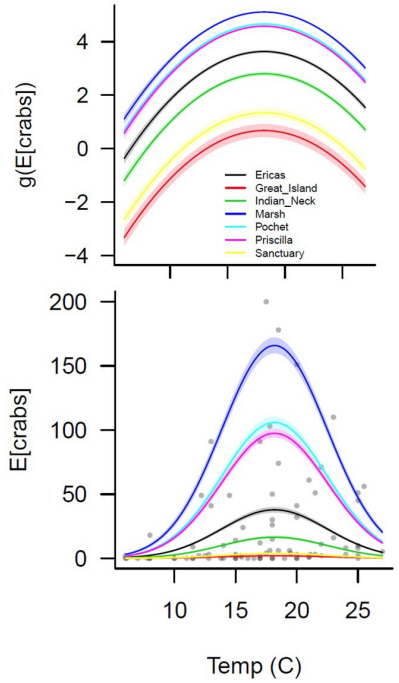
ECO 636
Applied Ecological Statistics

Week 3 – Linear models with multiple categorical predictors



Meg Graham MacLean, PhD

Department of Environmental Conservation

mgmaclean@umass.edu

2021 - Spring

The Week

Tuesday

- Review last week's recorded material
- Linear models with multiple categorical predictors examples

Wednesday (Lab)

- One-way ANOVA

Thursday

- Continue with multiple-categorical predictors
- Model selection: AIC

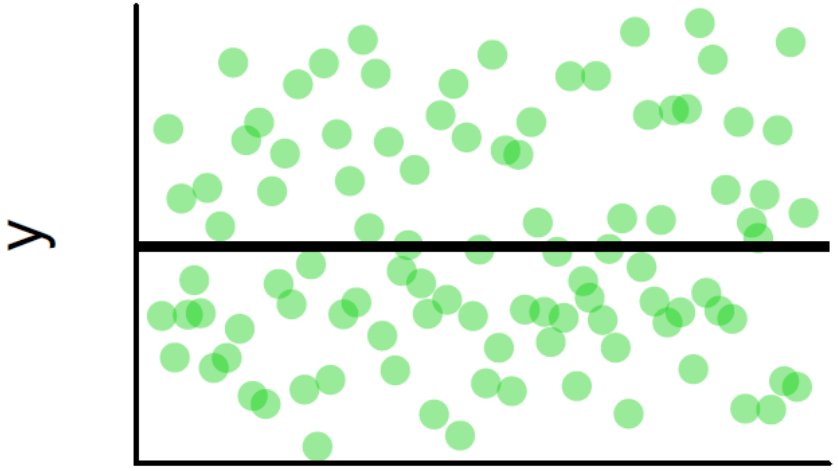
Review!

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?



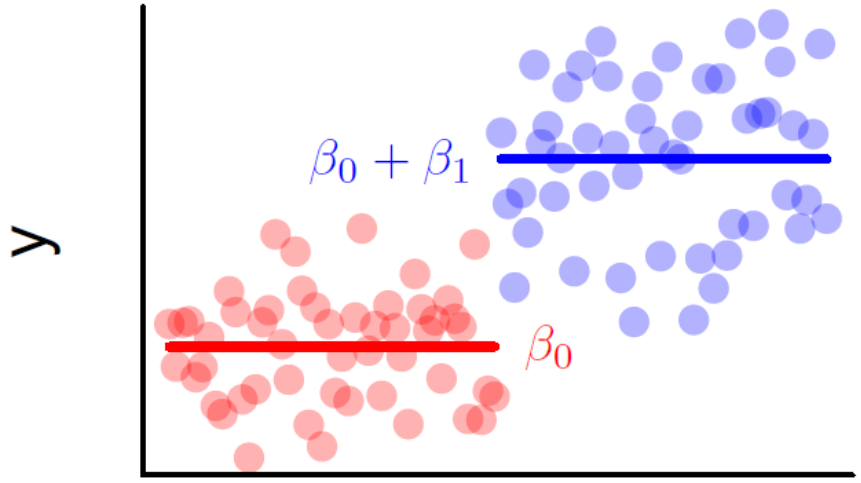
Review!

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?



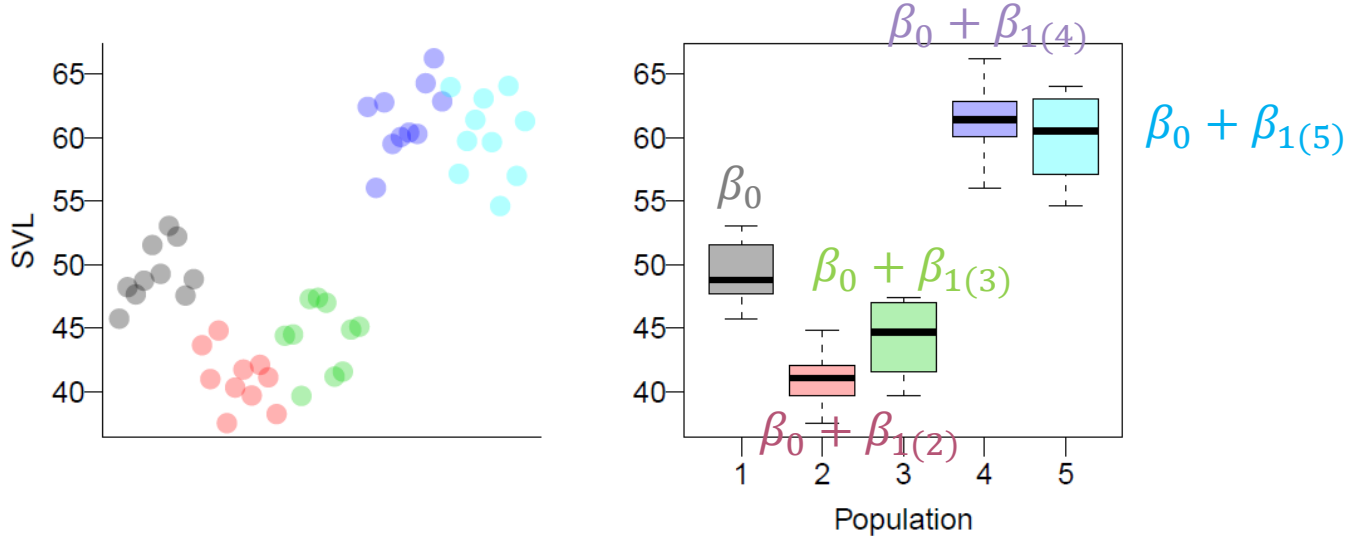
Review!

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?



Example

6. Interpret results

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

```
> 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

6. 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

F-statistic

$$F = \frac{explained\ variance}{residual\ variance} = \frac{SS_{model}}{SS_{residuals}}$$

Does the model outperform "random noise"? The higher the F, the more likely it does.

H_0 = the model is not significantly different than random noise at predicting y

```
> anova(lm(DBH~Stand,data = tree))
Analysis of Variance Table

Response: DBH
           Df Sum Sq Mean Sq F value    Pr(>F)
Stand       1  13374 13373.8  38.416 1.248e-07 ***
Residuals 48  16710   348.1
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

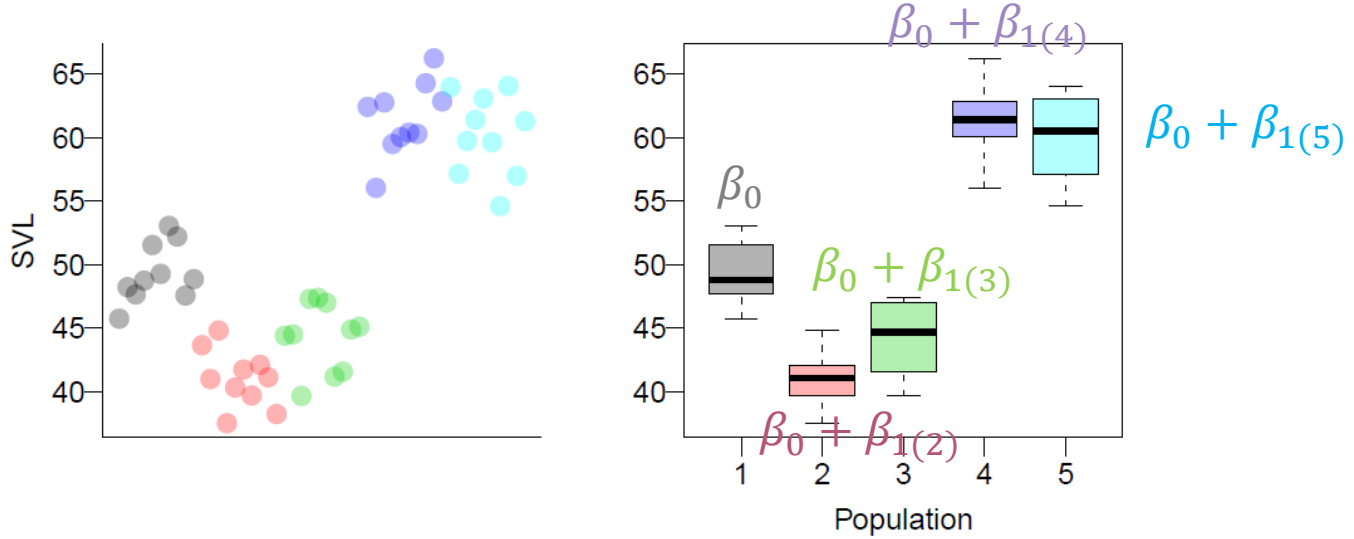
Review!

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?



One-way ANOVA as a linear model

ANOVA

- Single categorical explanatory variable (one-way)
- Generalization of the t-test for >2 groups
- Comparing differences in population means across multiple groups

Example: salamander lengths again! (in week 1 on Moodle)

- 4 salamander populations/sites of interest
- Question: Does SVL differ among salamander populations?

Example

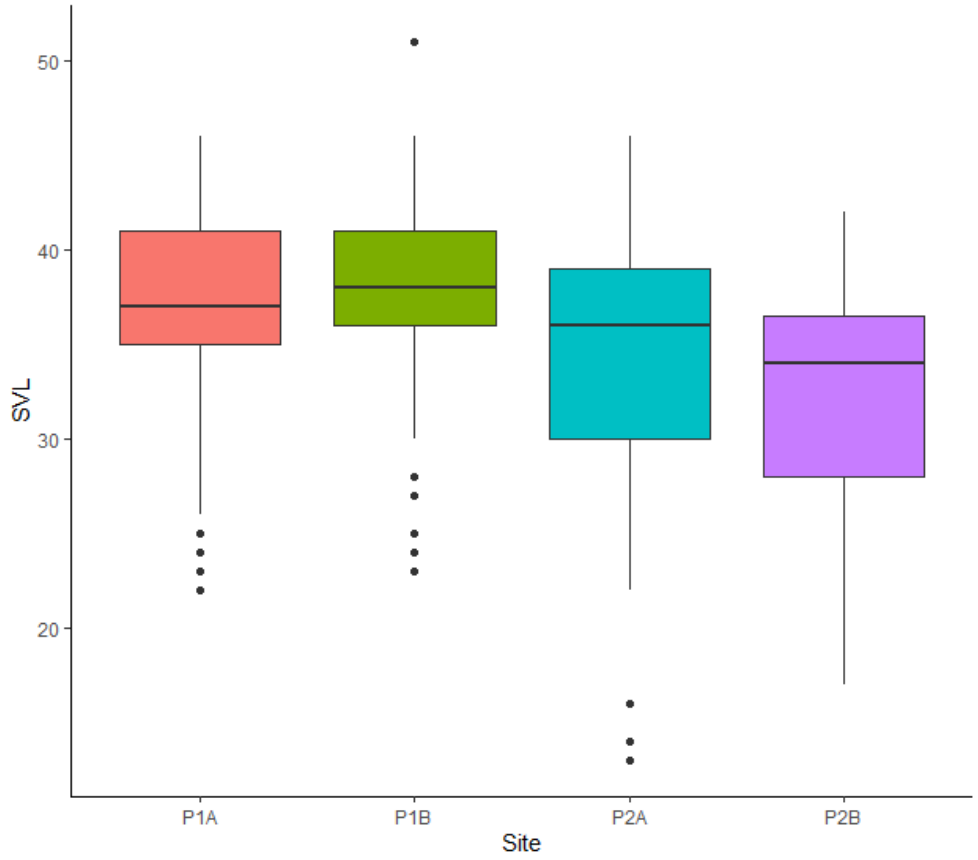
Modeling process:

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

Example

1. State the question:

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



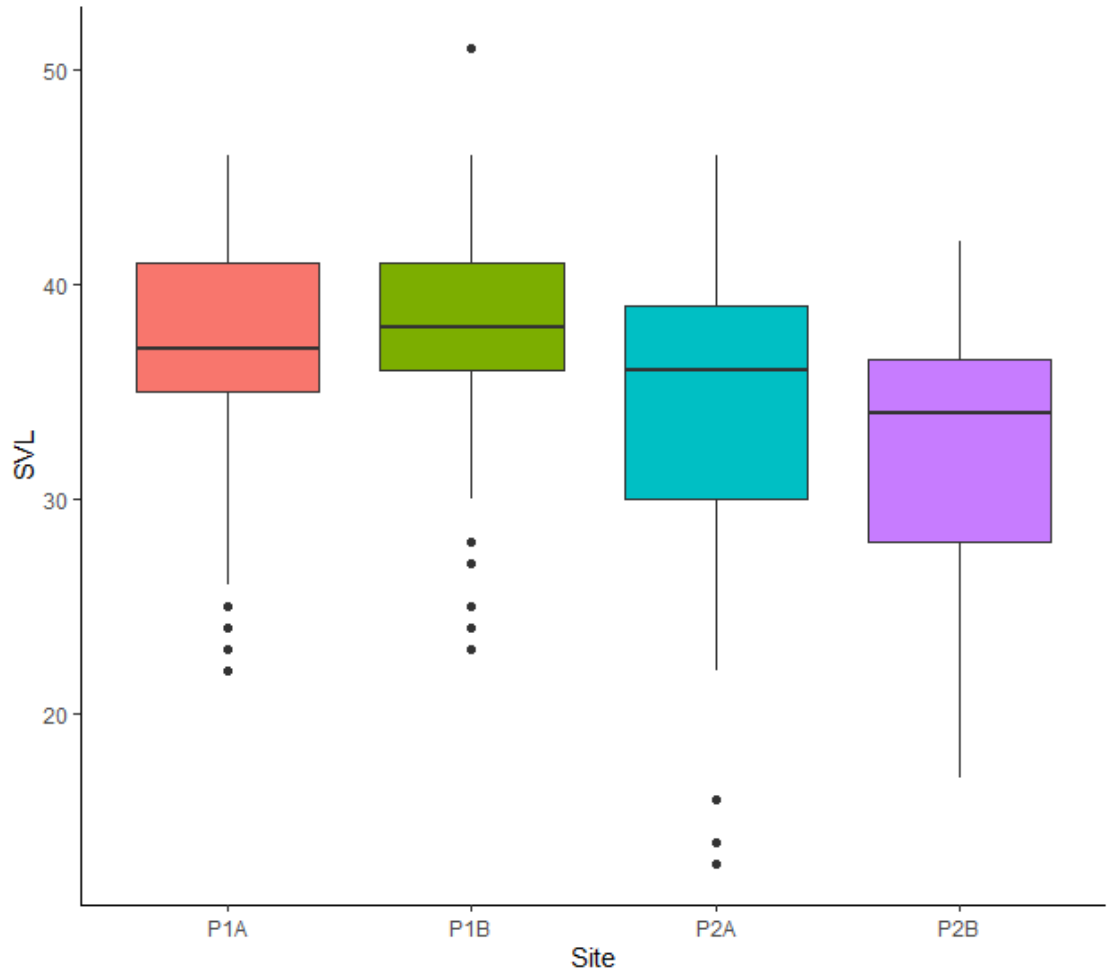
Example

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

Example

2. Data exploration
   - P1A mean: 37.03
   - P1B mean: 38.11
   - P2A mean: 34.79
   - P2B mean: 32.85



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

Example

3. Describe the model:

- In words:
  - Is there a difference between the 4 population means of SVL?
  - H_0: there is no difference!
- In mathematical form:
  - $y_i = \beta_0 + \beta_{1(g)} X_{i(g)} + e_i$
  - y_i is SVL
  - $X_{i(g)}$ is the population (aka Site)

?

Multiple samples

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

- β_0 is the mean of group 1
- $\beta_{1(g)}$ is the difference between group g and group 1
  - In our example we have: $\beta_{1(g=2)}$, $\beta_{1(g=3)}$, $\beta_{1(g=4)}$
- $e_i \sim N(0, \sigma)$

Example

3. Describe the model:

- In words:
  - Is there a difference between the 4 population means of SVL?
  - H_0: there is no difference!
- In mathematical form:
  - $y_i = \beta_0 + \beta_{1(g)} X_{i(g)} + e_i$
  - y_i is SVL
  - $X_{i(g)}$ is the population (aka group)
- What are the model assumptions?
  - Residuals are normally distributed
  - Constant variance (homogeneity)
  - Observations are independent
  - Predictors measured without error (fixed X)

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

Example

4. Fit the model

- Algebraically
  - $y_i = \beta_0 + \beta_{1(g)} X_{i(g)} + e_i$
  - y_i is SVL
  - $X_{i(g)}$ is the population (aka Site)
- In R:

In groups, duplicate the code from the exercise last week and modify it to run with the salamanders and interpret the output!

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

Example

5. Evaluate the output

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

```
summary(mSite)
```

```
##
## Call:
## lm(formula = SVL ~ Site, data = sally)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -21.7857  -3.0276   0.8925   3.9724  12.8925
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  37.0276     0.4140  89.441  < 2e-16 ***
## SiteP1B       1.0799     0.7106   1.520    0.129
## SiteP2A      -2.2419     0.5567  -4.027 6.40e-05 ***
## SiteP2B      -4.1743     0.7649  -5.458 7.21e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.57 on 569 degrees of freedom
##   (5 observations deleted due to missingness)
## Multiple R-squared:  0.08546,    Adjusted R-squared:  0.08064
## F-statistic: 17.72 on 3 and 569 DF,  p-value: 5.2e-11
```

Example

5. Evaluate the output

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

```
mSite <- lm(SVL ~ Site, data = sally)
#estimated coefficients
coef(mSite)
```

```
## (Intercept)     SiteP1B     SiteP2A     SiteP2B
##   37.027624    1.079903   -2.241910   -4.174291
```

```
#empirical means
obs.means
```

```
##       P1A       P1B       P2A       P2B
## 37.02762 38.10753 34.78571 32.85333
```

Example

5. Evaluate the output

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

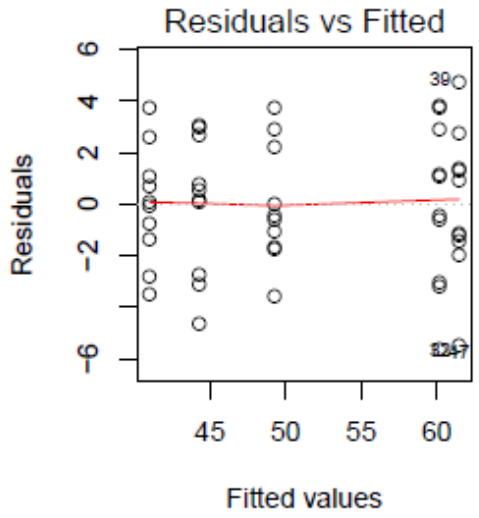
- Check assumptions:
  - Residuals are normally distributed
  - Constant variance (homogeneity)
  - Observations are independent
  - Predictors measured without error (fixed X)

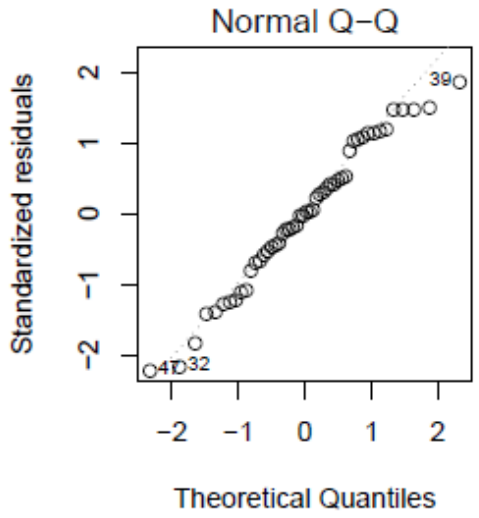
Example

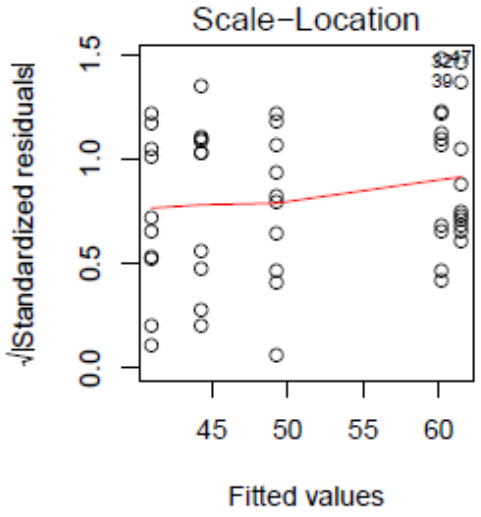
5. Evaluate the output

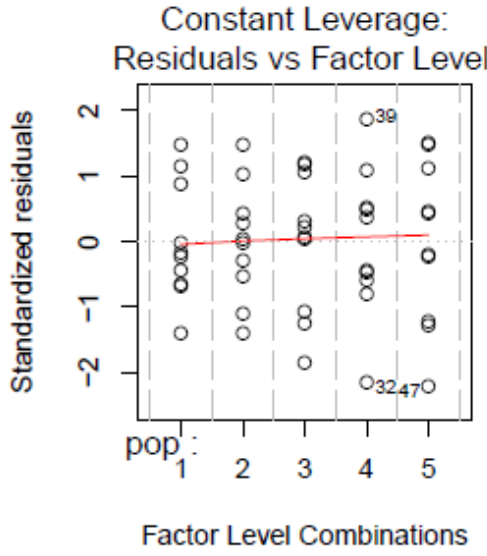
- Check assumptions:

```
> par(mfrow=c(2,2), oma=c(0,0,0,0))
> plot(mPop)
```





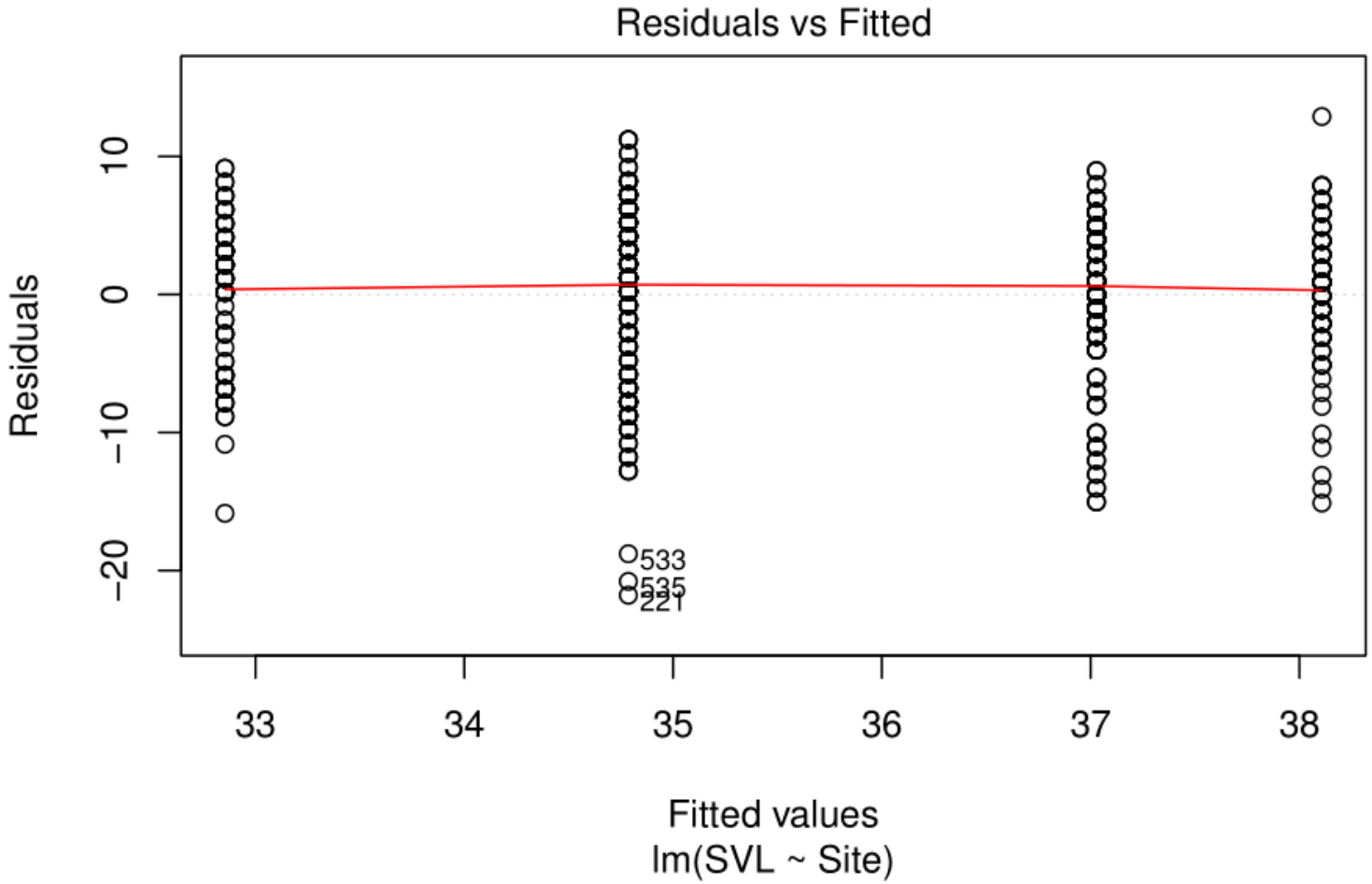




Example

5. Evaluate the output

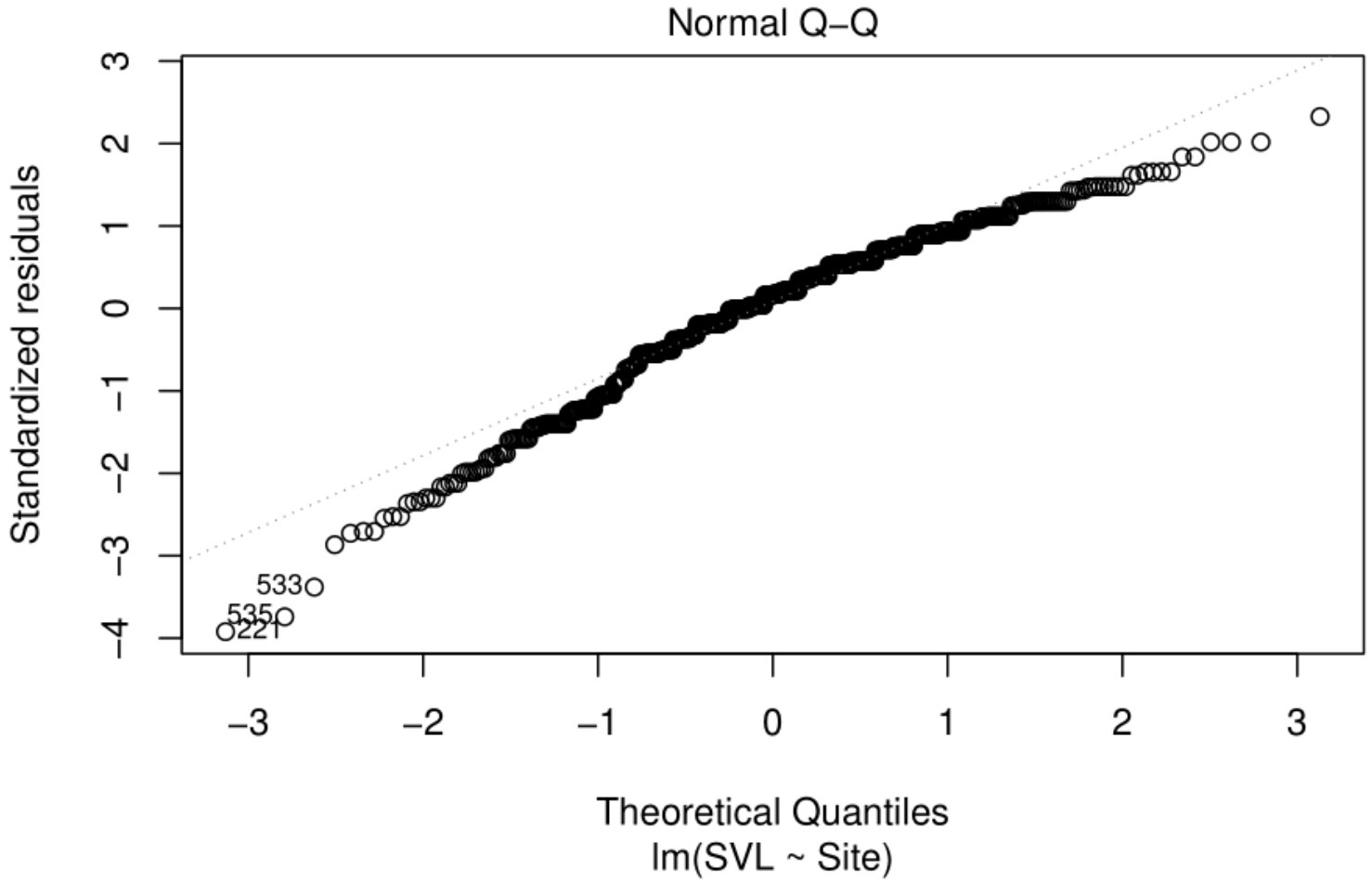
$$y_i = \beta_0 + \beta_{1(g)} X_{i(g)} + e_i$$



Example

5. Evaluate the output

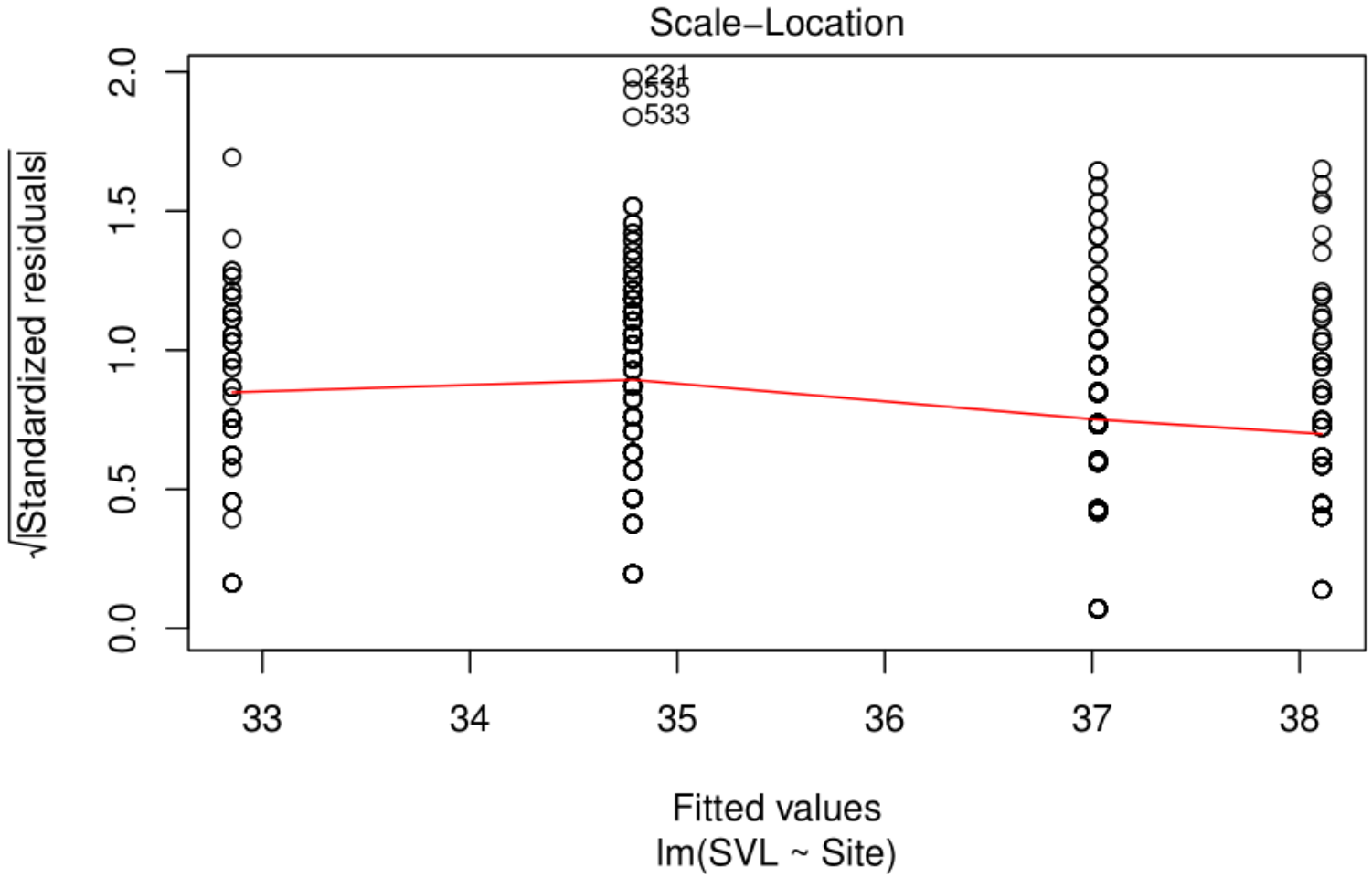
$$y_i = \beta_0 + \beta_{1(g)} X_{i(g)} + e_i$$



Example

5. Evaluate the output

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

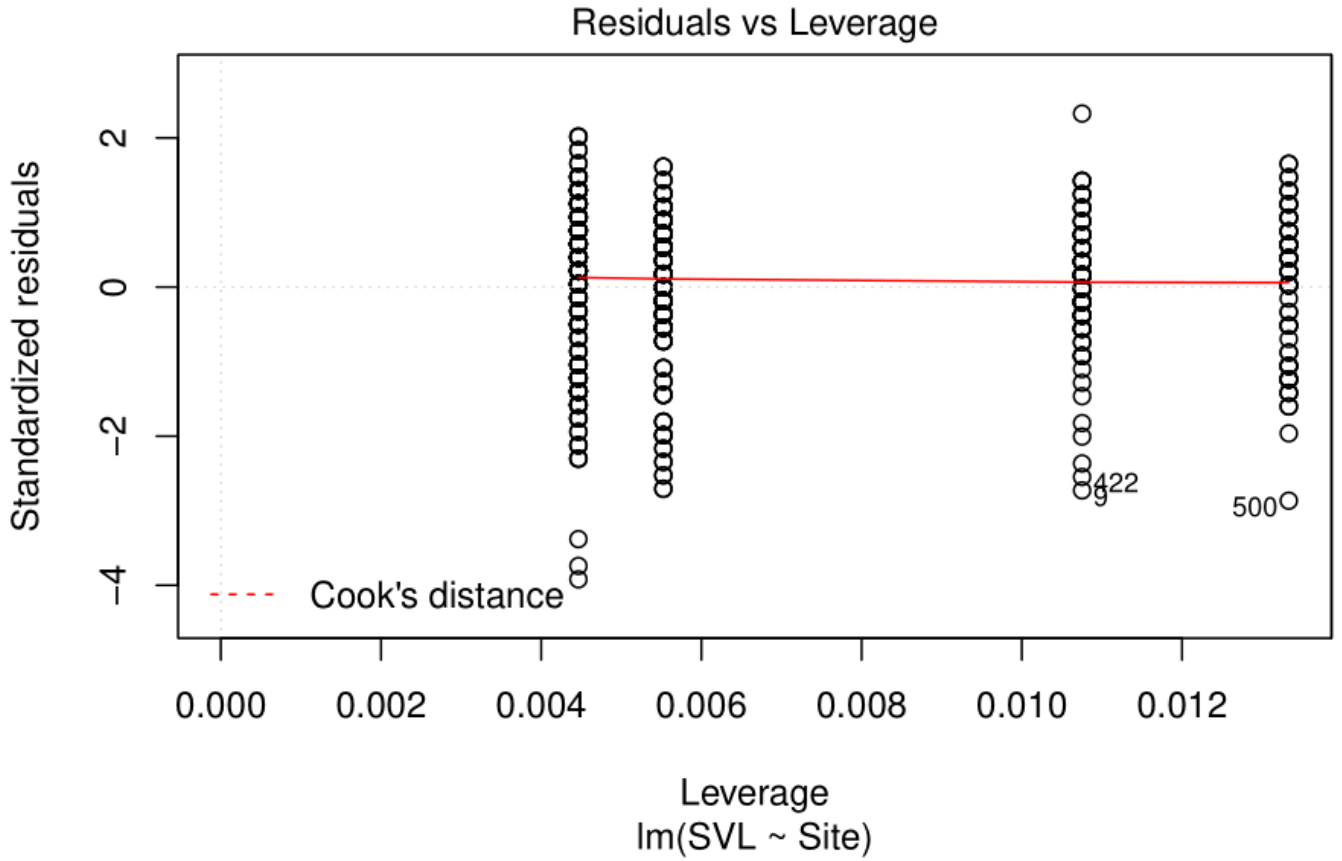


Example

5. Evaluate the output

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

Attempts to identify influential cases



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

Example

6. Interpret the results

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

```
summary(mSite)
```

```
##
## Call:
## lm(formula = SVL ~ Site, data = sally)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -21.7857  -3.0276   0.8925   3.9724  12.8925
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  37.0276     0.4140  89.441  < 2e-16 ***
## SiteP1B       1.0799     0.7106   1.520    0.129
## SiteP2A      -2.2419     0.5567  -4.027 6.40e-05 ***
## SiteP2B      -4.1743     0.7649  -5.458 7.21e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.57 on 569 degrees of freedom
##   (5 observations deleted due to missingness)
## Multiple R-squared:  0.08546,    Adjusted R-squared:  0.08064
## F-statistic: 17.72 on 3 and 569 DF,  p-value: 5.2e-11
```

Example

6. Interpret the results

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

```
summary(mSite)$coefficients
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 37.027624  0.4139881 89.441278 0.000000e+00
## SiteP1B      1.079903  0.7105942  1.519718 1.291374e-01
## SiteP2A     -2.241910  0.5566617 -4.027419 6.404214e-05
## SiteP2B     -4.174291  0.7648516 -5.457648 7.211364e-08
```

- Intercept is…
  - The mean of SiteP1A, or β_0
- What about the rest?
  - SiteP1B is the difference between SiteP1B and SiteP1A, or or $\beta_{1(2)}$
  - SiteP2A is the difference between SiteP2A and SiteP1A, or or $\beta_{1(3)}$
  - SiteP2B is the difference between SiteP2B and SiteP1A, or or $\beta_{1(4)}$

What if we are interested in the pairwise comparison between all of the means?

Example

6. Interpret the results – pairwise comparisons

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

- Option 1: releveling

```
sally2 <- sally
sally2$Site <- relevel(sally2$Site, ref = 2)
mSite2 <- lm(SVL ~ Site, data = sally2)
summary(mSite2)
```

```
##
## Call:
## lm(formula = SVL ~ Site, data = sally2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -21.7857  -3.0276   0.8925   3.9724  12.8925
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  38.1075     0.5775  65.982  < 2e-16 ***
## SiteP1A      -1.0799     0.7106  -1.520    0.129
## SiteP2A      -3.3218     0.6871  -4.835 1.72e-06 ***
## SiteP2B      -5.2542     0.8644  -6.078 2.22e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Example

6. Interpret the results – pairwise comparisons

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

- Option 1: releveling

```
sally2 <- sally
sally2$Site <- relevel(sally2$Site, ref = 2)
mSite2 <- lm(SVL ~ Site, data = sally2)
summary(mSite2)
```

- BUT! This can lead to Type I errors, we need to adjust for multiple comparisons!
  - What are you used to using with ANOVA models to look at pairwise comparisons?
    - TukeyHSD!

Example

6. Interpret the results – pairwise comparisons

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

- Option 1: releveling
  - Can produce Type I errors
- Option 2: Tukey HSD

```
aov.mSite <- aov(mSite)              # ANOVA table
(tuk.mSite <- TukeyHSD(aov.mSite)) # pairwise comparisons
```

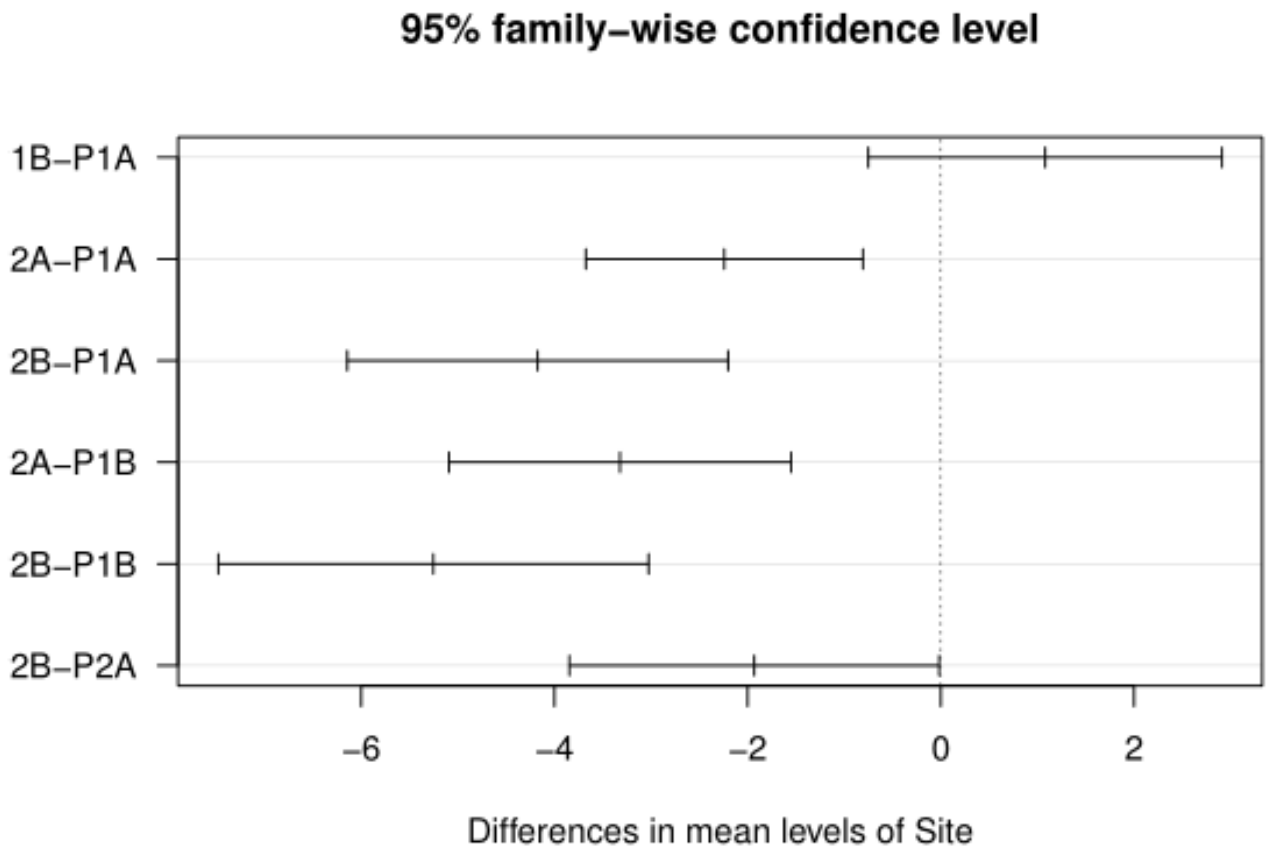
```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = mSite)
##
## $Site
##               diff        lwr         upr     p adj
## P1B-P1A  1.079903 -0.7510421  2.91084727 0.4262958
## P2A-P1A -2.241910 -3.6762263 -0.80759376 0.0003720
## P2B-P1A -4.174291 -6.1450375 -2.20354444 0.0000004
## P2A-P1B -3.321813 -5.0921044 -1.55152080 0.0000102
## P2B-P1B -5.254194 -7.4814142 -3.02697291 0.0000000
## P2B-P2A -1.932381 -3.8469087 -0.01785322 0.0469235
```

Example

6. Interpret the results – pairwise comparisons

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

```
plot(tuk.mSite)
```



Example

6. Interpret the results

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

```
summary(mSite)
```

```
##
## Call:
## lm(formula = SVL ~ Site, data = sally)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -21.7857  -3.0276   0.8925   3.9724  12.8925
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  37.0276     0.4140  89.441  < 2e-16 ***
## SiteP1B       1.0799     0.7106   1.520    0.129
## SiteP2A      -2.2419     0.5567  -4.027 6.40e-05 ***
## SiteP2B      -4.1743     0.7649  -5.458 7.21e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.57 on 569 degrees of freedom
##   (5 observations deleted due to missingness)
## Multiple R-squared:  0.08546,    Adjusted R-squared:  0.08064
## F-statistic: 17.72 on 3 and 569 DF,  p-value: 5.2e-11
```

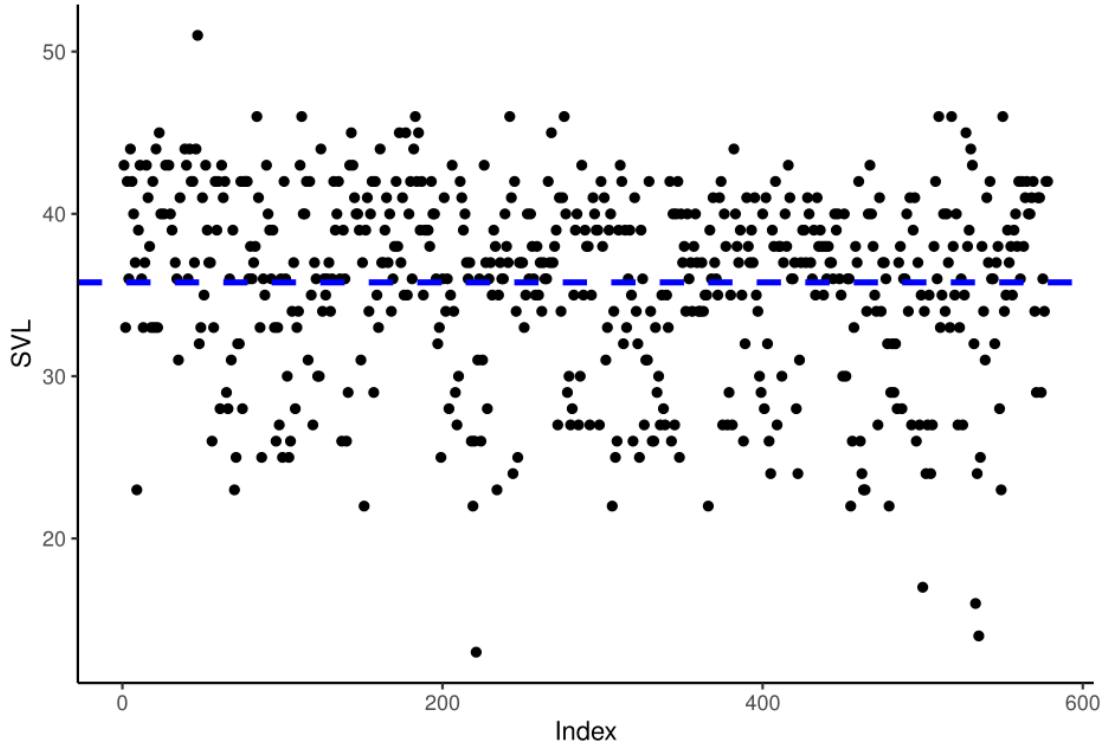
Example

6. Interpret the results

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

- Is the grouping variable (population) useful in predicting mean?
  - i.e. is the grouping significant?
  - Without the grouping variable we have the null model: $y_i = \beta_0 + e_i$

```
m0 <- lm(SVL ~ 1, data =  sally)
```



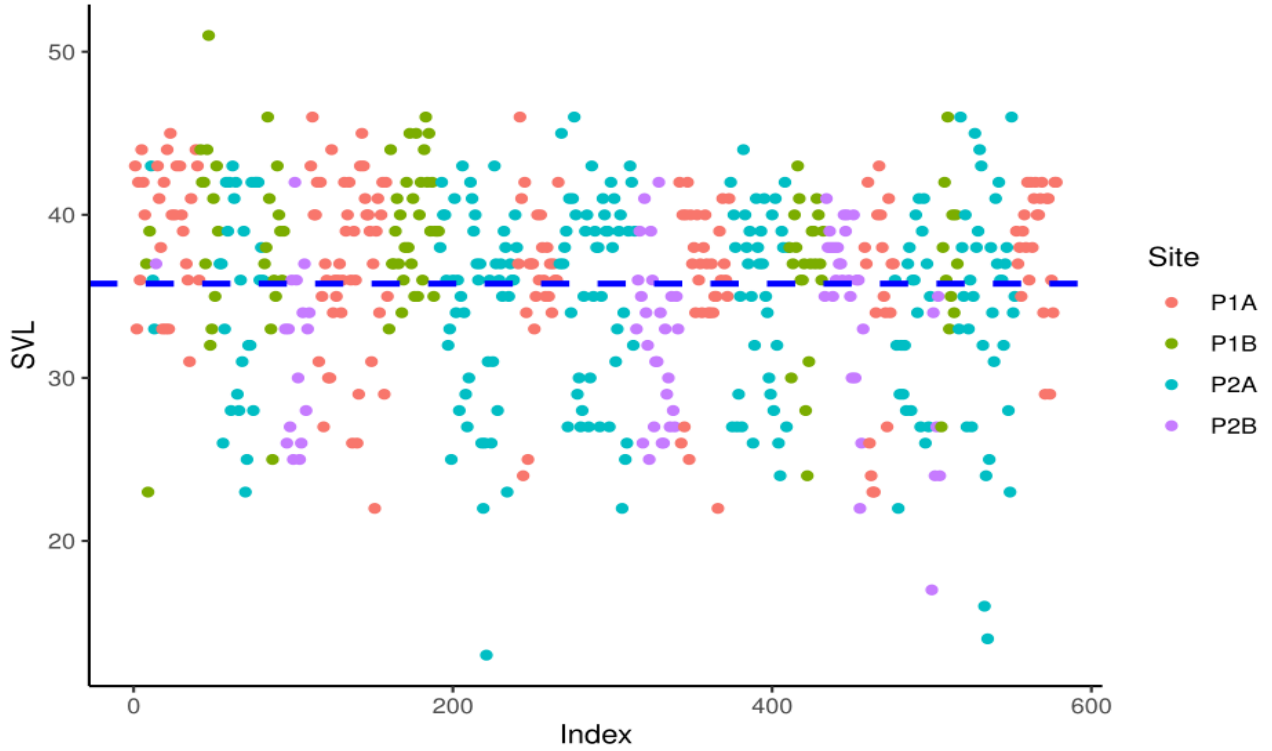
Example

6. Interpret the results

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

- Is the grouping variable (population) useful in predicting mean?
  - i.e. is the grouping significant?
  - Without the grouping variable we have the null model: $y_i = \beta_0 + e_i$

```
m0 <- lm(SVL ~ 1, data =  sally)
```

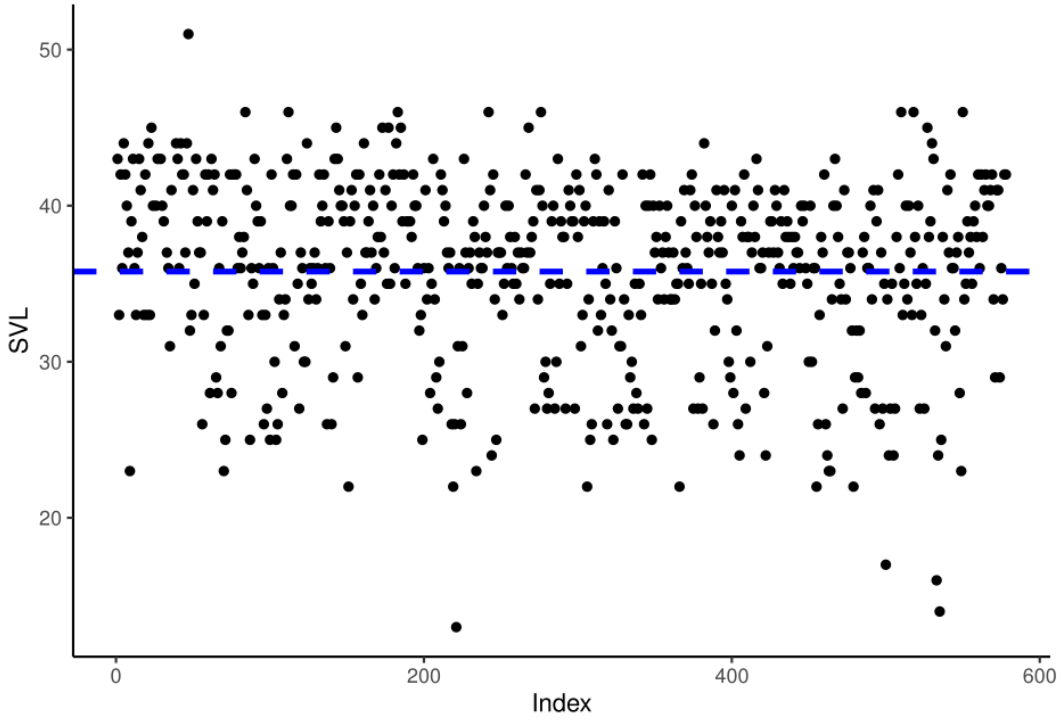


Example

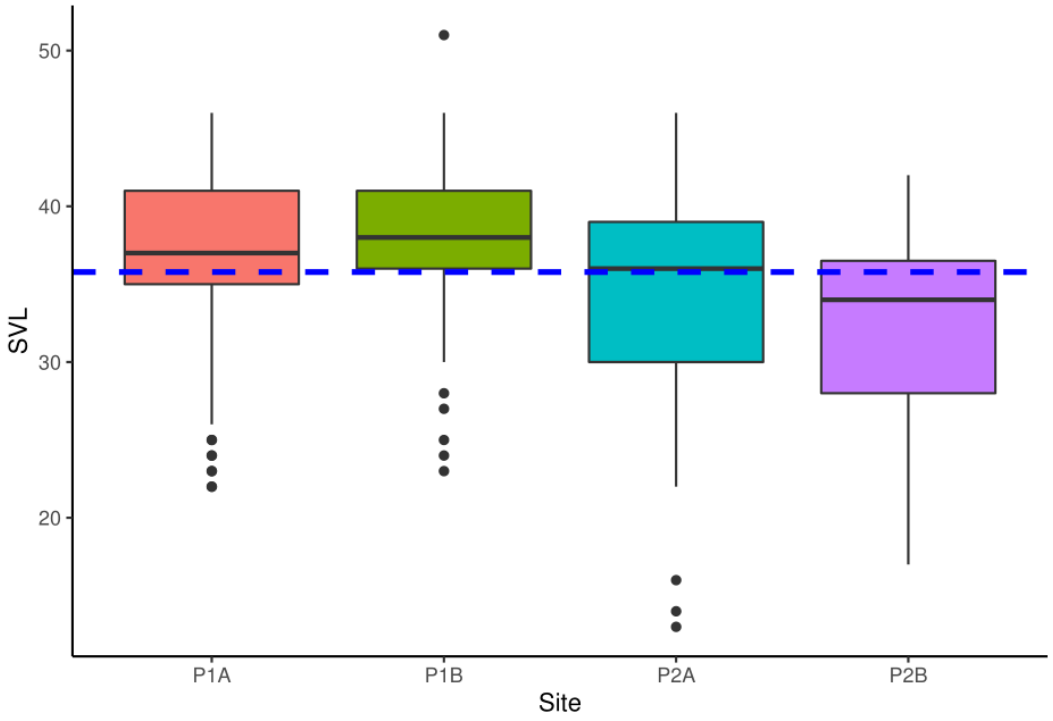
6. Interpret the results

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

- Is the grouping variable (population) useful in predicting mean?
  - i.e. is the grouping significant?
  - Without the grouping variable we have the null model: $y_i = \beta_0 + e_i$



Multiple



Example

6. Interpret the results

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

- Is the grouping variable (population) useful in predicting mean?
  - i.e. is the grouping significant?
  - Without the grouping variable we have the null model: $y_i = \beta_0 + e_i$
  - To test significance of grouping factor, we use sum of squares
    - Which is ANOVA!

```
anova(mSite)
```

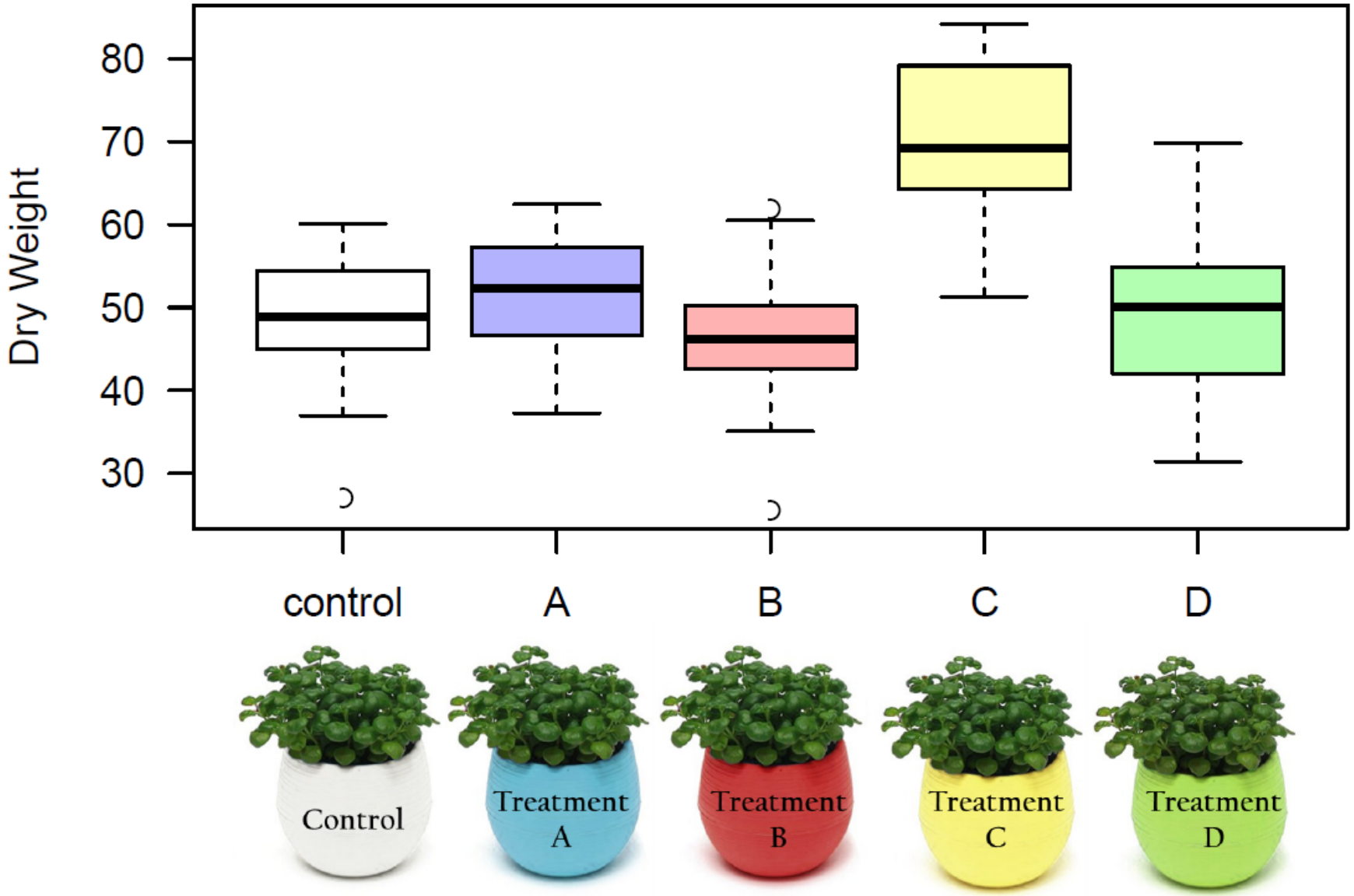
```
## Analysis of Variance Table
##
## Response: SVL
##            Df  Sum Sq Mean Sq F value  Pr(>F)
## Site        3  1649.4  549.80  17.724 5.2e-11 ***
## Residuals 569 17650.9   31.02
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Example 2 - controls

What if we have a case where we have treatments and a control?

- For example, we have four nutrient treatments to plants, and one control (no added nutrients), and we measure productivity (dry mass in grams).
  - Question: do our treatments influence biomass production?

Example 2 - controls



Example 2 - controls

Are there significant differences between the control and the treatment dry weights?

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

```
> treatMod <- lm(dryWt ~ treatment)
```

Example 2 - controls

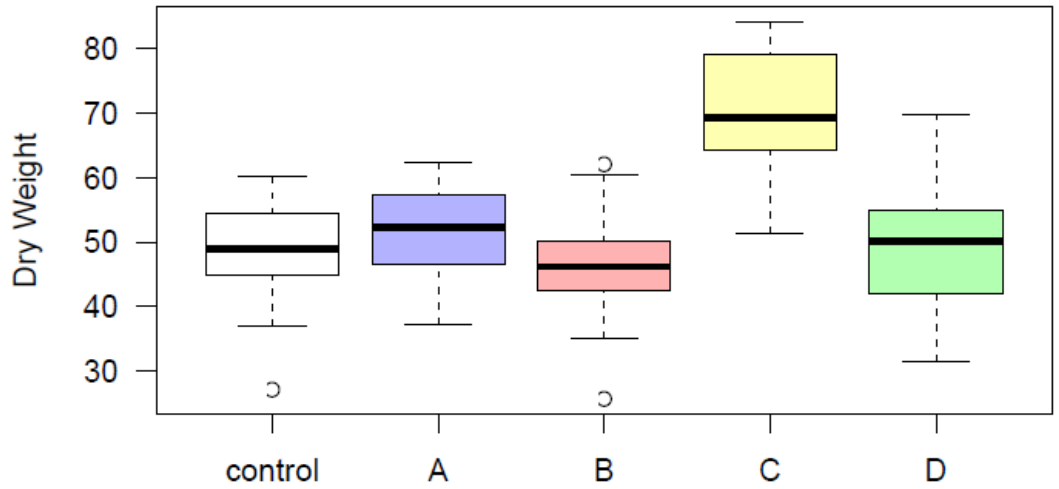
$$y_i = \beta_0 + \beta_{1(g)} X_{i(g)} + e_i$$

```
> summary(treatMod)

Call:
lm(formula = dryWt ~ treatment)

Residuals:
     Min       1Q   Median       3Q      Max
-21.5753  -4.9573   0.5088   5.7503  20.3114

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)        51.589      1.476  34.941  < 2e-16 ***
treatmentB         -4.930      2.088  -2.361   0.0196 *
treatmentC         18.297      2.088   8.763 4.61e-15 ***
treatmentControl   -2.877      2.088  -1.378   0.1704
treatmentD         -2.050      2.088  -0.982   0.3279
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.087 on 145 degrees of freedom
Multiple R-squared:  0.5306,  Adjusted R-squared:  0.5177
F-statistic: 40.98 on 4 and 145 DF,  p-value: < 2.2e-16
```

Example 2 - controls


```
> treatment <- relevel(treatment,ref="Control")
> treatMod <- lm(dryWt ~ treatment)
> round(summary(treatMod)$coefficients,3)
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    48.712      1.476  32.993    0.000
treatmentA      2.877      2.088   1.378    0.170
treatmentB     -2.053      2.088  -0.983    0.327
treatmentC     21.174      2.088  10.141    0.000
treatmentD      0.827      2.088   0.396    0.693
```

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



>1 explanatory variables multiple samples!

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

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)`

>1 explanatory variables multiple samples!

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

Next!

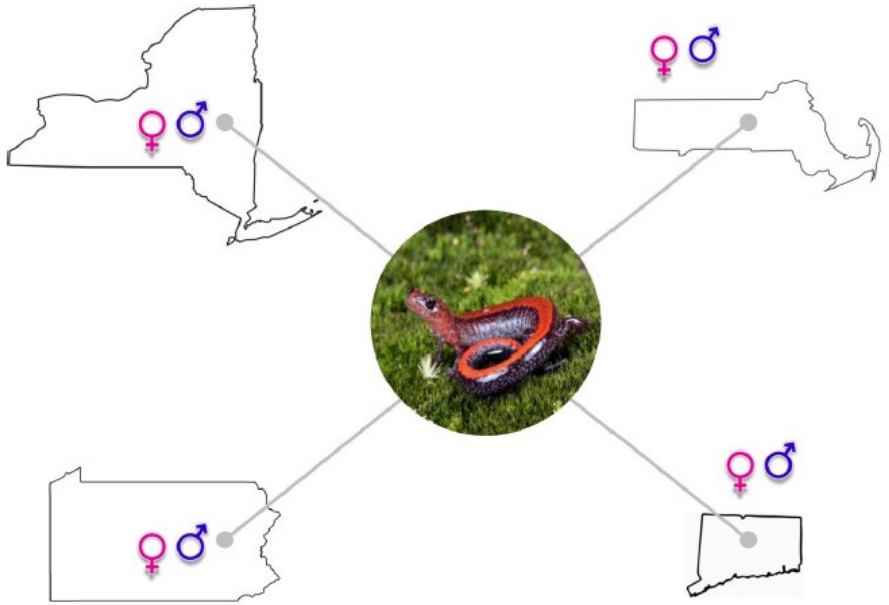
Response (Y)	Explanatory (X)	Model	In R
Continuous	None	Intercept-only/null	`lm(y~1)`
Continuous	Single two-level factor	*t*-test	`lm(y~x)`
Continuous	Single multi-level factor	One-way ANOVA	`lm(y~x)`
Continuous	>1 multi-level factor (+)	Two-way ANOVA	`lm(y~x1+x2)`

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

Two-way ANOVA as a linear model

Example for salamander lengths!

- 4 salamander populations of interest
- 2 sexes of interest
- Question: Does 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 as a linear model

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

- 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
- $e_i \sim N(0, \sigma)$

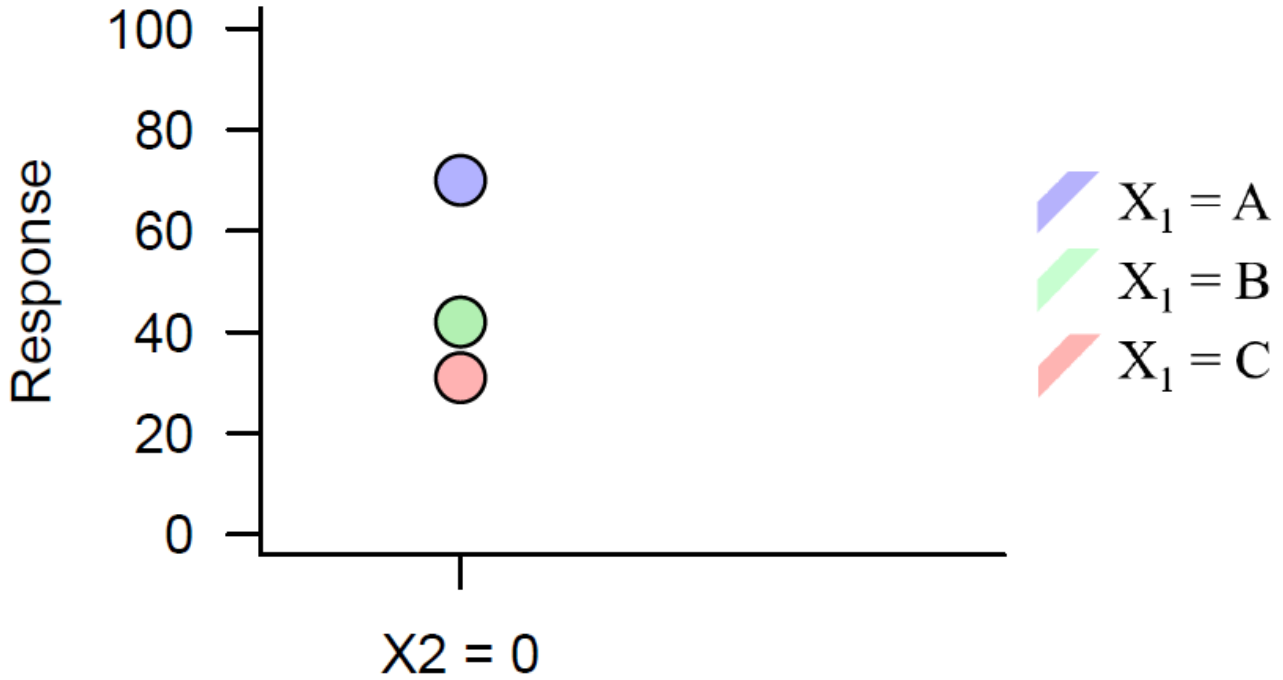
Two-way ANOVA as a linear model

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

What if we explore this graphically…

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

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

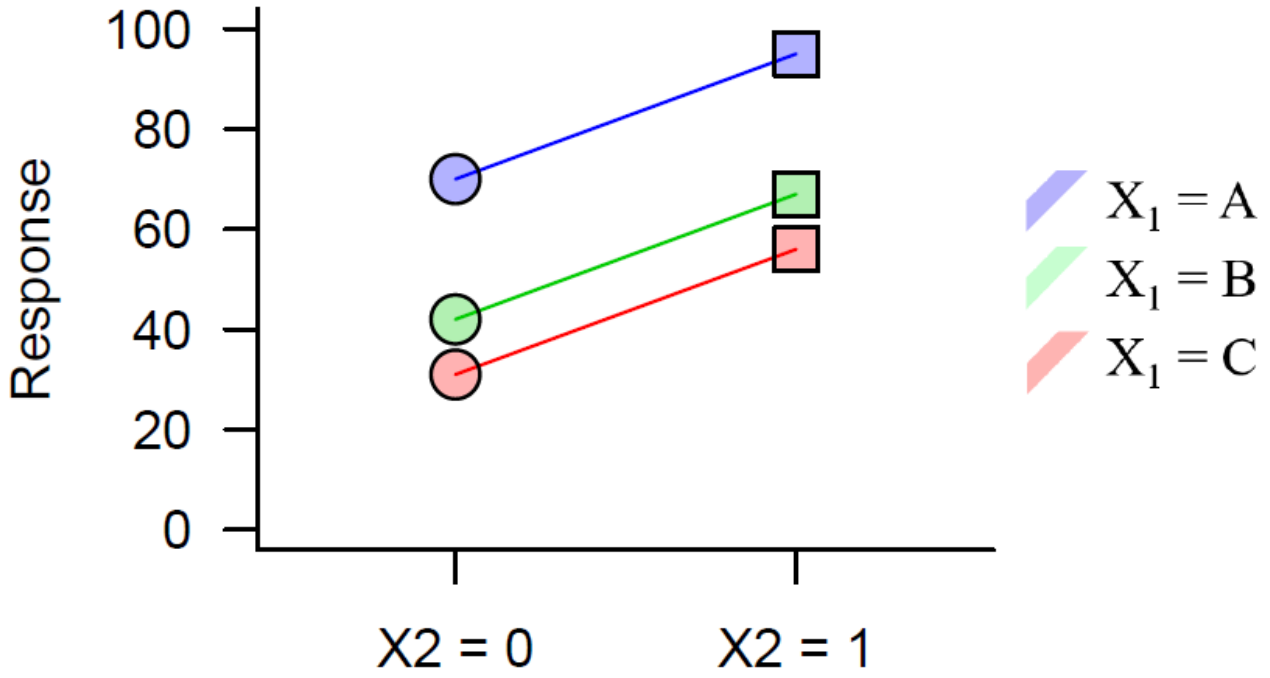


Two-way ANOVA as a linear model

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

What if we explore this graphically…

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

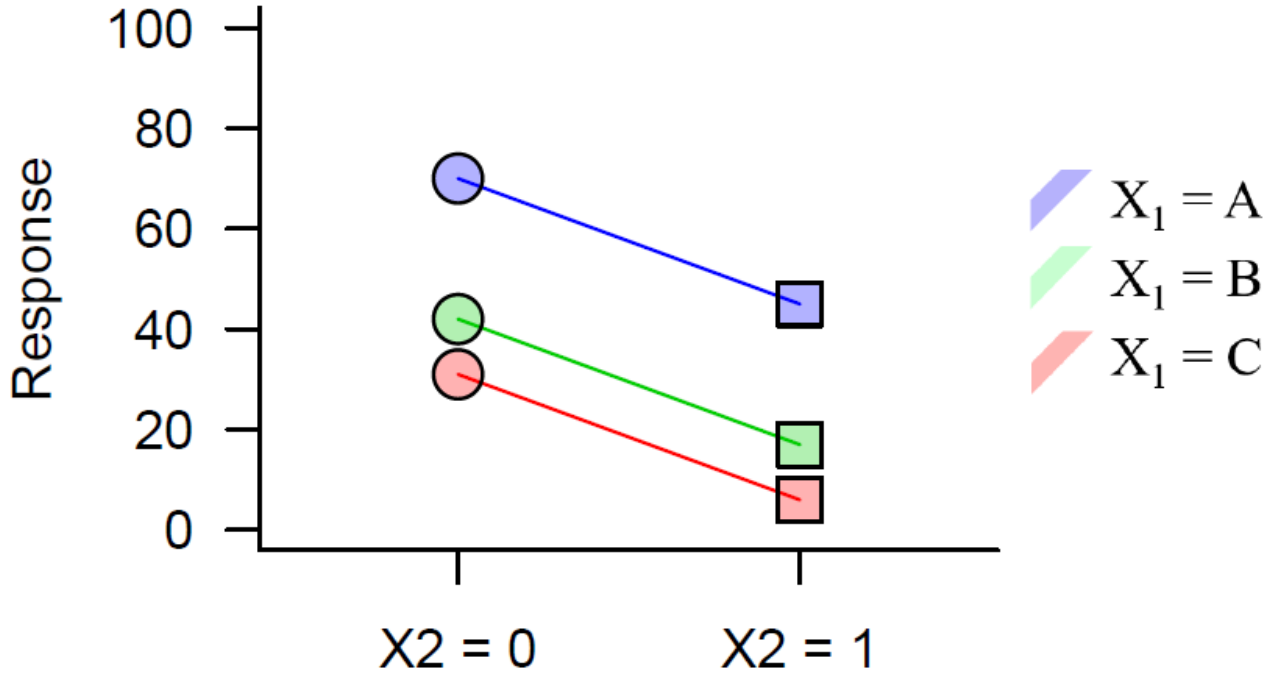


Two-way ANOVA as a linear model

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

What if we explore this graphically…

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



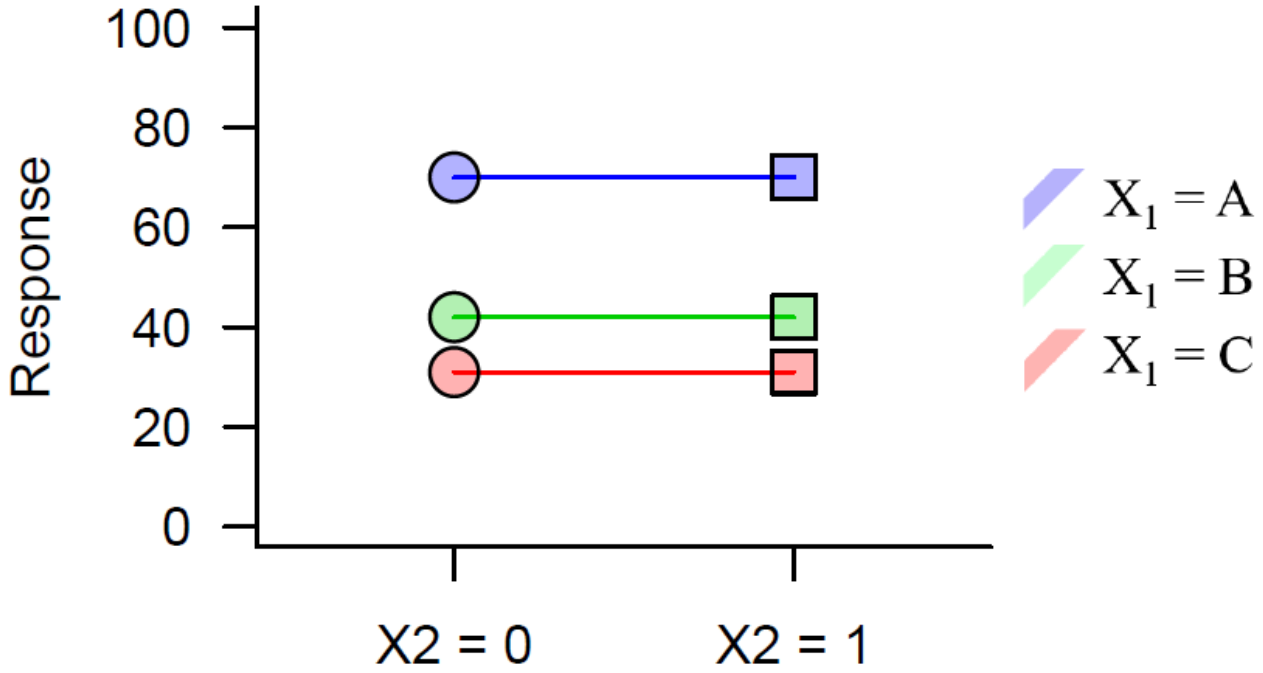
Two-way ANOVA as a linear model

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

What if we explore this graphically…

$$\beta_{2(g)} = 0$$

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



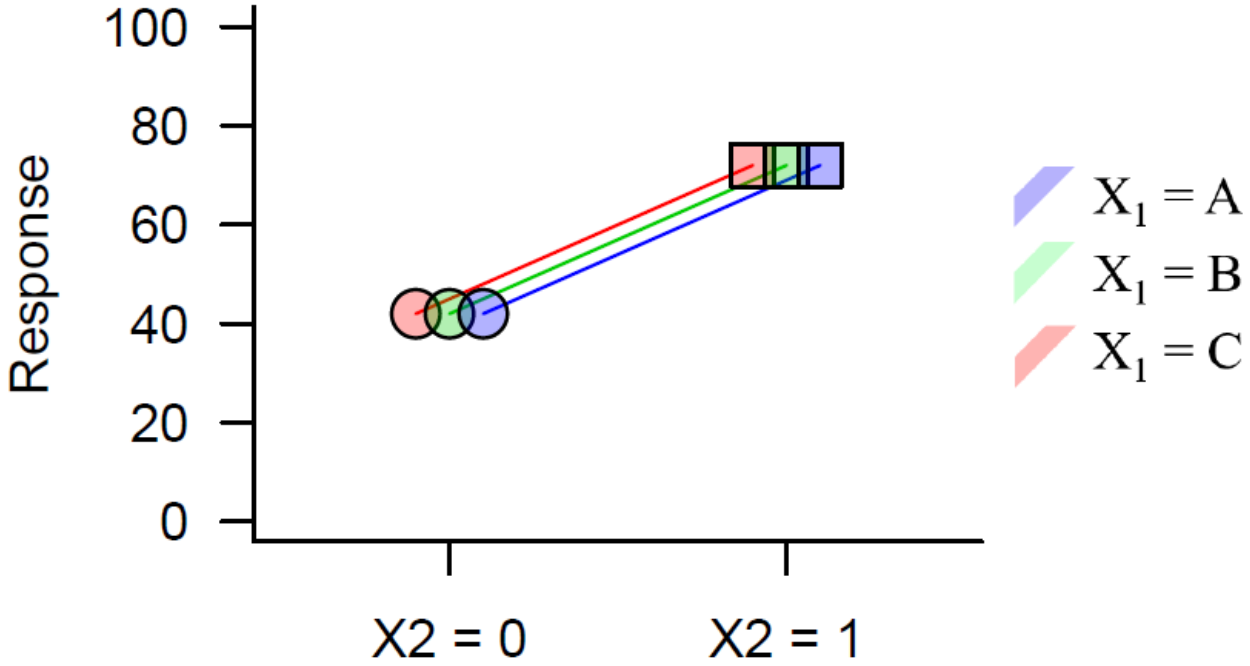
Two-way ANOVA as a linear model

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

What if we explore this graphically…

$$\beta_{1(g)} = 0$$

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



Two-way ANOVA as a linear model

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

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")
```

Two-way ANOVA as a linear model

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

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

Two-way ANOVA as a linear model

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

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

Two-way ANOVA as a linear model

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

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

```
> 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

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

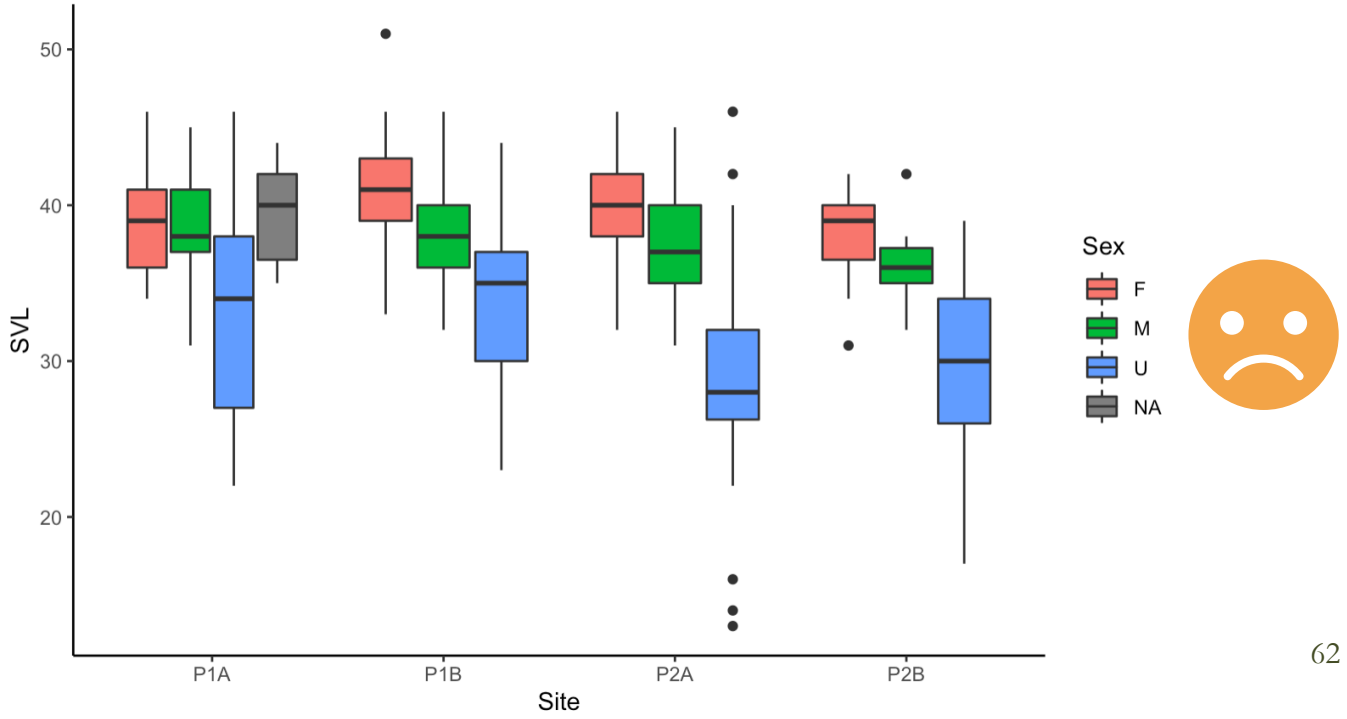
- β_0 is the mean of the first combination of factors – the reference level
  - **Females in Population A** <- super important to know your reference level
- What are the slopes ($\beta_{1(g)}$ and $\beta_{2(g)}$)?
- $\beta_{1(g)}$ is the group 1 contrasts – relate to the sex effect
  - $\beta_{1(g=male)}$ - 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

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

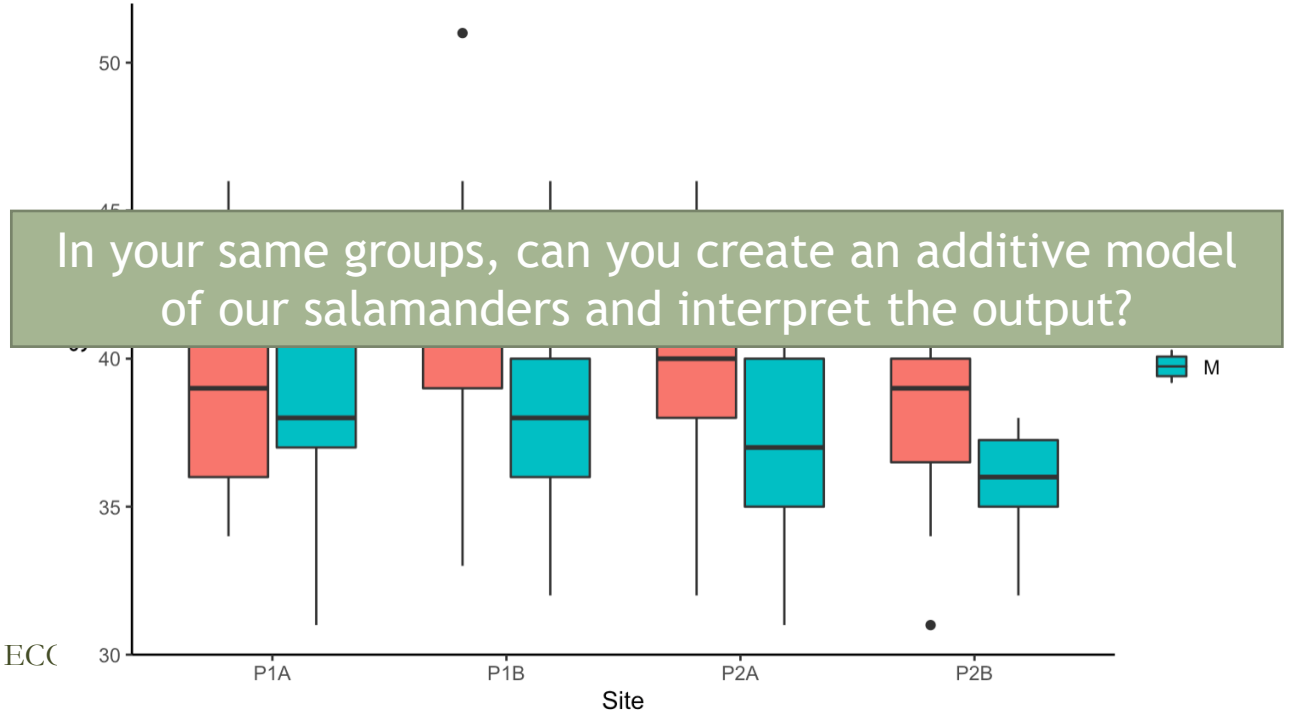


Two-way ANOVA as a linear model

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

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)



For next week:

1) Read section 5.2 in the Zuur book (2007).
2) Watch the recorded lecture and do the exercise
3) OPTIONAL: Read sections 7.1-7.7 in AndyFieldsRBook posted in Moodle R help files.
4) Finish the lab in week 3.
5) Complete the individual assessment on Moodle by 11:55pm Monday night.