DATA ANALYSIS

Week 10: Modeling Relationships

logistics: PS4

- no chapter 12 problems yet (moved to PS5)
- Instead, you have two additional problems
  - mtcars
  - creativity and intelligence
- see doc + template

Additional problem (mtcars):

You will use the “mtcars” dataset (data available in worksheet template) openly available in R. The data was extracted from the 1974 Motor Trend US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973–74 models). We will focus on two key variables, miles per gallon (mpg) and horsepower (hp).

Additional problem (creativity & intelligence):

- **Additional problem (creativity and intelligence):** The following table is from a study conducted by Benedek and colleagues (2020, link here) and looks at the relationship between creativity and intelligence. They find a correlation between **creativity** (as measured by the Bi-Association task where two adjectives were presented and participants should find a concept that is semantically related to both cues and links them in an original way (e.g., red - round: clown nose) and **fluid intelligence** (Gf, as measured by performance on a matrix pattern task). The reported correlation coefficient is 0.36 with a sample size of 102 participants.

- Verbally describe each of the following and calculate statistics wherever required:
  - key variable(s) & research question
  - sample and population
  - sample statistic
  - null & alternative hypothesis
  - sampling distribution
  - critical region, test statistic, p-value
  - statistical significance
  - type I and type II error
  - power
  - effect size

Table 1
Descriptive statistics and correlations of all measures.

	M	*SD*	1	2	3	4	5	6	7	8	9
1 Com-Assoc	1.07	0.11	–								
2 Orig-Assoc	1.20	0.21	**.33**	–							
3 Bi-Assoc	1.62	0.18	.17	**.44**	–						
4 DT Creativity	1.61	0.32	.15	**.38**	**.25**	–					
5 DT Fluency	8.03	2.55	.12	**.30**	**.22**	**.37**	–				
6 C-Activity	1.42	0.52	.05	.17	**.20**	**.25**	.17	–			
7 Openness	2.92	0.51	.12	**.32**	**.33**	**.22**	.13	**.40**	–		
8 Gr	13.26	2.04	.17	**.42**	**.29**	**.39**	**.51**	.18	**.25**	–	
9 Gf	11.54	2.78	.07	**.31**	**.36**	.07	.06	.07	.16	.16	–
10 W-Speed	12.41	1.75	.10	-.08	-.06	.01	.04	-.01	**.19**	.09	.09

Notes. Com-Assoc = common association, Orig-Assoc = original association, Bi-Assoc = bi-association, DT = Divergent thinking, C-Activity = Creative activities, Gr = Broad retrieval ability, Gf = fluid intelligence, W-Speed = Writing speed. For $n = 102$, correlations of $r \geq 0.19$ are significant at $p < .05$, correlations of $r \geq 0.25$ are significant at $p < .01$, and correlations of $r \geq 0.31$ are significant at $p < .001$. Significant correlations ($p < .05$) are indicated in bold.

where are we going next?

data = model + error

thus far

- data = mean + error
- data = X (interval/ratio) + error

after break

- data = X (interval/ratio/nominal) + error
- data = X + Y + error
- data (NOIR) = model (NOIR) + error

today’s agenda

- hypothesis testing for regression
- hypothesis testing for nominal variable

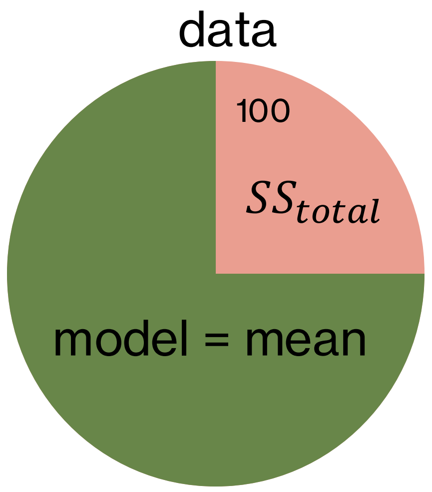
data come in all forms

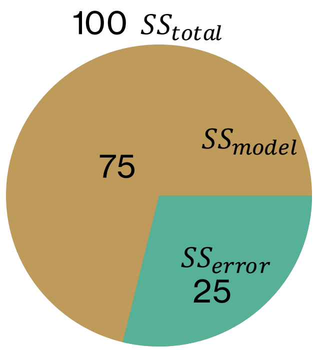
- think back to scales of measurement (NOIR): what kinds of data have we worked with so far?

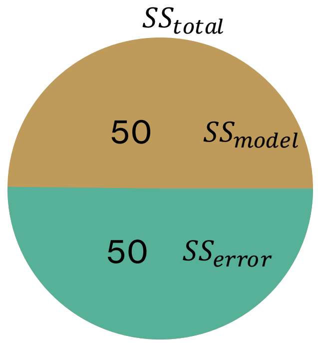
	independent variable (X)		
dependent variable (Y)	nominal	ordinal	interval/ ratio
nominal			
ordinal			
interval/ratio			r or b

linear regression: model fit in samples

- coefficient of determination (R^2): percentage of variance explained in Y due to X
    - $R^2 = \frac{SS_{model}}{SS_{total}}$
- standard error: “average” error left over in Y
    - standard error of estimate: $SE_{model} = \sqrt{\frac{SS_{error}}{df}} = \sqrt{\frac{SS_{error}}{n-2}}$
    - standard error of correlation: $SE_r = s_r = \sqrt{\frac{1-r^2}{n-2}}$

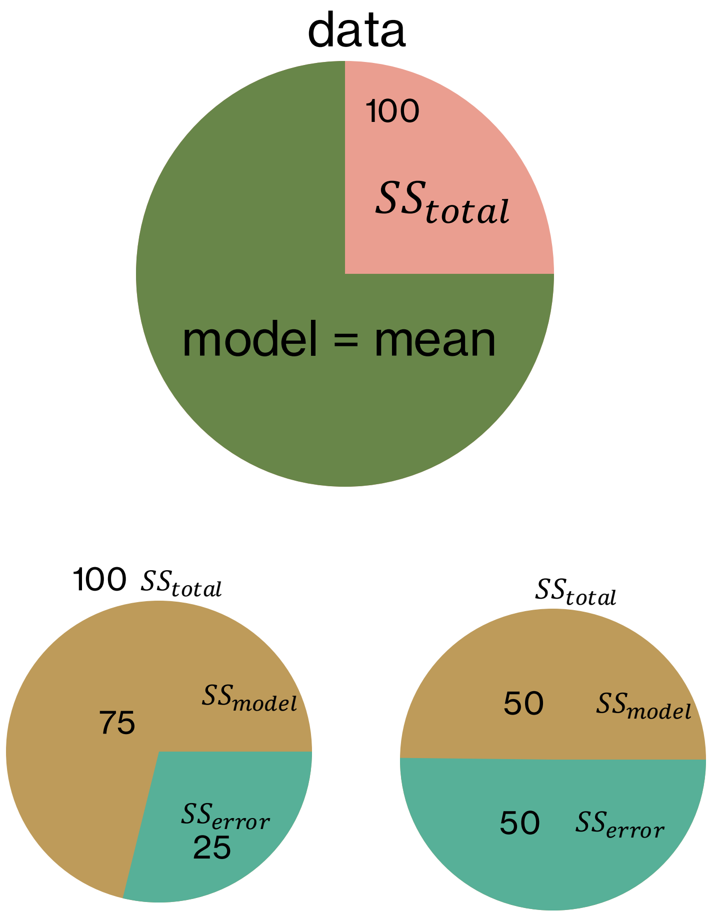


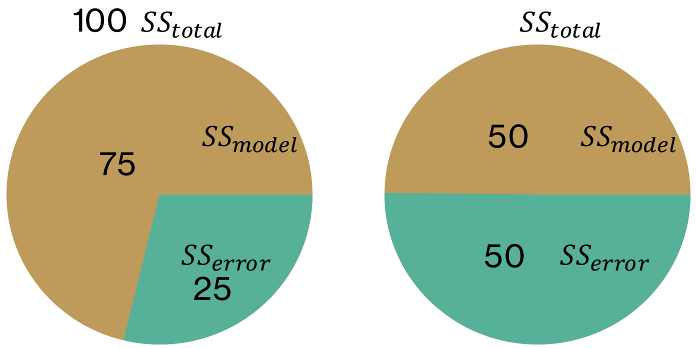




hypothesis test for populations: ANOVA

- an **an**alysis **o**f **va**riance (ANOVA) tests whether a variable explains significantly more variance in another variable than chance
  - $SS_{total} = SS_{model} + SS_{error}$
- we can calculate the ratio between the variance explained by the model and the variance expected/left over
  - 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 both errors by *degrees of freedom* *(see end of slide deck)*





F ratio

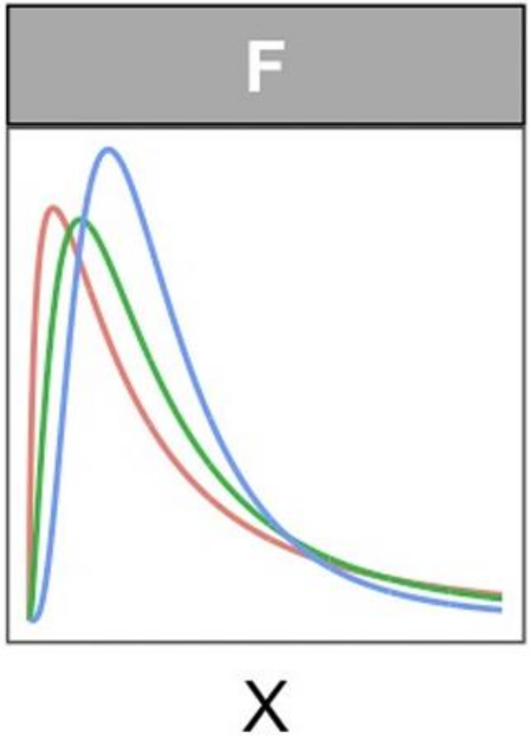
- The F ratio compares the “average” squared error between model (explained variance) and the natural (unexplained) variance (data = model + error)

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

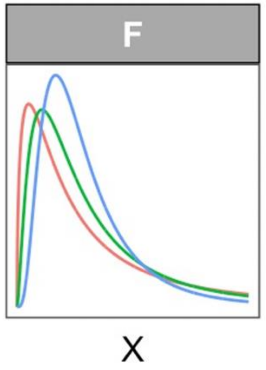
- obtaining SS_{model} and SS_{error}
  - $SS_{error} = \sum(Y - \hat{Y})^2$ and $SS_{total} = \sum(Y - M_y)^2$
  - $SS_{model} = SS_{total} - SS_{error} = \sum(\hat{Y} - M_y)^2$
- obtaining df_{model} and df_{error}
  - k denotes the number of levels of the independent variable OR number of estimated parameters
  - $df_{model} = k - 1$ (also called df_1 or $df_{numerator}$)
  - $df_{error} = n - k$ (also called df_2 or $df_{denominator}$)

interpreting F values

- The F-distribution is a positively skewed distribution
- defined by two parameters (df_1 and df_2) that determine the exact form/shape
- F-values are typically **non-negative**: why??
    - $F = \frac{MS_{model}}{MS_{error}} = \frac{SS_{model}/df_{model}}{SS_{error}/df_{error}}$
    - $F = 1$: $MS_{model} = MS_{error}$ i.e., the model does not do any better than random chance
    - $F > 1$: more variance explained by model than random chance
- F ratios enable us to generalize our models to the population (in contrast to R^2 and standard error)



NHST for linear regression (F-test)



step 1: state the hypotheses

$H_0: \beta = 0$
$H_1: \beta \neq 0$

start with assuming that the "slope" is 0, i.e., there is no relationship between the two variables

step 2: set criteria for decision

$\alpha = .05$

find $F_{critical}$ based on **right** tailed test and degrees of freedom

$df_1 = k - 1$
$df_2 = n - k$

step 3: collect data

(1) compute SS_{model} and SS_{error}

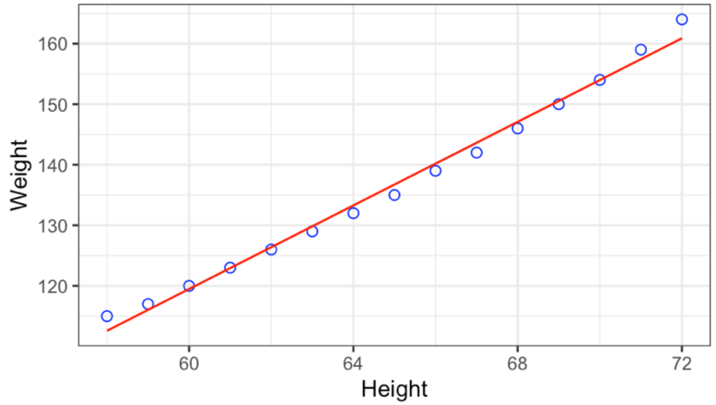
(2) compute $F_{observed} = \frac{MS_{model}}{MS_{error}}$

(3) find p-value for F-score

step 4: make a decision!

check whether $F_{observed}$ is beyond $F_{critical}$ and p-value < .05. if so, reject null hypothesis!

F-test for women dataset

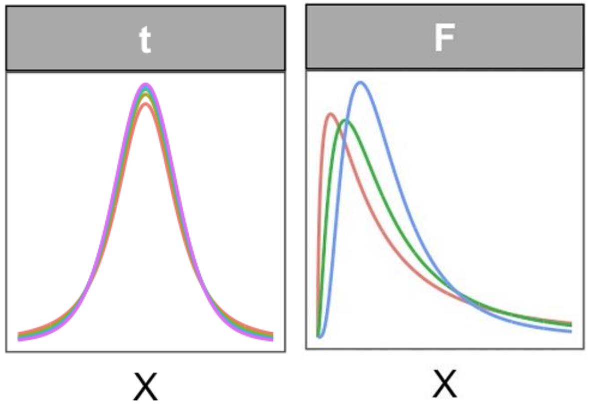


- **step 1: state the hypotheses**
  - H$_0$: β = 0, height explains no variance in weights for women
  - H$_1$: $\beta \neq$ 0, height explains some variance in weights for women
- **step 2: set criteria for decision**
  - $\alpha = .05, k = 2, n = 15$
  - $F_{critical}$

  $= F_{critical}(k-1, n-k) = F\ (1,13) = 4.667$
- **step 3: collect data**
  - SS_{error} = 30.23 and SS_{total} = 3362.93
  - thus, SS_{model} = $SS_{total} - SS_{error}$ = 3332.7
  - compute the F-statistic:
  - $F_{observed} = \frac{MS_{model}}{MS_{error}} = \frac{SS_{model}/df_{model}}{SS_{error}/df_{error}} = \frac{3332.7/1}{30.23/13} = 1433$
  - compute p-value: $p_{observed} < .0001$
- **step 4: decide!**
  - Height explains significantly more variance in weights than expected by chance, $b = 3.45$, $F\ (1,13) = 1433, p < .0001$.

[Sheets solution](#)
[video tutorial](#)

t and F relationship



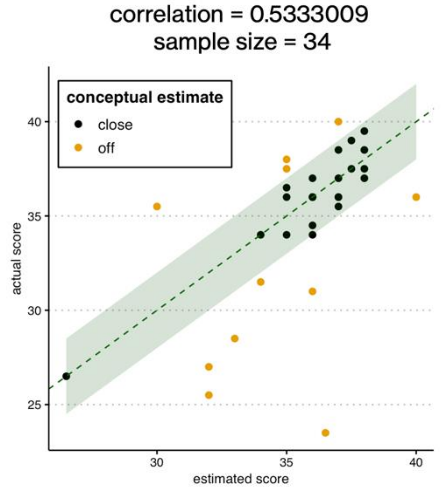
- regression test for women dataset
  - $F\,(1, 13) = 1433, p < .0001$
- conduct a correlation test for women dataset ($r = .995$, n = 15)
  - $r = .995, t(13) = 37.86, p < .001$
- what is t^2 ?
- **for the same data, $t^2 = F$!!**
- t tests are in original units of the sample statistic, F tests are in squared error units
- both tests have the same general conceptual form (observed / expected)

F-tables

- F-tests are typically represented in tables

		SS	df	MS	F	p-value
SS_{model}	regression	3332.7	1	3332.7	1433.02	<.0001
SS_{error}	residual	30.23	13	2.33		
SS_{total}	total	3362.93	14			

- knowing parts of the F table are sufficient for completing it!



review: conceptual exam t-test

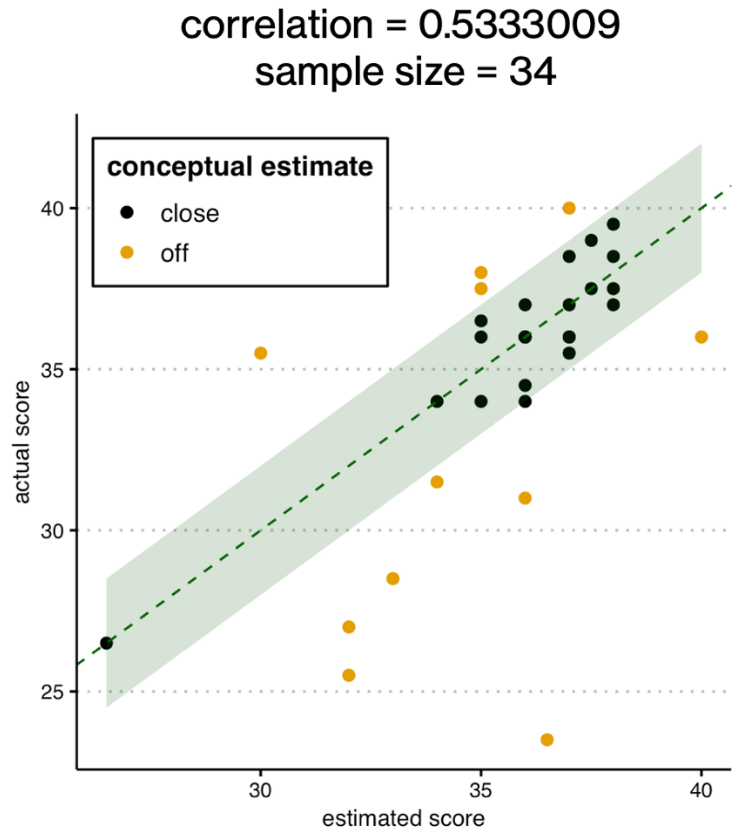
- **step 1: state the hypotheses**
  - H_0: $\rho = 0$, no correlation between estimate and actual score on conceptual exam
  - H_1: $\rho \neq 0$, a correlation between estimate and actual score on conceptual exam
- **step 2: set criteria for decision**
  - $t_{n-2} = t_{32} = t_{critical} = \pm 2.0369 \; at \; \alpha = .05$
- **step 3: collect data**
  - correlation r = 0.5333009
  - compute the standard error for correlation

$$SE_r = s_r = \sqrt{\frac{1 - r^2}{n - 2}} = 0.1496$$

  - compute the t-statistic:
  - $t_{observed} = \frac{r-0}{SE_r} = \frac{0.5333009}{0.1496} = 3.57$
  - compute p-value: $p_{observed} = .001$
- **step 4: decide!**
  - estimates significantly correlate with actual scores on the conceptual exam, r = .53, t(32) = 3.57, p = .001

W10 Activity 2: conduct F test

- data



```
> stats_model = lm(data = data_analysis,
+                  m1c_actual ~ m1c_estimate)
> car::Anova(stats_model)
Anova Table (Type II tests)

Response: m1c_actual
             Sum Sq Df F value   Pr(>F)
m1c_estimate 163.78  1  12.718 0.001164 **
Residuals    412.08 32
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

creating F-table

- F-tests are typically represented in tables

		SS	df	MS	F	p-value
SS_{model}	regression					
SS_{error}	residual					
SS_{total}	total					

creating F-table

```
> stats_model = lm(data = data_analysis,
+                  m1c_actual ~ m1c_estimate)
> car::Anova(stats_model)
Anova Table (Type II tests)

Response: m1c_actual
             Sum Sq Df F value   Pr(>F)
m1c_estimate 163.78  1  12.718 0.001164 **
Residuals    412.08 32
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- F-tests are typically represented in tables

		SS	df	MS	F	p-value
SS_{model}	regression	163.78	1	163.78	12.718	.001
SS_{error}	residual	412.08	32	12.877		
SS_{total}	total	575.86	33			

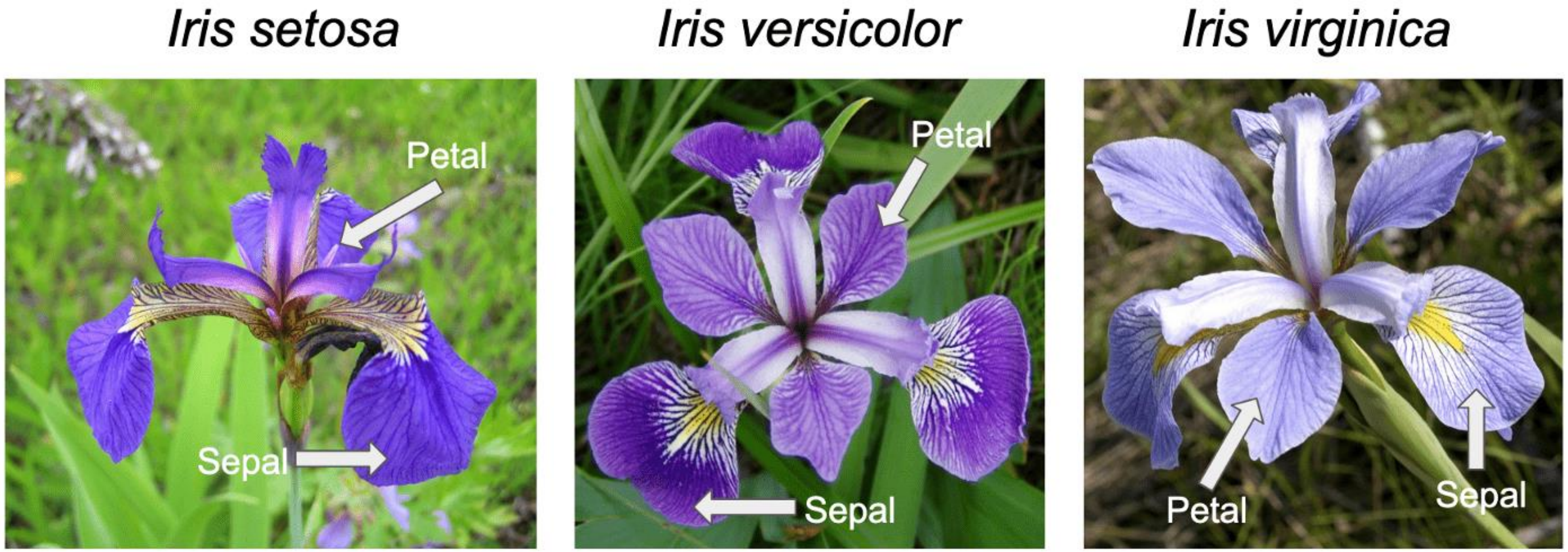
data come in all forms

- think back to NOIR: what kinds of data have we worked with so far?
- when data are **not interval/ratio**, the same general framework of can be applied, with a few modifications

	independent variable		
dependent variable	nominal	ordinal	interval/ ratio
nominal			
ordinal			
interval/ratio			r or b

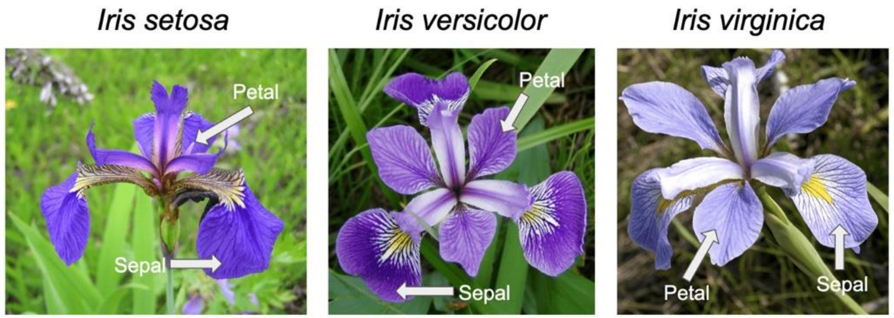
example: iris dataset

- the iris dataset contains petal and sepal dimensions for three species (setosa, virginica, and versicolor)



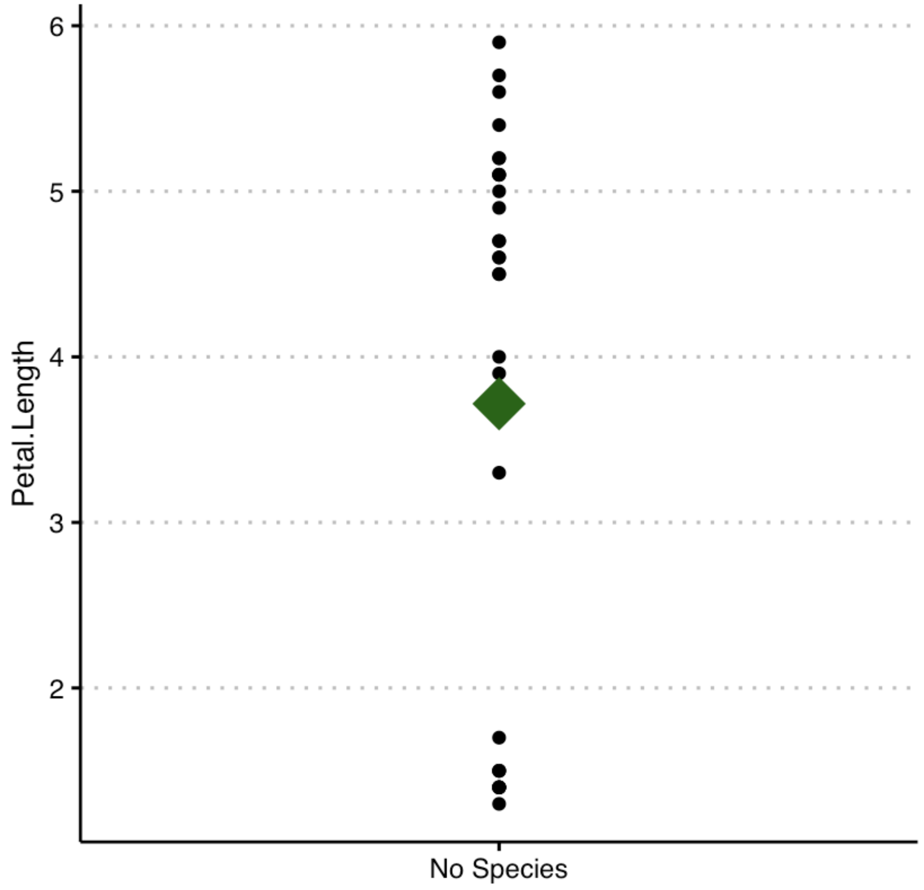
example: iris dataset

- our goal is to build the best model for petal lengths
- if there were no species labels in this dataset, what would be the best model of petal lengths?



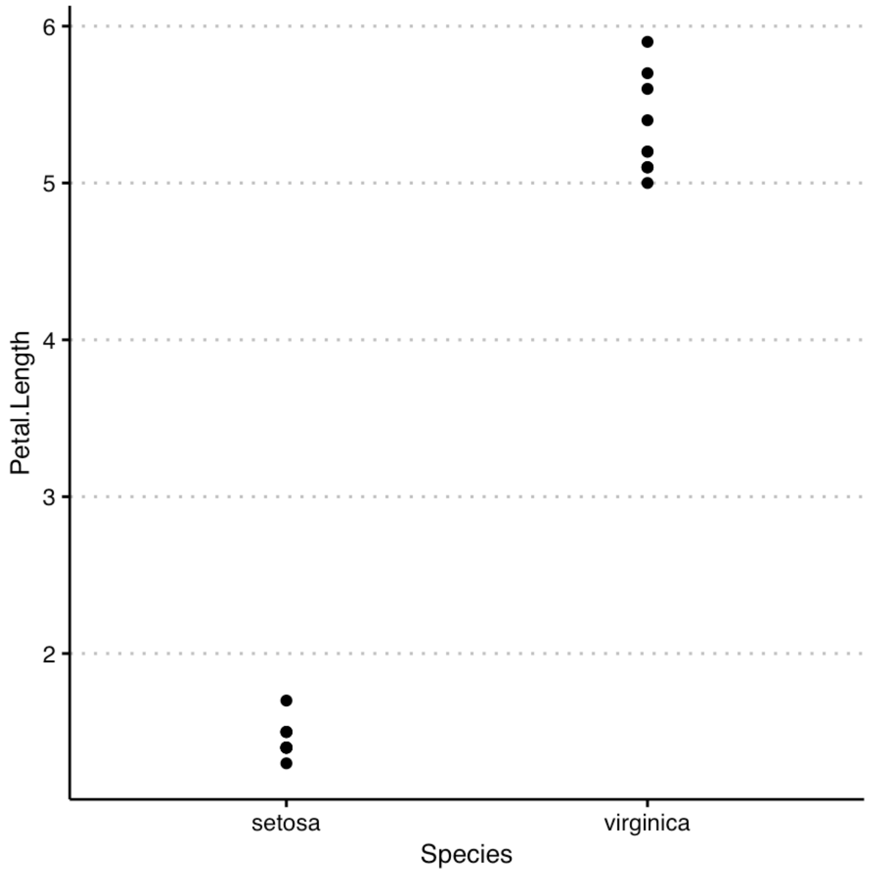
example: iris dataset

- if there were no species labels in this dataset, the overall or “grand mean” of all petal lengths would be the best model for the data
- this “grand mean” will provide our baseline, i.e., how much better can we do than the grand mean in fitting a model to the data?



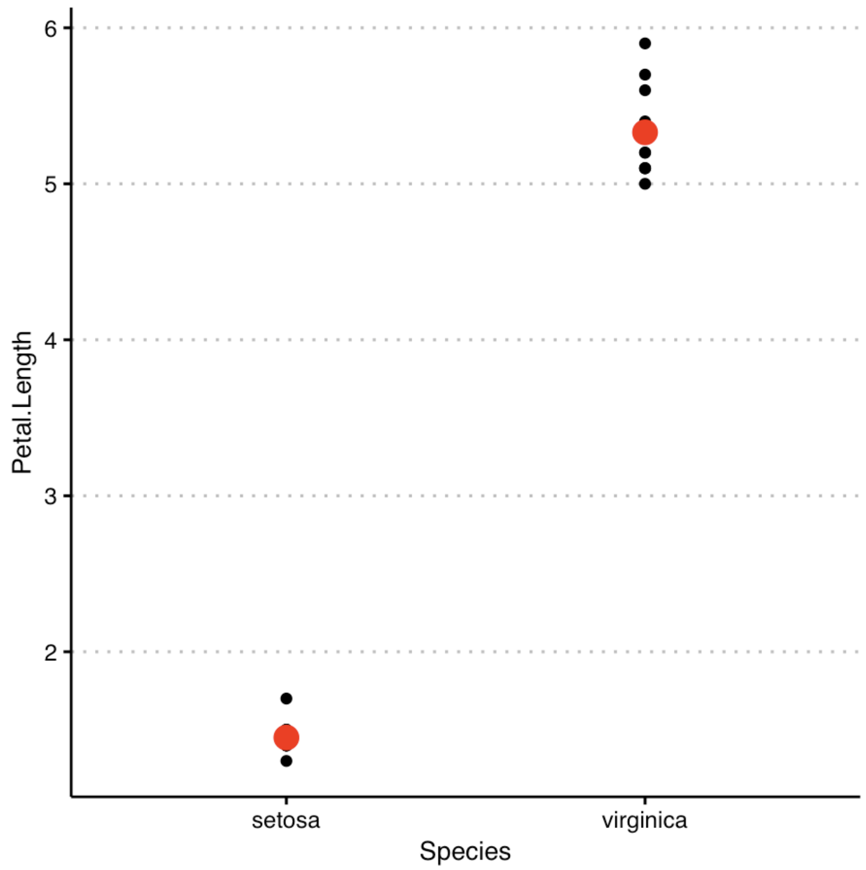
comparing two groups

- our goal is to fit a different model to the data that includes species as additional information and evaluate how much better we can do than the grand mean
- Y (petal lengths) = X (species) + error
- instead of a continuous scale of values, X (species) can only take two values: setosa and virginica
- how can we build a model using species information?



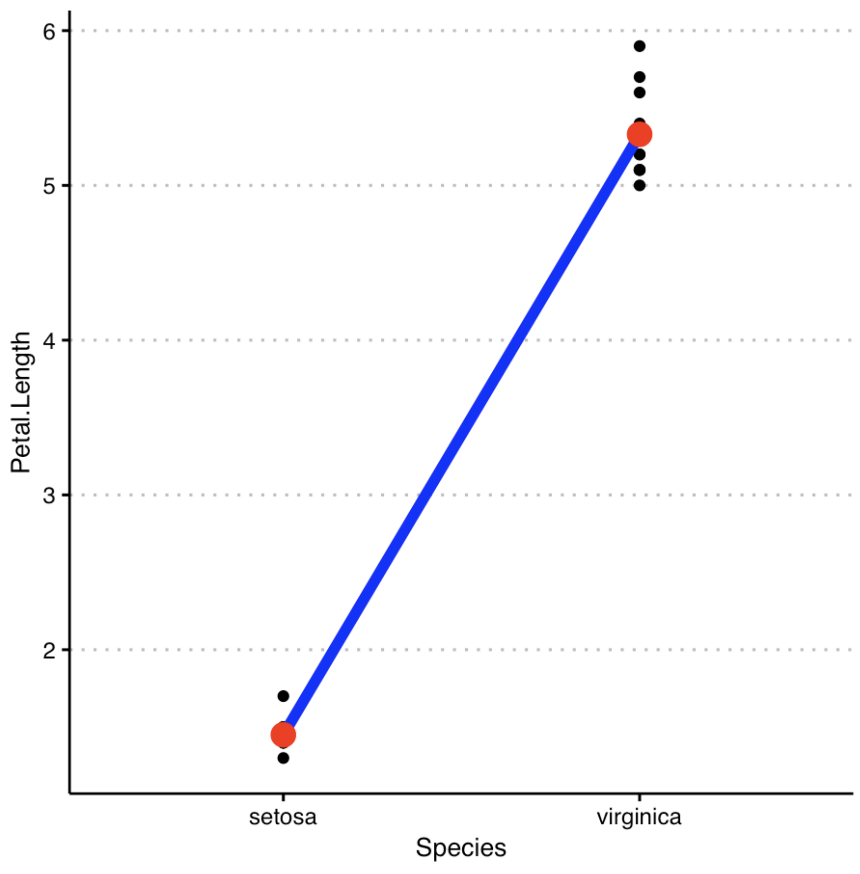
comparing two groups

- we could take the mean of each group



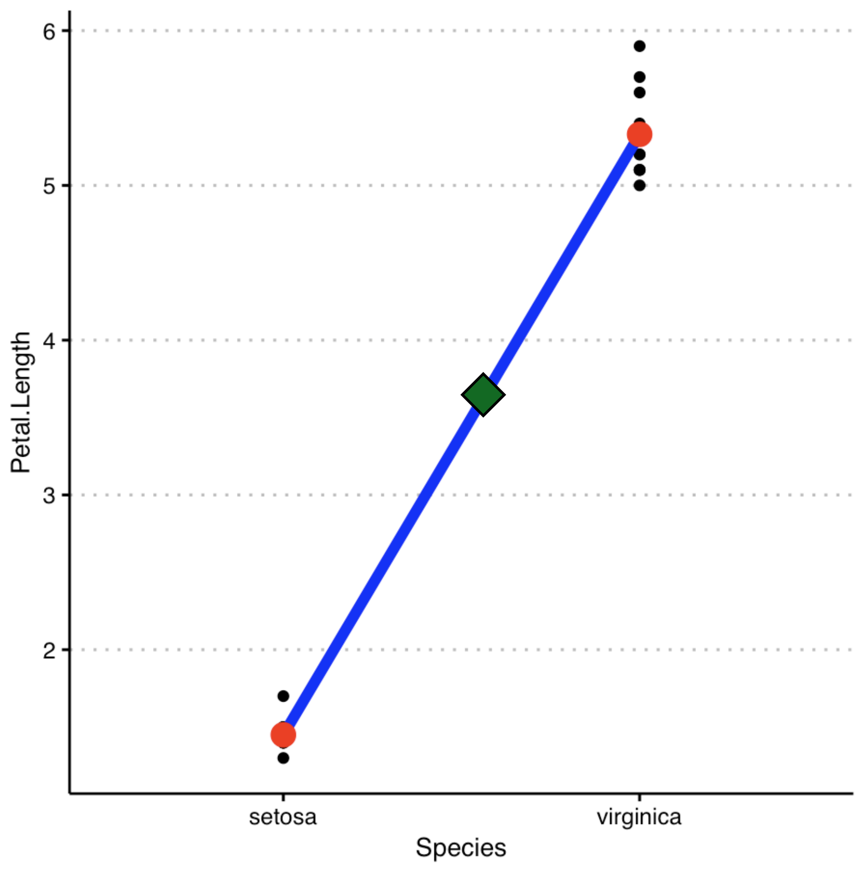
comparing two groups

- essentially, we are “fitting” a model to the data that substitutes the mean of the individual species instead of the grand mean
- the model is the species means instead of the grand mean



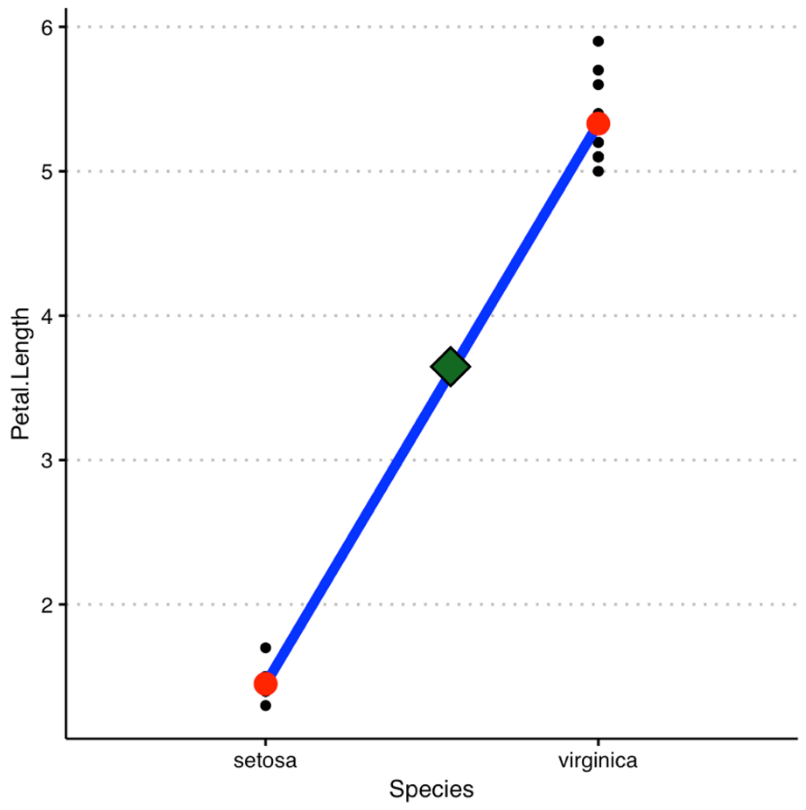
comparing two groups

- essentially, we are “fitting” a model to the data that substitutes the mean of the individual species instead of the grand mean
- the model is the species means instead of the grand mean
- data = model + error
- Y = $\hat{Y}$ + error
- $\hat{Y}$ = the mean of the group to which the data point belongs



F-test for two groups

- just as we did an “overall” test for linear regression, we can do the same here for the iris dataset, where we compare the grand mean model with the species mean model
- recall that $F = \frac{MS_{model}}{MS_{error}} = \frac{SS_{model}/df_{model}}{SS_{error}/df_{error}}$
- and SS_{total} = SS_{model} + SS_{error}
- how do we obtain SS_{total}, SS_{error}, and SS_{model}?



F-test for two groups

- SS_{total} represents score deviations from grand mean (M_Y)

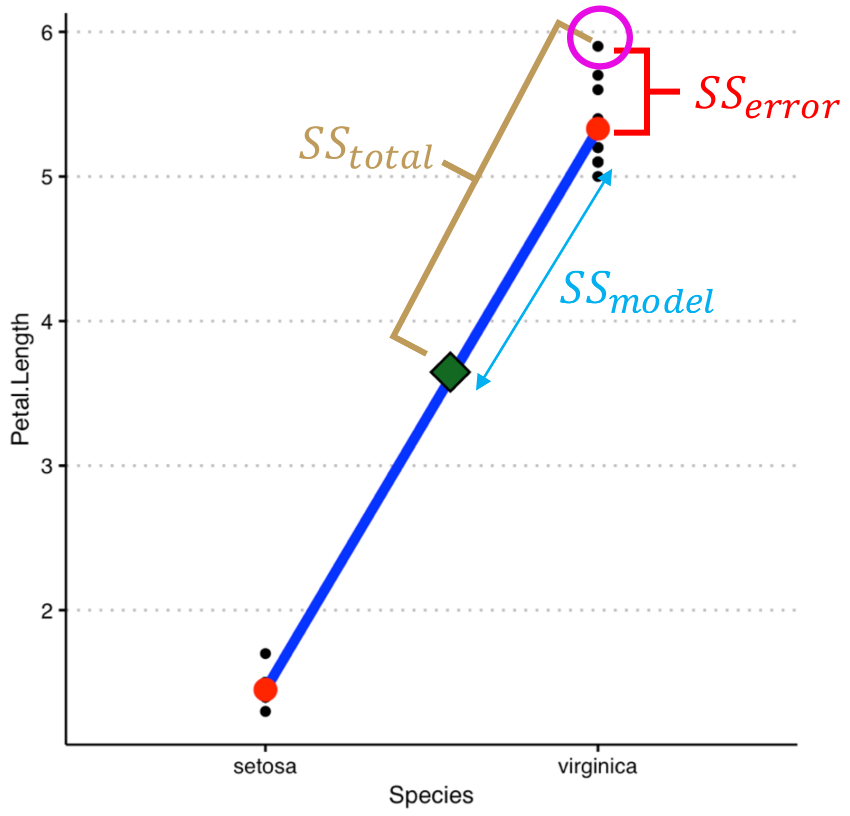
$$SS_{total} = \sum (Y - M_Y)^2$$

- SS_{error} represents the deviations of each score from its group mean

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

- SS_{model} represents the ***gains*** we get if we substitute each score with the group mean instead of the grand mean

$$SS_{model} = \sum \sum n_i \, (M_{group} - M_Y)^2 = SS_{total} - SS_{error}$$



NHST for two independent groups (F-test)

step 1: state the hypotheses

step 2: set criteria for decision

step 3: collect data

step 4: make a decision!

$H_0: \mu_2 - \mu_1 = 0$
$H_1: \mu_2 - \mu_1 \neq 0$

start by assuming that group information is not useful, i.e., the two groups do not have differences

$\alpha = .05$
find $F_{critical}$ based on **right** tailed test and degrees of freedom
$df_1 = k - 1$
$df_2 = n - k$

(1) compute SS_{model} and SS_{error}

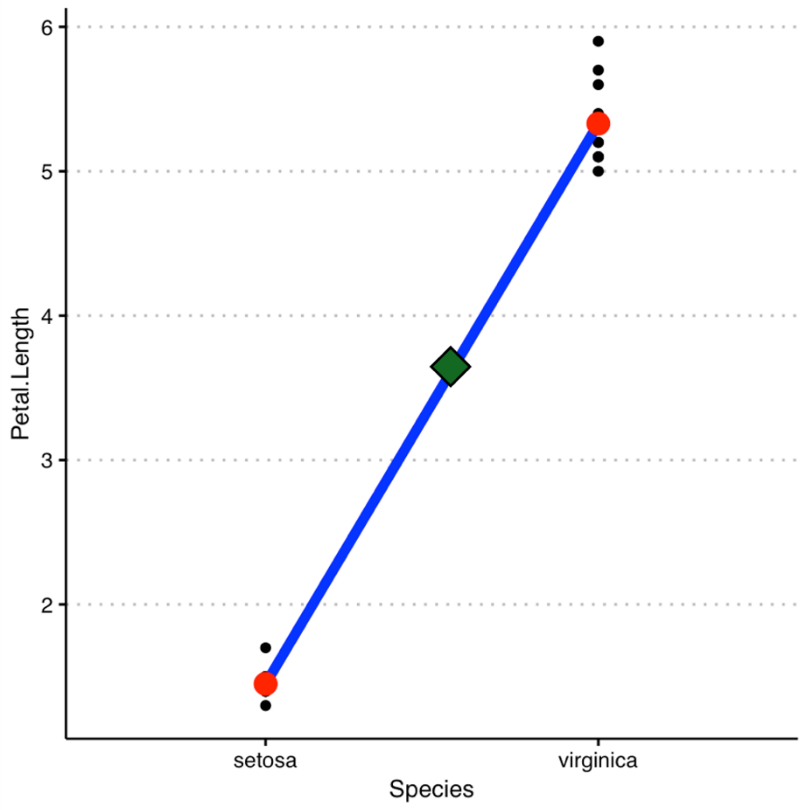
(2) compute $F_{observed} = \frac{MS_{model}}{MS_{error}}$

(3) find p-value for F-score

check whether F is beyond $F_{critical}$ and p-value < .05. if so, reject null hypothesis!

activity: F-test for iris dataset

- conduct the F test for the [iris dataset]
- **step 1: state the hypotheses**
- **step 2: set criteria for decision**
  - find $F_{critical}$
- **step 3: collect data**
  - compute SS_{total}, SS_{model} and SS_{error}
  - compute $F_{observed} = \frac{MS_{model}}{MS_{error}}$
  - find p-value for F-score
- **step 4: decide**



F-test for iris dataset

- **step 1: state the hypotheses**
  - H_0: $\mu_{virginica} - \mu_{setosa} = 0$: petal lengths for both species are equal
  - H_1: $\mu_{virginica} - \mu_{setosa} \neq 0$: petal lengths for species are different
- **step 2: set criteria for decision**

$k = 2$: number of levels of independent variable OR estimated parameters

$$df_1 = k - 1 = 2 - 1 = 1$$

$$df_2 = n - k = 20 - 2 = 18$$

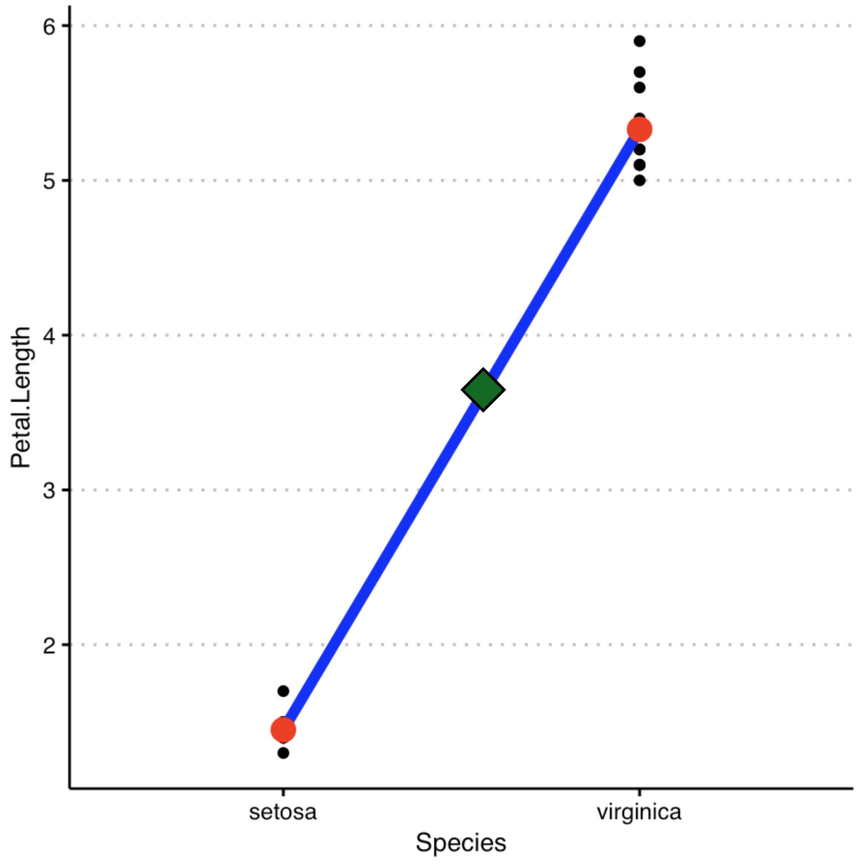
$$F(df_1, df_2) = F(1,18) = F_{critical} = 4.414$$

step 3a: obtaining SS_{total}

- what is SS_{total}? SS_{total} is the error left over after the grand mean has been fit to the data

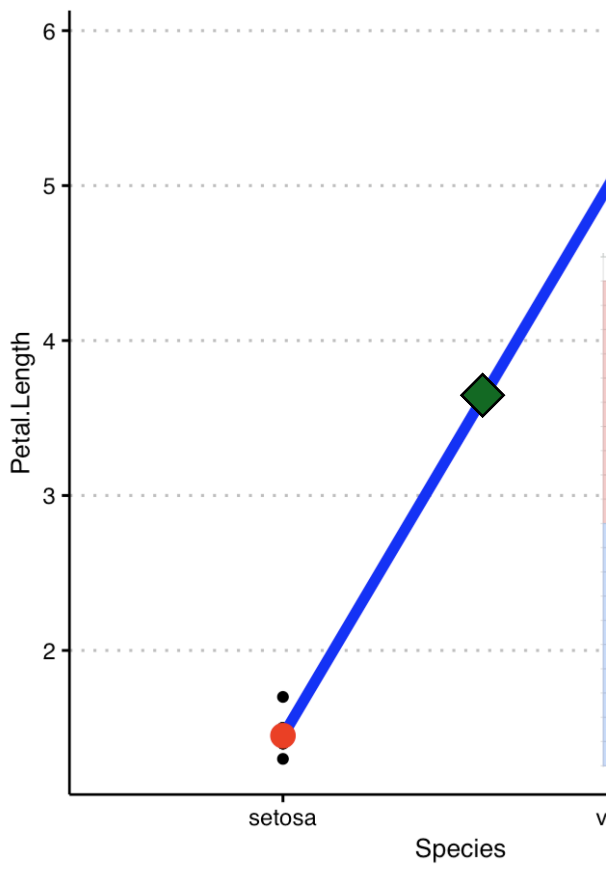
$$SS_{total} = \sum (Y - M_Y)^2$$

- for iris, SS_{total}= 76.218



step 3b: obtaining SS_{error}

- SS_{error} is the error that is left over after our species model has been fit
- our species model substitutes each raw score with the mean of the specific species
- $SS_{error} = \sum(Y - \hat{Y})^2 = \sum(Y - M_{group})^2$
- for iris, SS_{error}= .946



Species	Petal.Length(Y)	Yhat
setosa	1.4	1.45
setosa	1.4	1.45
setosa	1.3	1.45
setosa	1.5	1.45
setosa	1.4	1.45
setosa	1.7	1.45
setosa	1.4	1.45
setosa	1.5	1.45
setosa	1.4	1.45
setosa	1.5	1.45
virginica	5.6	5.33
virginica	5.1	5.33
virginica	5.1	5.33
virginica	5.9	5.33
virginica	5.7	5.33
virginica	5.2	5.33
virginica	5	5.33
virginica	5.2	5.33
virginica	5.4	5.33
virginica	5.1	5.33

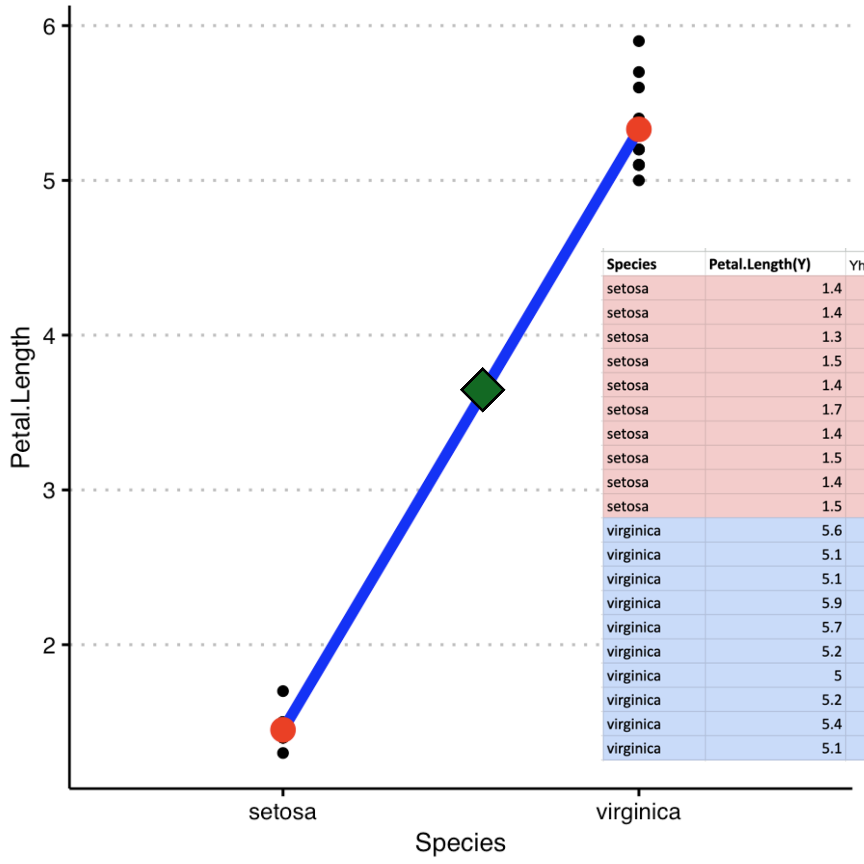
step 3c: obtaining SS_{model}

- how can we obtain SS_{model}?

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

thus, $SS_{model} = SS_{total} - SS_{error}$

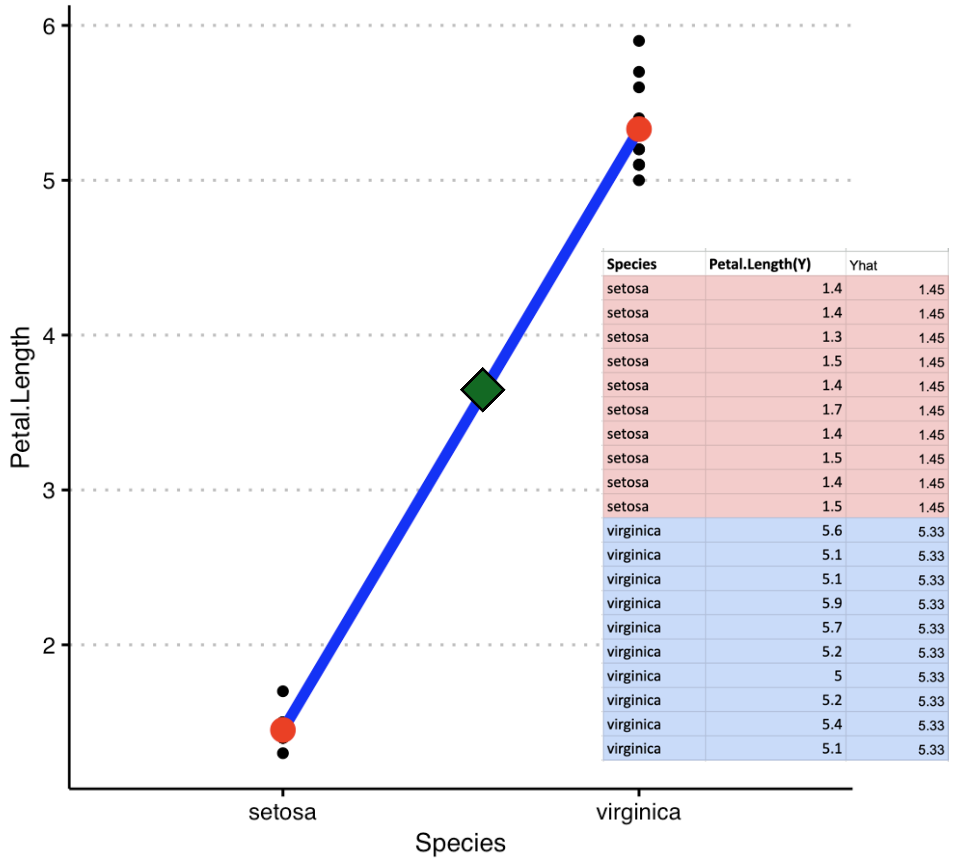
- for iris, SS_{total}= 76.218 and SS_{error}= .946
- SS_{model}= 75.272



Species	Petal.Length(Y)	Yhat
setosa	1.4	1.45
setosa	1.4	1.45
setosa	1.3	1.45
setosa	1.5	1.45
setosa	1.4	1.45
setosa	1.7	1.45
setosa	1.4	1.45
setosa	1.5	1.45
setosa	1.4	1.45
setosa	1.5	1.45
virginica	5.6	5.33
virginica	5.1	5.33
virginica	5.1	5.33
virginica	5.9	5.33
virginica	5.7	5.33
virginica	5.2	5.33
virginica	5	5.33
virginica	5.2	5.33
virginica	5.4	5.33
virginica	5.1	5.33

step 3d: obtaining $F_{observed}$

- $F_{observed} = \frac{MS_{model}}{MS_{error}} = \frac{SS_{model}/df_{model}}{SS_{error}/df_{error}}$ = 1432.24
- p-value = <.0001
- $F_{critical} = 4.414$
- thus, $F(1,18) =$ 1432.24, p < .0001
  - we can reject the null hypothesis
  - petal lengths of setosa and virginica are significantly different



Species	Petal.Length(Y)	Yhat
setosa	1.4	1.45
setosa	1.4	1.45
setosa	1.3	1.45
setosa	1.5	1.45
setosa	1.4	1.45
setosa	1.7	1.45
setosa	1.4	1.45
setosa	1.5	1.45
setosa	1.4	1.45
setosa	1.5	1.45
virginica	5.6	5.33
virginica	5.1	5.33
virginica	5.1	5.33
virginica	5.9	5.33
virginica	5.7	5.33
virginica	5.2	5.33
virginica	5	5.33
virginica	5.2	5.33
virginica	5.4	5.33
virginica	5.1	5.33

F-table

		SS	df	MS	F	p-value
SS_{model}	species	75.272	1	75.272	1432.24	<.0001
SS_{error}	residual	0.946	18	0.0526		

```
Response: Petal.Length
          Sum Sq Df F value    Pr(>F)
Species   75.272  1  1432.2 < 2.2e-16 ***
Residuals  0.946 18
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

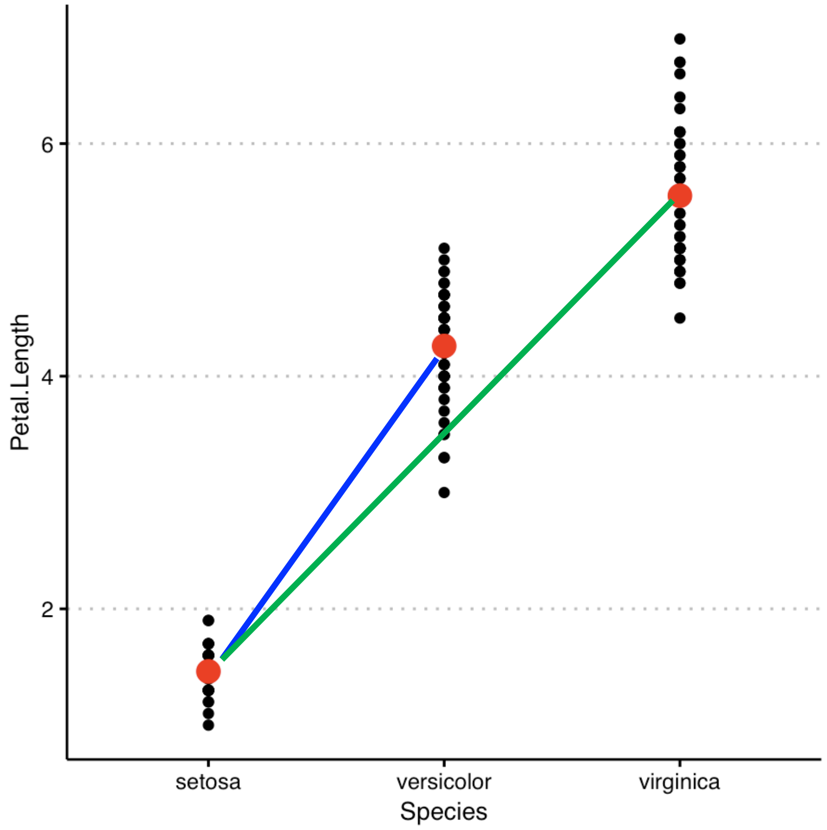
Sheets solution
video tutorial

types of ANOVAs = complex linear models

- n (independent variables)
  - one-way: one independent variable
  - two-way / three-way
  - (n>3)-way: crazy land
- within or between subjects
  - between subjects: regular ANOVA
    - independent observations: each raw score comes from *different* individuals!
  - within-subjects: repeated measures ANOVA
    - non-independent observations: multiple raw scores from from the same individuals

revisiting iris

- recall that the iris dataset actually contains information about three species (setosa, virginica, and versicolor)
- when more than two groups are involved, we need to expand our model to include multiple groups



NHST for one-way ANOVA

- **step 1: state the hypotheses**
  - H_0: no change in mean petal lengths due to species, i.e., $\mu_{setosa} = \mu_{virginica} = \mu_{versicolor}$
  - H_1: there is at least one mean difference (no claims about where!)
- **step 2: set criteria for decision**

$$F(df_1, df_2) = F_{critical}$$

- **step 3: collect data**
- **step 4: make a decision!**

NHST for one-way ANOVA

step 1: state the hypotheses

step 2: set criteria for decision

step 3: collect data

step 4: make a decision!

$H_0: \mu_1 = \mu_2 = \cdots = \mu_n$
$H_1: at\ least\ one\ mean\ difference$

$\alpha = .05$
find $F_{critical}$ based on **right** tailed test and degrees of freedom
$df_1 = k - 1$
$df_2 = n - k$

(1) compute SS_{model} and SS_{error}

(2) compute $F_{observed} = \frac{MS_{model}}{MS_{error}}$

(3) find p-value for F-score

check whether F is beyond $F_{critical}$ and p-value < .05. if so, reject null hypothesis!

next time

- special cases (independent t-tests, z-tests, etc.)

Here are the to-do's for this week:

- Submit Week 10 Quiz
- Submit Problem Set 4
- Submit any lingering questions here!
- Extra credit opportunities:
    - Submit Exra Credit Questions
    - Submit Optional Meme Submission

Before Thursday

- Review W10 Activity 1 Solutions.
- Watch: Hypothesis Testing (Linear Regression). (ok to watch after Tuesday!)
    - Practice Data
    - Solution Sheet
- Watch: Hypothesis Testing (Two groups F Test).
    - Practice Data
    - For solution, see the three groups solution below and adapt to two groups.

After Thursday

- Watch: Hypothesis Testing (One-way ANOVA).
    - Practice Data
    - Solution Sheet
- Watch: Completing F tables.
- See Apply section.

a puzzle

- how many pieces of information do you need to definitely guess the color of the traffic light?
- light is not green
- light is not red
- 2 pieces of information is enough

https://www.youtube.com/watch?v=rATNoxKg1yA

a puzzle

- the mean of quiz scores for 5 students is 9 points.
- what are the scores?
- what if I told you some of the numbers?
- four students’ scores are 8, 10, 8, and 9, what is the score of the fifth student?

degrees of freedom (df)

- main idea: how many pieces of information are needed to obtain a statistic?
- mean = $M = \frac{\sum X}{n}$
  - all values in a dataset are needed
  - why? because changing even a single score would change M
  - $df = n$
- standard deviation = $\frac{\sum(X-M)^2}{n-1}$
  - computing M restricts the scores that went into the calculation
  - if M is known, you only need to know n-1 scores to find the last score
  - only $n - 1$ scores are free to vary once M is known
  - for SD, effectively only $n - 1$ deviations are free to vary
  - $df = n - 1$

degrees of freedom (df)

- correlations
  - what is needed to calculate $t_{observed} = \frac{r - \rho}{SE_r}$?
  - *r*, which need two means to be estimated (everything else follows)
  - $df = n - 2$ for t-distribution of correlations
- another way to think about df : number of estimated parameters

degrees of freedom (df)

- simple linear regression ($\hat{Y} = a + bX$ where $b = r\,\frac{s_y}{s_x}$ and $a = M_y - bM_x$)
  - $F = \frac{MS_{model}}{MS_{error}} = \frac{SS_{model}/df_{model}}{SS_{error}/df_{error}}$
  - $SS_{model} = \sum(\hat{Y} - M_y)^2$
    - $k = 2$ total estimated parameters (b and a)
    - but knowing b restricts a so we lose one degree of freedom
    - $df_{model} = k - 1$
  - $SS_{error} = \sum(Y - \hat{Y})^2$
    - n observations and 2 total estimated parameters to compute $\hat{Y}$ (b and a)
    - $df_{error} = n - k$

F test for linear regression in R

```r
data("women")
View(women)

weight_model = lm(data = women, weight ~ height)
summary(weight_model)
car::Anova(weight_model)
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -87.51667    5.93694  -14.74 1.71e-09 ***
height        3.45000    0.09114   37.85 1.09e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

		SS	df	MS	F	p-value
SS_{model}	IV	3332.7	1	3332.7	1433.02	<.0001
SS_{error}	residual	30.23	13	2.33		

```
Anova Table (Type II tests)

Response: weight
          Sum Sq Df F value    Pr(>F)
height    3332.7  1    1433 1.091e-14 ***
Residuals   30.2 13
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```