CogLab: Making Inferences

WEEK 9

formative assignment #2

- descriptive statistics and plotting in R
  - due Nov 4

9	Tuesday, October 29, 2024	W9: Project Work
9	Thursday, October 31, 2024	W9 continued…
9	Monday, November 4, 2024	**Formative Assignment (R Descriptive) Due**
10	Tuesday, November 5, 2024	Weeks 10-12: Data Collection
10	Thursday, November 7, 2024	Weeks 10-12: Data Collection
10	Monday, November 11, 2024	**Project Milestone #5 (Pre-Registration + Checklist) Due**
11	Tuesday, November 12, 2024	Weeks 10-12: Data Collection
11	Thursday, November 14, 2024	Weeks 10-12: Data Collection
11	Sunday, November 17, 2024	**Formative Assignment (R Descriptive) Resubmission Due**
11	Monday, November 18, 2024	**Formative Assignment (R Inferential) Due**
12	Tuesday, November 19, 2024	Weeks 10-12: Data Collection
12	Thursday, November 21, 2024	**Psychonomics Conference: NO CLASS**
12	Friday, November 22, 2024	**Project Milestone #6 (Analyses: Deadline 1) Due**
13	Tuesday, November 26, 2024	**THANKSGIVING BREAK!!! NO CLASS**
13	Thursday, November 28, 2024	**THANKSGIVING BREAK!!! NO CLASS**
14	Monday, December 2, 2024	**Project Milestone #6 (Analyses: Deadline 2) Due**

project checklist

- project checklist

	Pilot 1	Pilot 2	Pilot 3
Which browser were you using?			
Which operating system (Mac / Windows / iPad, etc.)			
Date of piloting			
Were instructions clear? Please note down which instructions had typos / were unclear			
How long did the task take you?			
Was there a consent form?			
Was the demographic survey displayed correctly?			
Did you see the data being displayed at the end of the study?			
What do you think the experiment was about?			
Any other comments?			

CogLab Project Checklist

Task	Check if done
Sanity Check ☐ Is the attention check response being recorded? ☐ Is the free association response being recorded? ☐ Can you differentiate between training / attention / association / prime / target? ☐ Can you differentiate between prime and target trials? ☐ Can you differentiate practice and test trials? ☐ Is subject ID being recorded? ☐ Is RT being recorded?	☐
For the **demographic survey**, how are you showing these questions? Are there multiple answers people can pick or is it a binary choice? Are people able to select multiple answers when they should not be?	☐
For the **demographic survey,** what questions are being shown on the same screen? What questions should be on different screens?	☐
For the **demographic survey**, how are the data being recorded and is it being recorded? Also, do you have all the questions you need?	☐

Before Pre-Registration:
☐ Are you providing accuracy feedback on priming practice trials?
☐ Have you addressed ALL the feedback from Milestone 4?
    ☐ Feedback 1
    ☐ Feedback 2
    ☐ Feedback 3
☐ Are you recording IP addresses?
☐ Are you commenting the condition definition inside cognition.run
☐ Have you piloted your experiment with Uma + other group + 5 friends)
☐ Have they completed the pilot feedback sheet?
☐ Have you sent the cognition.run link by Nov 10?
☐ Have you finalized the analysis plan + sample size?
☐ Have you created and submitted a pre-registration draft?
	☐

Analysis ☐ Did you confirm/correct all datatypes? ☐ Did you figure out how to “filter” certain types of trials? ☐ Did you fix all typos in attention responses? ☐ Have you computed mean attention accuracy? ☐ Have you applied exclusions based on accuracy AND RTs? ☐ Have you created an RT bar graph? ☐ Have you fit a statistical model?	☐

pre-registration + project checklist

- milestone #5: pre-registration + project checklist + piloting (Nov 10)

1. **Data Collection**: Have any data been collected for this study already?
2. **Main Question**: What is the main question being asked or hypothesis being tested in this study?
3. **Dependent Variable(s)**: Describe the key dependent variable(s) specifying how they will be measured.
4. **Condition(s)**: How many and which conditions will participants be assigned to? Please include an example trial of each type of condition you have in your experiment. Please also specify which independent variable will be within-participants or between-participants.
5. **Analyses**: Specify exactly which analyses you will conduct to examine the main question/hypothesis.
6. **Outliers & Exclusions**: Describe exactly how outliers will be defined and handled, and your precise rule(s) for excluding observations.
7. **Predicted Plot**: Please submit a predicted plot for your study based on what you expect the pattern to look like for your main hypothesis.
8. **Sample Size**: How many observations will be collected or what will determine sample size? No need to justify the decision, but be precise about exactly how the number will be determined.
9. **Exploratory details**: Anything else you would like to pre-register? (e.g., secondary analyses, variables collected for exploratory purposes, unusual analyses planned?)

recap

- what we covered:
  - manipulating data using tidyverse verbs
  - project work
- your to-do's were:
  - *work on:* formative assignment #2 (R descriptive)
  - *work on:* project checklist + pre-registration

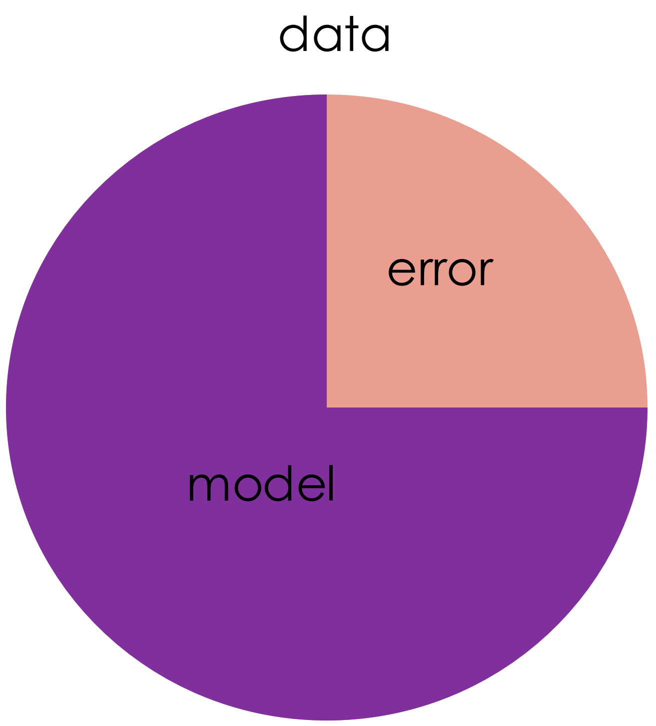
today’s agenda

- making statistical inferences from data

what is the goal of statistics?

data = model + error

- the goal of statistics is to find a simple explanation to the observed data (Y, dependent variable), i.e., build a model of the data that approximates/explains it as well as possible –
- what is a *good* model? one that represents the data really well
- how do we start building models?



some simple models

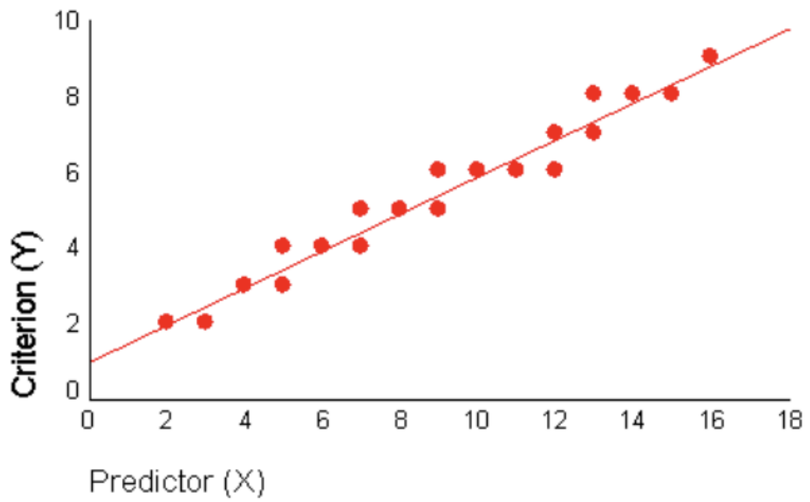
- central tendencies (mean/median/mode)
  - derived from the key dependent variable (data = Y) itself
  - no other variables needed for this
  - the mean is the <u>best</u> model if no other variables are available
- measures of variation: estimates of model fit
  - sum of squared errors (SSE or SS): $\sum_{i=1}^{N}(Y_i-\mu)^2$
  - mean of squared errors (MSE): $\frac{\sum_{i=1}^{N}(Y_i-\mu)^2}{N} = \frac{SS}{N}$
  - root mean squared error (RMSE): $\sqrt[2]{\frac{\sum_{i=1}^{N}(Y_i-\mu)^2}{N}} = \sqrt{MSE}$

slightly more complex models…

- using one more variable (X) to “explain” the dependent variable/data (Y)
- how does knowing something about X impact what we know about Y?
- what types of “models” are these?

linear regression

- a linear regression (or a linear model) is a model that fits a line to the data
- Y = a + bX + error
- slope: $b = r\frac{s_y}{s_x}$
- intercept: $a = M_y - bM_x$



exploring the data

- open your project + Rmd
- new heading # linear models
- load the dataset women
- make a scatterplot of the data
    - x = weight
    - y = height
- fit a line to the data via geom_smooth()

first-R-notebook.Rmd × | women ×

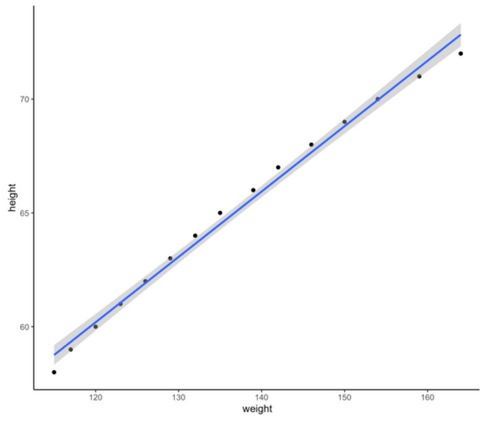
	height	weight
1	58	115
2	59	117
3	60	120
4	61	123
5	62	126
6	63	129
7	64	132
8	65	135
9	66	139
10	67	142
11	68	146
12	69	150
13	70	154
14	71	159
15	72	164

```
# linear models
```

````
```{r}

data(women)
````

```
women %>%
  ggplot(aes(x= weight, y = height))+
  geom_point() +
  geom_smooth(method = "lm")+
  theme_classic()
```


```

linear regression in R

- predict height by weight
- print the summary of the model
- what is the equation of the line?

```
women_model = lm(data = women, height ~ weight)
```

```
summary(women_model)
```

```
Call:
lm(formula = height ~ weight, data = women)

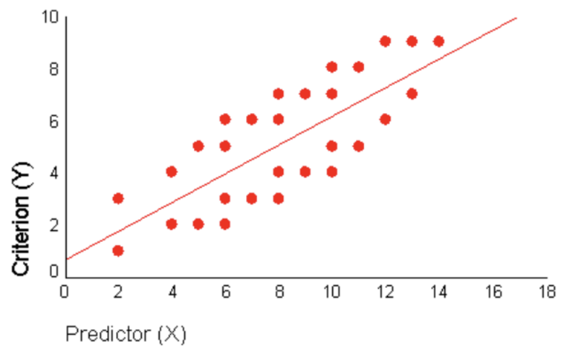
Residuals:
     Min       1Q   Median       3Q      Max 
-0.83233 -0.26249  0.08314  0.34353  0.49790 

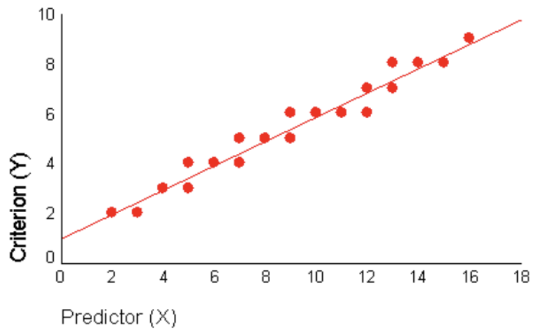
Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 25.723456   1.043746   24.64 2.68e-12 ***
weight       0.287249   0.007588   37.85 1.09e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.44 on 13 degrees of freedom
Multiple R-squared:  0.991,	Adjusted R-squared:  0.9903 
F-statistic:  1433 on 1 and 13 DF,  p-value: 1.091e-14
```

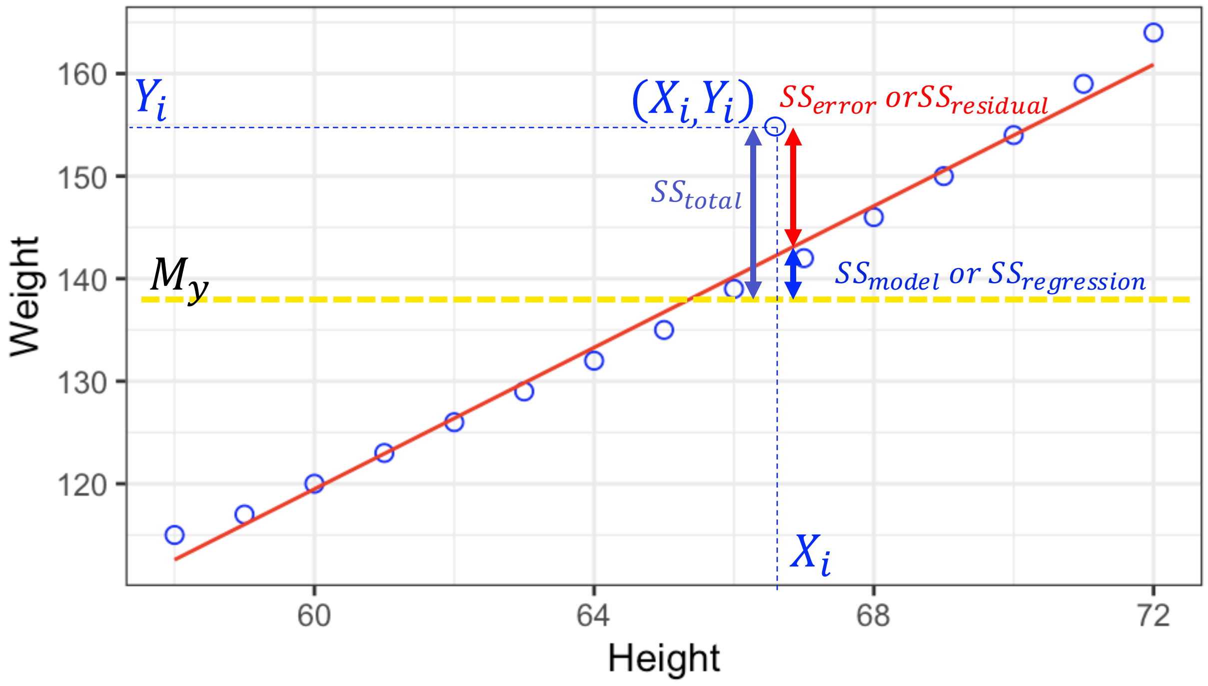
assessing model fit

- let's say we find a line of best fit
  - data = model + error
  - Y = $a + bX$ + error
  - $\hat{Y} = a + bX$ = predictions
  - Y = $\hat{Y}$ + error
- how well does the line fit our data?
- $SS_{error} = \sum_{i=1}^{n}(y_i - a - bx_i)^2 = \sum(Y - \hat{Y})^2$





understanding goodness/errors



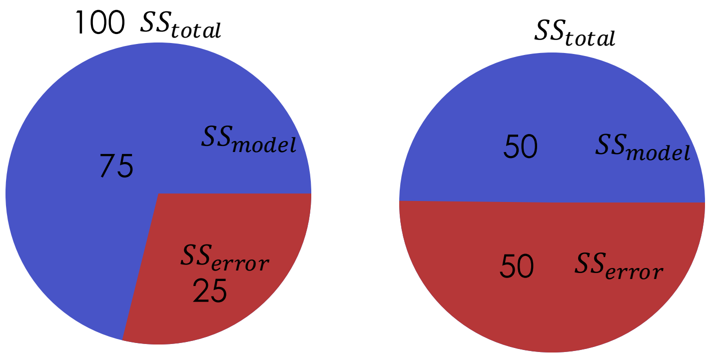
$$SS_{total} = SS_{model} + SS_{error}$$

$$SS_{total} = \sum (Y - M_y)^2$$

$$SS_{error} = \sum (Y - \hat{Y})^2$$

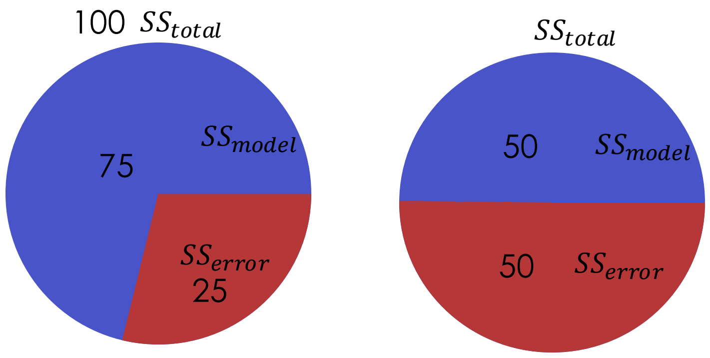
$$SS_{model} = \sum (\hat{Y} - M_y)^2$$

overall test of model (ANOVA)



- **an**alysis **o**f **va**riance assesses the overall fit of the model
- $SS_{total} = SS_{model} + SS_{error}$
- we calculate the ratio between the variance explained by the model and the natural variance expected/left over in the dependent variable
  - if $\frac{SS_{model}}{SS_{error}}$ is high, the model explains **more** variance than expected
  - if $\frac{SS_{model}}{SS_{error}}$ is low, the model explains **less** variance than expected
- typically, we want the "average" variance explained, so we also divide this by *df*

F ratio



- The F ratio compares the "average" squared error between model (explained variance) and the natural (unexplained) variance (data = model + error)

$$F = \frac{explained\ variance}{unexplained\ variance} = \frac{MS_{model}}{MS_{error}} = \frac{SS_{model}/df_{model}}{SS_{error}/df_{error}}$$

- obtaining SS_{model} and SS_{error}
  - $SS_{error} = \sum(Y - \hat{Y})^2$ and $SS_{total} = \sum(Y - M_y)^2$
  - $SS_{model} = SS_{total} - SS_{error}$
- obtaining df_{model} and df_{error}
  - k denotes the number of levels of the independent variable OR number of estimated parameters
  - $df_{model} = k - 1$
  - $df_{error} = n - k$

ANOVA for women dataset

- install the car package
- use the Anova() function
- how do we report this F test?
- weight significantly predicted height, $F(1,13) = 1433$, $p < .001$.

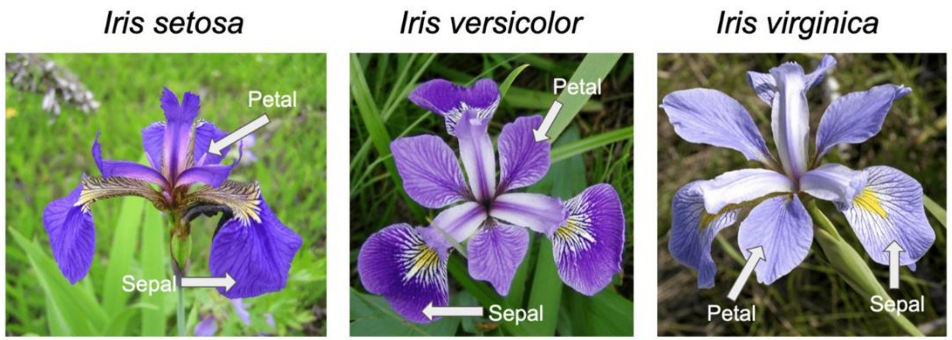
```
car::Anova(women_model)
```

```
> car::Anova(women_model)
Anova Table (Type II tests)

Response: height
           Sum Sq Df F value    Pr(>F)
weight    277.483  1    1433 1.091e-14 ***
Residuals   2.517 13
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

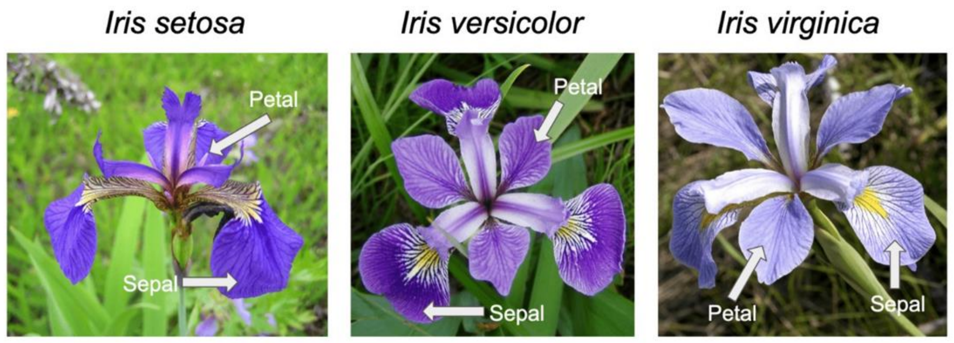
ANOVAs for categorical IVs

- the same logic applies to problems where the independent variable is not continuous
- research question: what explains the variation in petal lengths (Y)?
  - data (Y) = model (X)+ error
  - petal lengths (Y) = species (X) + error



descriptive exercise

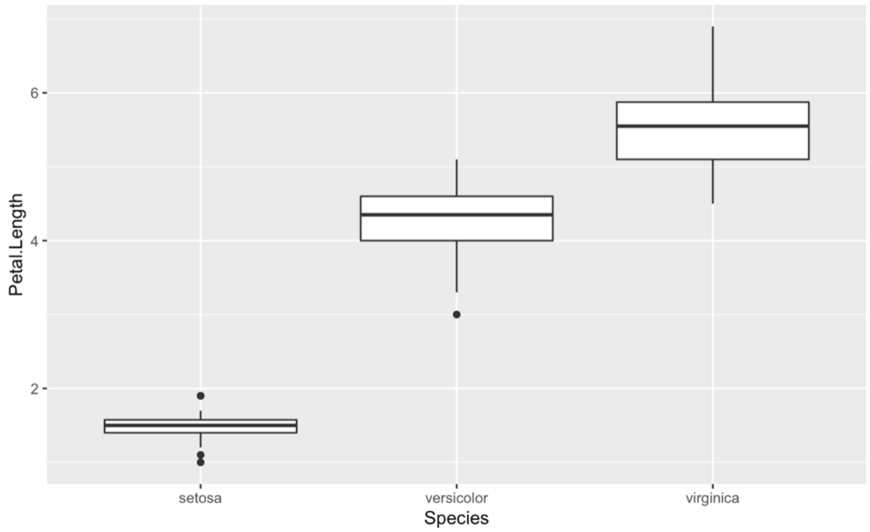
- obtain the mean petal length for each species in the iris dataset using tidyverse functions




```
# A tibble: 3 × 2
  Species    mean_length
  <fct>            <dbl>
1 setosa            1.46
2 versicolor        4.26
3 virginica         5.55
```

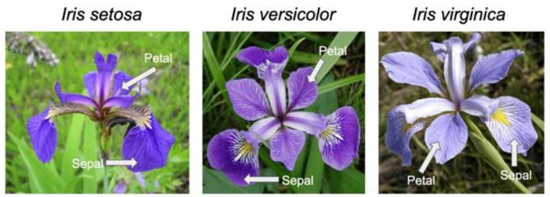
plotting exercise

- make a boxplot of petal lengths by species



ANOVA for iris

- load and view iris
- fit a model to petal lengths
- view car::Anova() results
- does species explain the variation in petal lengths?
- which species are different from each other?




```
## anova

data(iris)
View(iris)
```

```
iris_model = lm(data=iris, Petal.Length ~ Species)
```

```
car::Anova(iris_model)
```

```
> car::Anova(iris_model)
Anova Table (Type II tests)

Response: Petal.Length
          Sum Sq  Df F value    Pr(>F)
Species   437.10   2  1180.2 < 2.2e-16 ***
Residuals  27.22 147
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

follow-up tests

- when more than two groups are present, it can be useful to understand exactly which groups differ from each other
- install emmeans package
- load the package inline and compute pairwise differences

```
emmeans::emmeans(iris_model,
                 pairwise ~ Species,
                 adjust = "tukey")
```

```
$emmeans
 Species    emmean     SE  df lower.CL upper.CL
 setosa       1.46 0.0609 147     1.34     1.58
 versicolor   4.26 0.0609 147     4.14     4.38
 virginica    5.55 0.0609 147     5.43     5.67

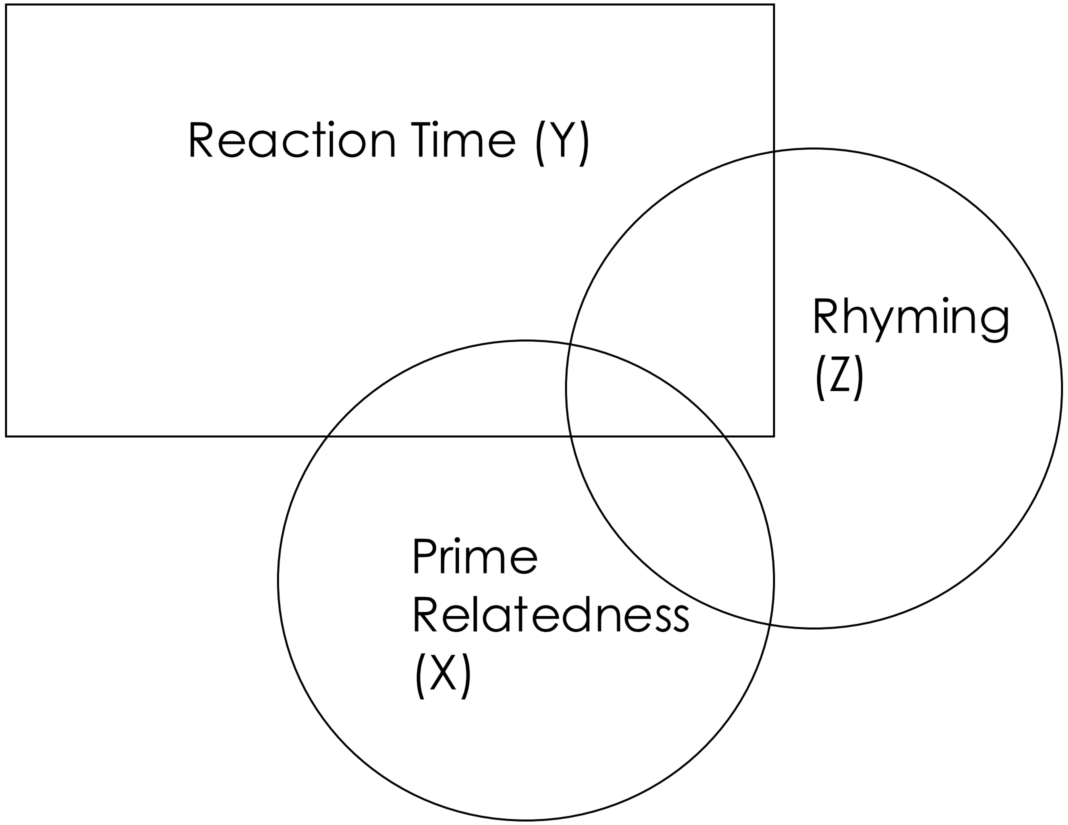
Confidence level used: 0.95

$contrasts
 contrast               estimate     SE  df t.ratio p.value
 setosa - versicolor       -2.80 0.0861 147 -32.510  <.0001
 setosa - virginica        -4.09 0.0861 147 -47.521  <.0001
 versicolor - virginica    -1.29 0.0861 147 -15.012  <.0001

P value adjustment: tukey method for comparing a family of 3 estimates
```

even more complex models…

- what if the variation in our data (Y) could be explained further?
- data = model + error
  - *one IV*: $Y = a + bX + \text{error}$
  - *multiple IVs*: $Y = a + b_1X_1 + b_2X_2 + \ldots + \text{error}$
- central idea remains the same, but more complex relationships are possible



next class

- **before** class
  - *apply:* formative assignment #2 (due Monday)
  - *apply:* pre-registration + checklist (due Nov 10)
- **during** class
  - complex models + project work