit the model

- Algebra: $y_i = \beta_0 + \beta_1 X_{urbi} + \beta_2 X_{upi} + \beta_3 X_{weti} + e_i$
- R:

```
> mG <- lm(log.HR ~ urban1.50 + upland1.50 + wetland1.50, data = indigos)
> model.matrix(~ urban1.50 + upland1.50 + wetland1.50, data = indigos)
```

	(Intercept)	urban1.50	upland1.50	wetland1.50
1	1	0.36	0.17	0.45
2	1	0.00	0.67	0.35
3	1	0.01	0.67	0.24
4	1	0.06	0.60	0.35
5	1	0.00	0.55	0.46
6	1	0.30	0.23	0.40

Example

4. Fit the model

$$y_i = \beta_0 + \beta_1 X_{urbi} + \beta_2 X_{upi} + \beta_3 X_{weti} + e_i$$

```
> summary(mG)

Call:
lm(formula = log.HR ~ urban1.50 + upland1.50 + wetland1.50, data = indigos)

Residuals:
     Min       1Q   Median       3Q      Max
-2.63249 -0.52193  0.06665  0.59191  1.69868

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   4.6091     0.3502  13.160  < 2e-16 ***
urban1.50    -1.8465     0.5392  -3.424 0.000938 ***
upland1.50    0.9171     0.4999   1.835 0.069945 .
wetland1.50   0.6711     0.6086   1.103 0.273153
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8476 on 88 degrees of freedom
Multiple R-squared:  0.3091,  Adjusted R-squared:  0.2855
F-statistic: 13.12 on 3 and 88 DF,  p-value: 3.694e-07
```

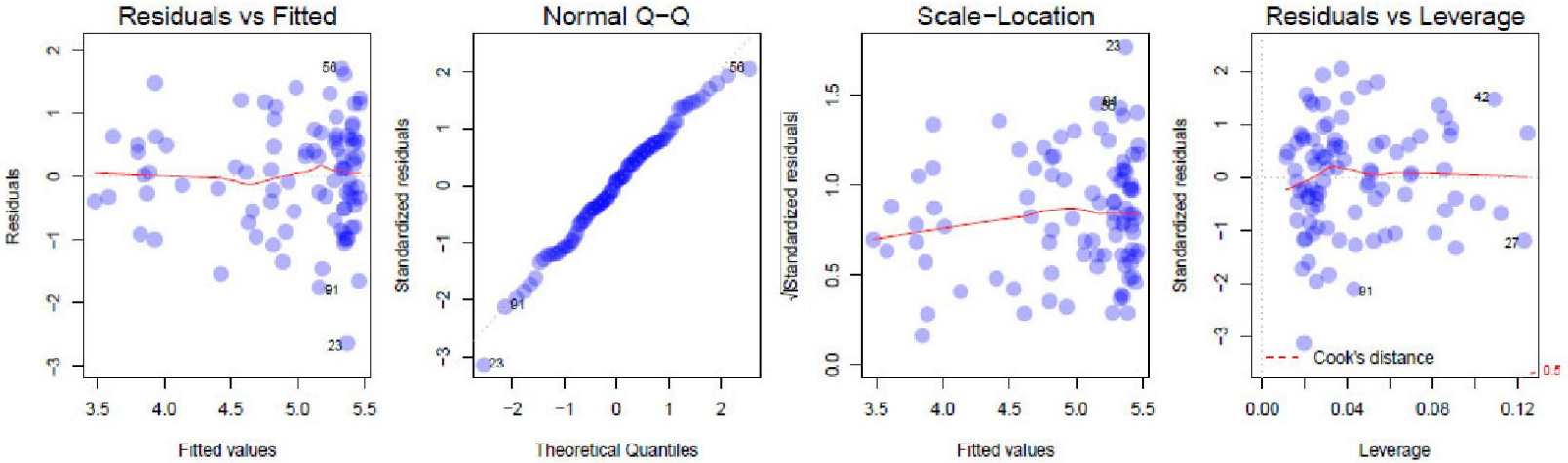
Example

5. Evaluate the output

- Model validation - check assumptions!

```
> mG <- lm(log.HR ~ urban1.50 + upland1.50 + wetland1.50, data = indigos)
> summary(mG)
> plot(mG)
```

Do we meet our assumptions?



Example

5. Evaluate the output

- Model selection
  - What is the best model? What other candidate models are there?
    - Full (global) model: $y_i = \beta_0 + \beta_1 X_{urbi} + \beta_2 X_{upi} + \beta_3 X_{weti} + e_i$
    - Single predictor models:
      - $y_i = \beta_0 + \beta_1 X_{urbi} + e_i$
      - $y_i = \beta_0 + \beta_2 X_{upi} + e_i$
      - $y_i = \beta_0 + \beta_3 X_{weti} + e_i$
    - Null model: $y_i = \beta_0 + e_i$
    - Combination models!
      - $y_i = \beta_0 + \beta_1 X_{urbi} + \beta_2 X_{upi} + e_i$
      - $y_i = \beta_0 + \beta_1 X_{urbi} + \beta_3 X_{weti} + e_i$
      - $y_i = \beta_0 + \beta_2 X_{upi} + \beta_3 X_{weti} + e_i$

Example

5. Evaluate the output

- Model selection
  - What is the best model? What other candidate models are there?

```
> UrUpWe<- lm(log.HR ~ urban1.50  + upland1.50 + wetland1.50, data = indigos)
> UrUp   <- lm(log.HR ~ urban1.50  + upland1.50,  data = indigos)
> UrWe   <- lm(log.HR ~ urban1.50  + wetland1.50, data = indigos)
> UpWe   <- lm(log.HR ~ upland1.50 + wetland1.50, data = indigos)
> Ur     <- lm(log.HR ~ urban1.50,   data=indigos)
> Up     <- lm(log.HR ~ upland1.50,  data=indigos)
> We     <- lm(log.HR ~ wetland1.50, data=indigos)
> m0     <- lm(log.HR ~ 1, data=indigos)
```

How do we pick?

Example

5. Evaluate the output

- Model selection
  - What is the best model?
    - AIC

```
> fitList <- list(
+  UrUpWe = lm(log.HR ~ urban1.50  + upland1.50 + wetland1.50, data = indigos),
+  UrUp   = lm(log.HR ~ urban1.50  + upland1.50,  data = indigos),
+  UrWe   = lm(log.HR ~ urban1.50  + wetland1.50, data = indigos),
+  UpWe   = lm(log.HR ~ upland1.50 + wetland1.50, data = indigos),
+  Ur     = lm(log.HR ~ urban1.50,   data=indigos),
+  Up     = lm(log.HR ~ upland1.50,  data=indigos),
+  We     = lm(log.HR ~ wetland1.50, data=indigos),
+  m0     = lm(log.HR ~ 1, data=indigos)
+ )
```

Example

5. Evaluate the output

- Model selection
  - What is the best model? What other candidate models are there?

```
> (modtab <- aictab(fitList))

  Model selection based on AICc:

        K   AICc Delta_AICc AICcWt Cum.Wt      LL
UrUp    4 236.29       0.00   0.38   0.38 -113.91
Ur      3 237.15       0.87   0.25   0.63 -115.44
UrUpWe  5 237.26       0.98   0.24   0.87 -113.28
UrWe    4 238.48       2.19   0.13   1.00 -115.01
UpWe    4 246.53      10.25   0.00   1.00 -119.04
Up      3 250.30      14.01   0.00   1.00 -122.01
We      3 259.50      23.21   0.00   1.00 -126.61
m0      2 264.72      28.43   0.00   1.00 -130.29
```

Huh, looks like maybe we should try the Urban and Upland model

Example

6. Interpret the results

- How do we visualize our model?

$$y_i = \beta_0 + \beta_1 X_{urbi} + \beta_2 X_{upi} + e_i$$

```
> summary(mTop)$coefficients
              Estimate Std. Error   t value     Pr(>|t|)
(Intercept)  4.854762  0.2705938 17.941140 2.600465e-31
urban1.50   -2.070404  0.5001409 -4.139642 7.885350e-05
upland1.50   0.863374  0.4981017  1.733329 8.650048e-02
```

$$y_i = 4.85 - 2.07 X_{urbi} + 0.86 X_{upi}$$

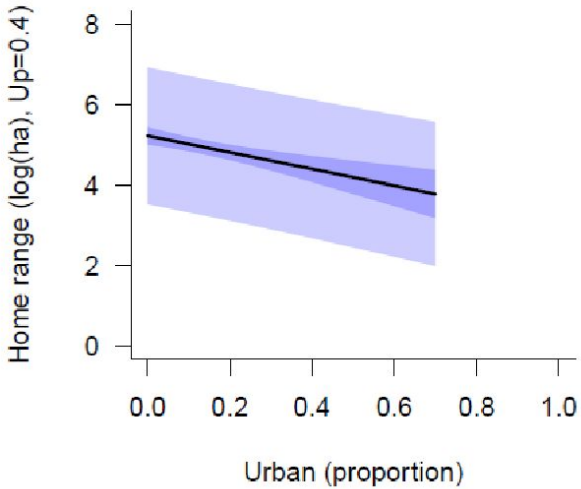
- We can visualize our model by holding all but one X constant
  - Typically we use the mean of Xs we are holding constant

Example

$$y_i = 4.85 - 2.07X_{urbi} + 0.86 \times 0.44$$

- Show relationship and uncertainty in relationship

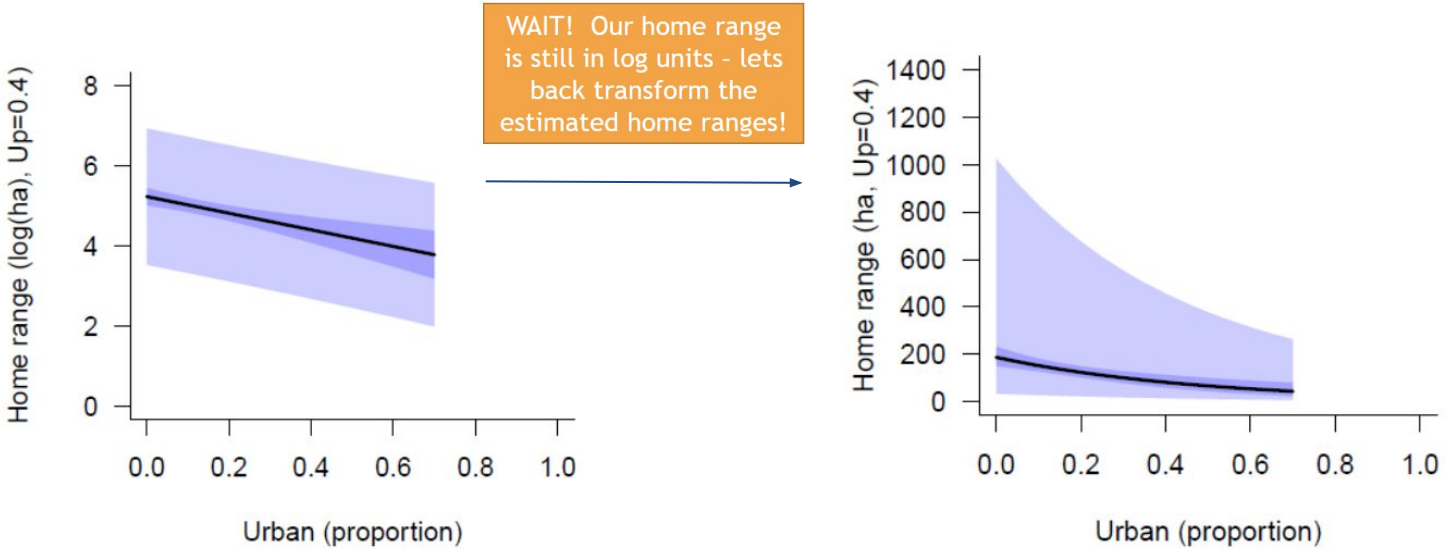
```
> #Predict Urban relationship
> CI.urb <- predict(mTop, newdata=urb.df, interval="confidence")
> PI.urb <- predict(mTop, newdata=urb.df, interval="prediction")
```



Example

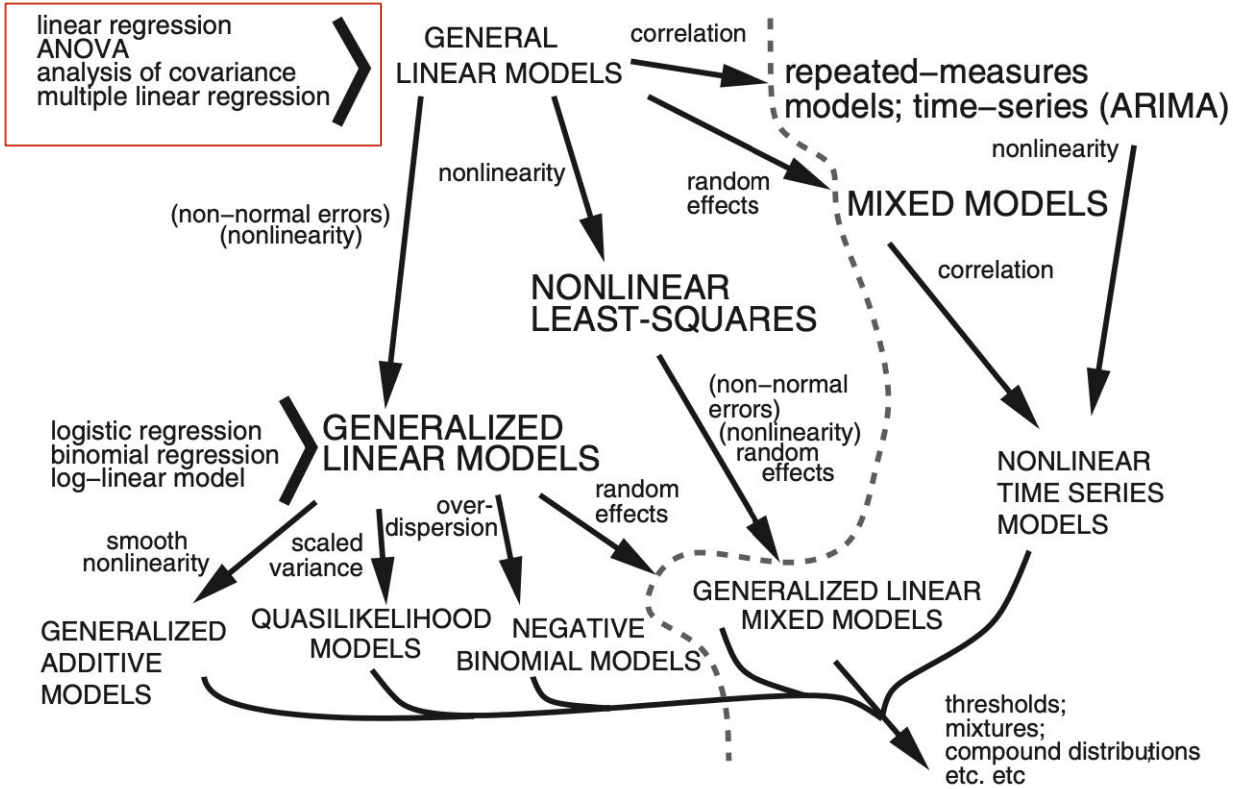
$$y_i = 4.85 - 2.07X_{urbi} + 0.86 \times 0.44$$

- Show relationship and uncertainty in relationship



2. Beyond the Linear Model

Covered Today!



General Linear Models

- GLM/GLMM: The General Linear Model/General Linear Mixed Model, or General Multivariate Regression Model is simply a compact way of simultaneously writing several multiple linear regression models.
- GLM is a flexible generalization of ordinary linear regression that **allows for response variables that have error distribution models other than a normal distribution**

HW

Use one of the techniques we discussed today to model a dataset of your interest.

Thanks for bearing with me!