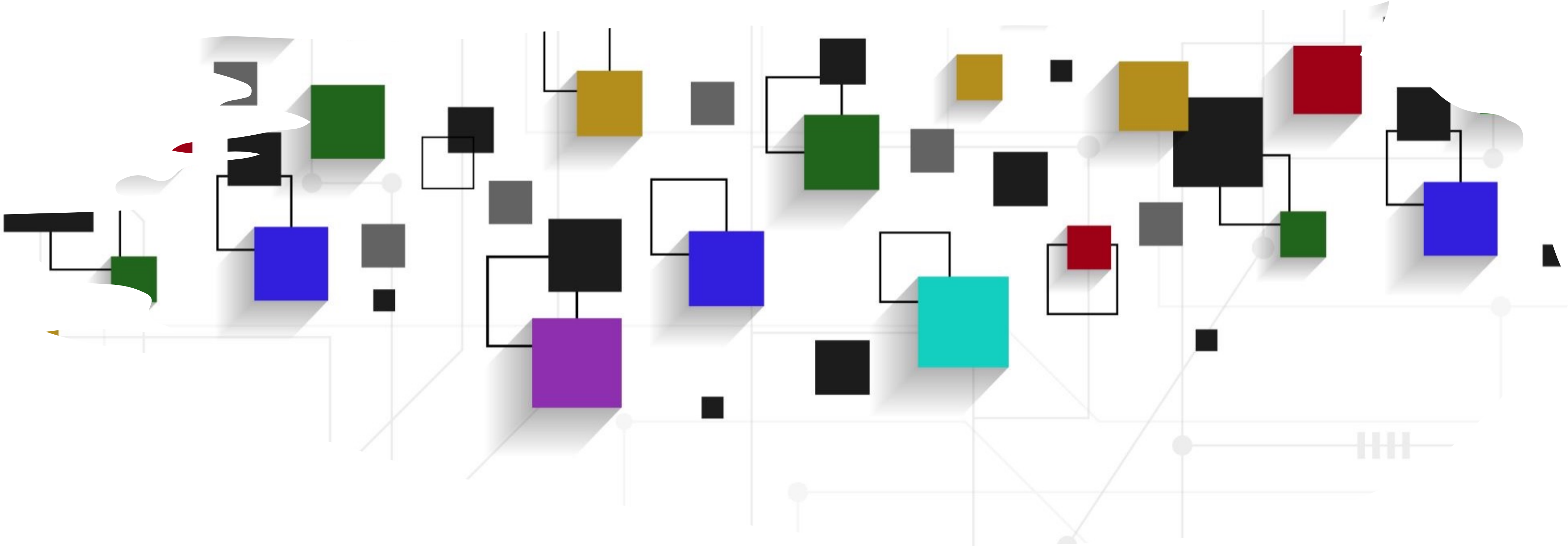


CogLab: Making Inferences

WEEK 11

recap: Oct 24/26, 2023

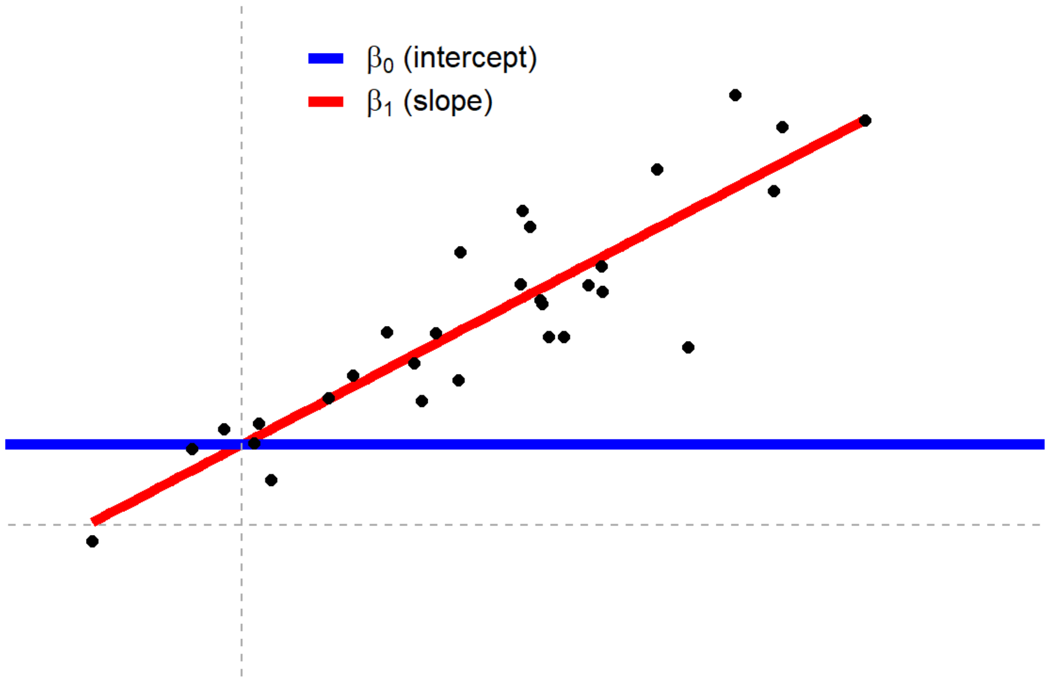
- what we covered:
  - manipulating data using tidyverse verbs
  - linear regression
- your to-do's were:
  - *prep*: complete all primers
  - *prep*: read about hypothesis testing
  - *schedule*: group meeting

today’s agenda

- linear regression continued
- two-way/multiple linear regression

linear regression

- a linear regression (or a linear model) is a model that fits a line to a set of data points
    - Y = aX + b
    - Y: dependent variable
    - X: independent variable
    - a? b?
- a: slope, b: intercept
- sometimes, we reorder this equation:
    - $y = \beta_0 + \beta_1 x$
    - β_0: intercept (where the line cuts the y-axis)
    - β_1: slope (the change in y due to x)
- in this framework, the null hypothesis (H_0) is that $\beta_1 = 0$, i.e., there is no change in y due to x
    - $H_0: \beta_1 = 0$



https://lindeloev.github.io/tests-as-linear/

linear regression in R

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

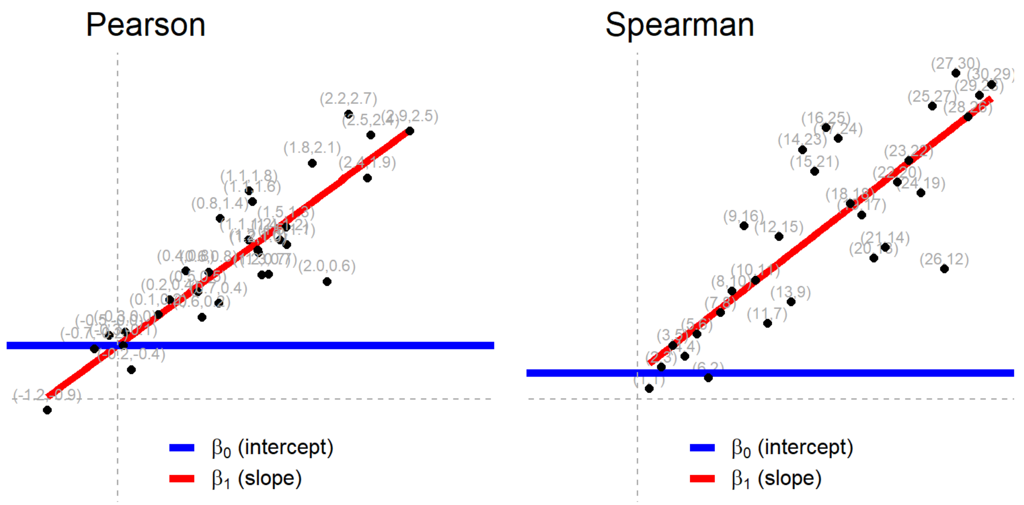
```
women_model = lm(data = women, height ~ weight)
```

```
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
```

linear regression and correlation

- correlations also describe the relationship between Y and X, so what's the difference?
- mathematically, correlations are equivalent to a linear model where a line is being fit to a set of data points
- two common correlation
  - Pearson's *r*: *r* = slope if x and y have the same standard deviation
  - Spearman's *rho* = same linear model but with ranks of x and Y
    - $\text{rank}(y) = \beta_0 + \beta_1\, \text{rank}(x)$



https://lindeloev.github.io/tests-as-linear/

linear regression and correlation

- compute the standard deviation of the height and weight columns
- create two new columns that contain the z-scored height and weight
- compute the standard deviation of the z-scored height and weight columns

```
sd(women$height)
sd(women$weight)
```

```
women = women %>%
  mutate(z_height = scale(height),
         z_weight = scale(weight))
```

```
sd(women$height)
sd(women$weight)
```

linear regression and correlation

- predict the z-scored height with the z-scored weight using linear regression
- now compute the correlation between the two columns using summarize() and cor()

```r
women_model_2 = lm(data = women, z_height ~ z_weight)
summary(women_model_2)
```

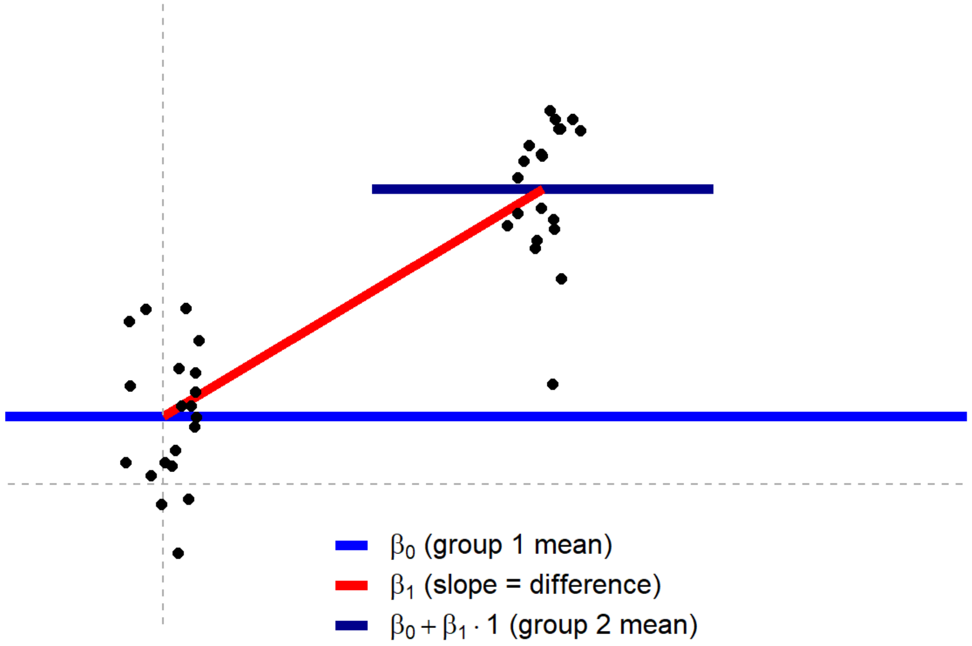
```
Call:
lm(formula = z_height ~ z_weight, data = women)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.18611 -0.05869  0.01859  0.07682  0.11133 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -8.268e-16  2.541e-02    0.00        1    
z_weight     9.955e-01  2.630e-02   37.85 1.09e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.0984 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
```

```r
women %>%
  summarise(r = cor(z_height, z_weight))
```

```
          r
1 0.9954948
```

linear regression and t-tests

- unpaired/independent samples t-test
  - $y = \beta_0 + \beta_1 x$
  - x = 0 or 1 (which group)
  - H_0: $\beta_1 = 0$
  - comparing paired differences and testing whether the difference is significantly different from 0
  - note that "x" here contains information about group membership for each y

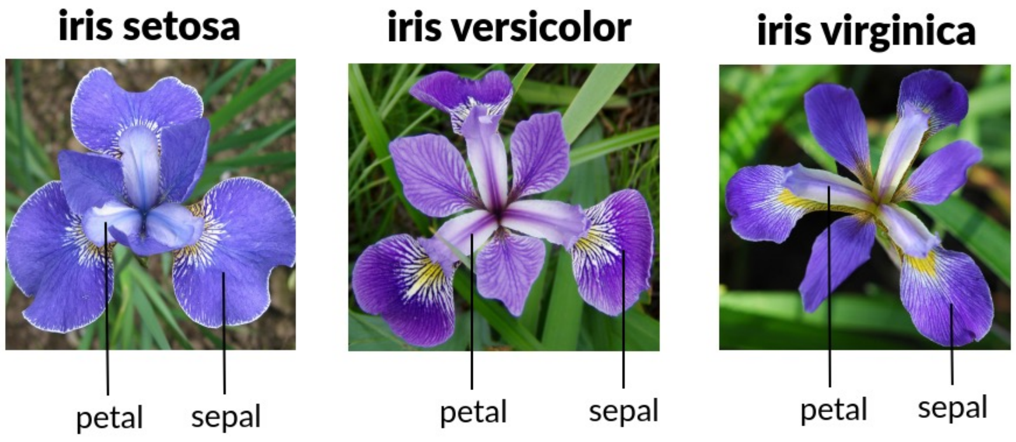


revisiting iris

- recall that iris contains flower petal and sepal information for three species

```
data("iris")
View(iris)
```

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
5.1	3.5	1.4	0.2	setosa
4.9	3.0	1.4	0.2	setosa
4.7	3.2	1.3	0.2	setosa
4.6	3.1	1.5	0.2	setosa
5.0	3.6	1.4	0.2	setosa
5.4	3.9	1.7	0.4	setosa
4.6	3.4	1.4	0.3	setosa
5.0	3.4	1.5	0.2	setosa
4.4	2.9	1.4	0.2	setosa
4.9	3.1	1.5	0.1	setosa
5.4	3.7	1.5	0.2	setosa
4.8	3.4	1.6	0.2	setosa
4.8	3.0	1.4	0.1	setosa
4.3	3.0	1.1	0.1	setosa
5.8	4.0	1.2	0.2	setosa
5.7	4.4	1.5	0.4	setosa



subset of iris

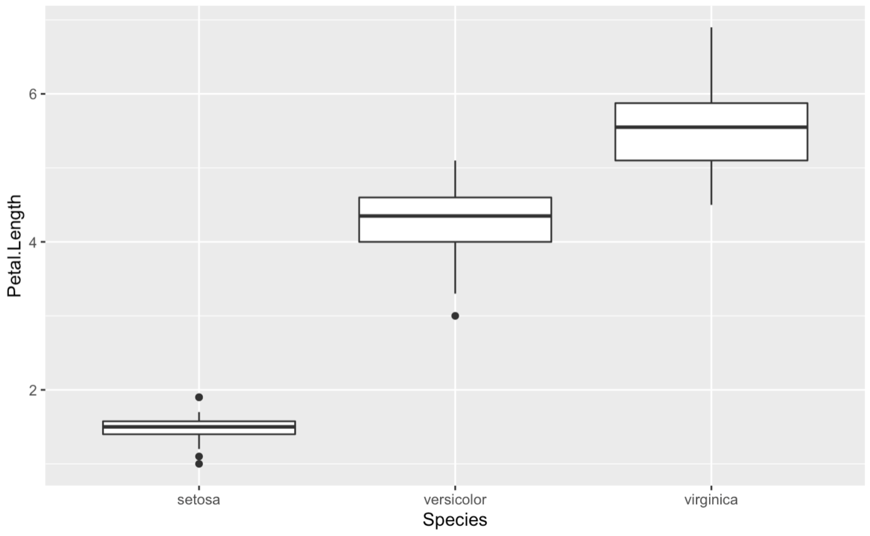
- create a subset of iris that only contains setosa and virginica
- plot the petal lengths by species in a boxplot

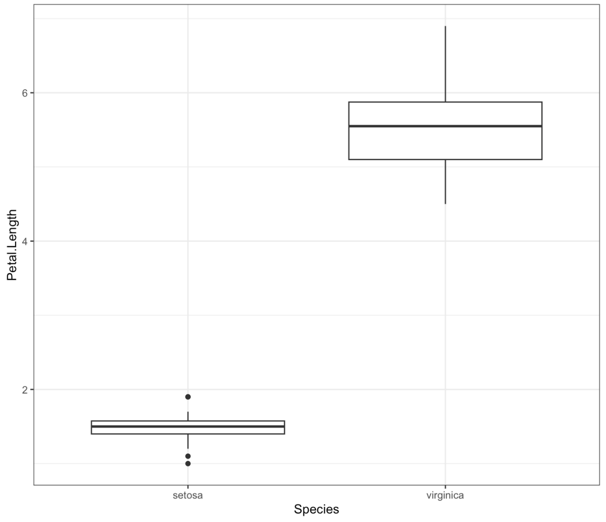
```
## t -test

```{r}
iris_subset = iris %>%
  filter(Species %in% c("setosa", "virginica"))

```
```

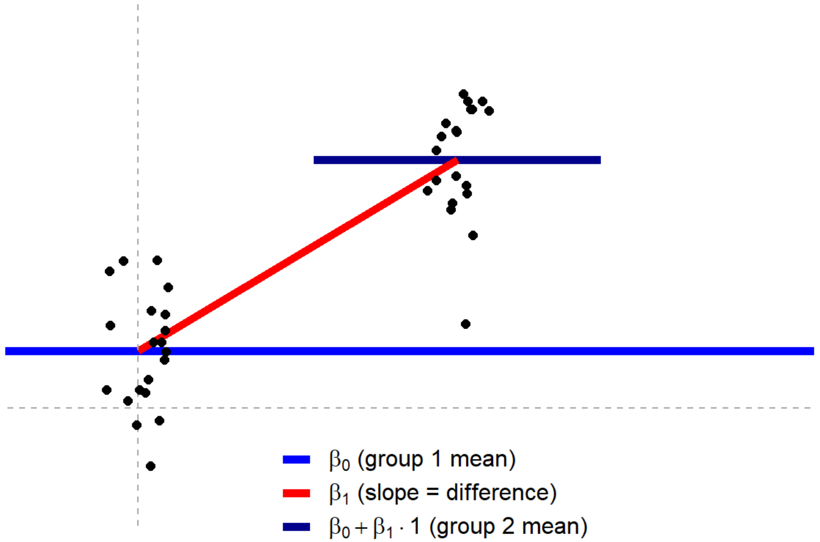
```
iris_subset %>%
  ggplot(aes(x = Species, y = Petal.Length))+
  geom_col()
```





comparing

- create linear model
- conduct t-test




```r
iris_subset_lm = lm(data = iris_subset, Petal.Length ~ Species)
summary(iris_subset_lm)
```

```
Call:
lm(formula = Petal.Length ~ Species, data = iris_subset)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.0520 -0.1620  0.0380  0.1405  1.3480 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)       1.46200    0.05786   25.27   <2e-16 ***
Speciesvirginica  4.09000    0.08182   49.99   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4091 on 98 degrees of freedom
Multiple R-squared:  0.9623,	Adjusted R-squared:  0.9619 
F-statistic:  2499 on 1 and 98 DF,  p-value: < 2.2e-16
```

```r
t.test(Petal.Length ~ Species, data = iris_subset)
```

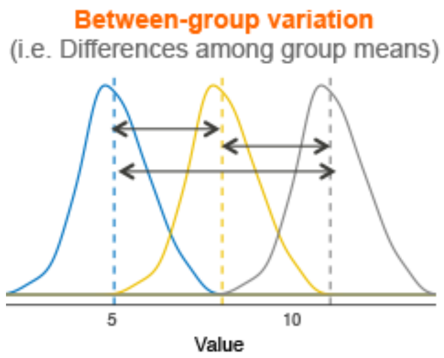
```
	Welch Two Sample t-test

data:  Petal.Length by Species
t = -49.986, df = 58.609, p-value < 2.2e-16
alternative hypothesis: true difference in means between group setosa and group virginica is not equal to 0
95 percent confidence interval:
 -4.253749 -3.926251
sample estimates:
   mean in group setosa mean in group virginica 
                  1.462                   5.552 
```

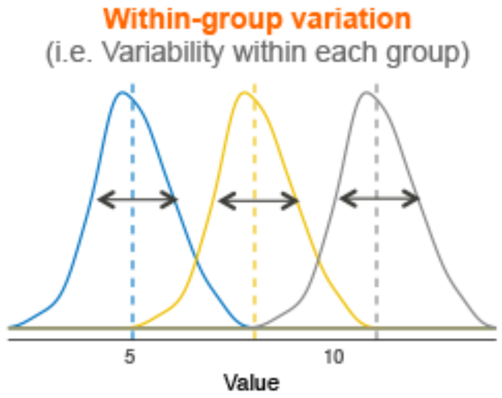
testing more than two groups

- a t-test is a special case of linear models
- it is *also* a special case of only comparing two groups
- example of comparing more than two groups?

ANOVA: Analysis of Variance

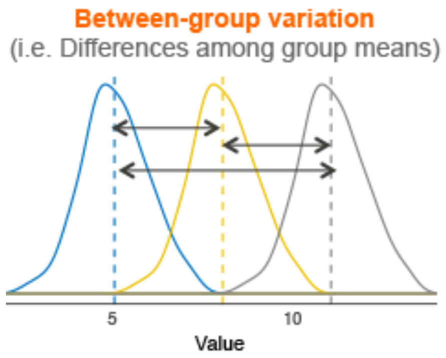
- a *generalized* t-test for more than two means/groups!
- key idea: we will try to understand the difference between groups and whether it can be attributed to our "conditions" or randomness
- $SS_{between}$ = variation between groups
- SS_{within} = variation within groups
- $F = SS_{between}/Ss_{within}$
- If F > 1, the group differences are greater than what would be expected as random variation within groups

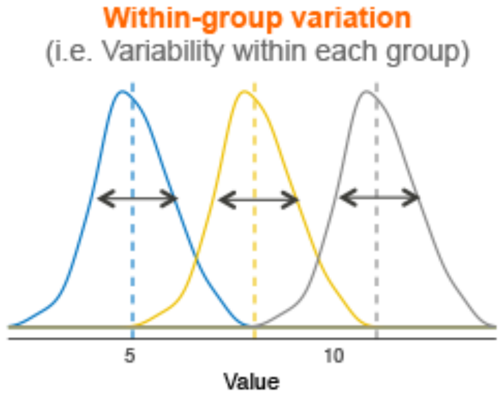




types of ANOVAs

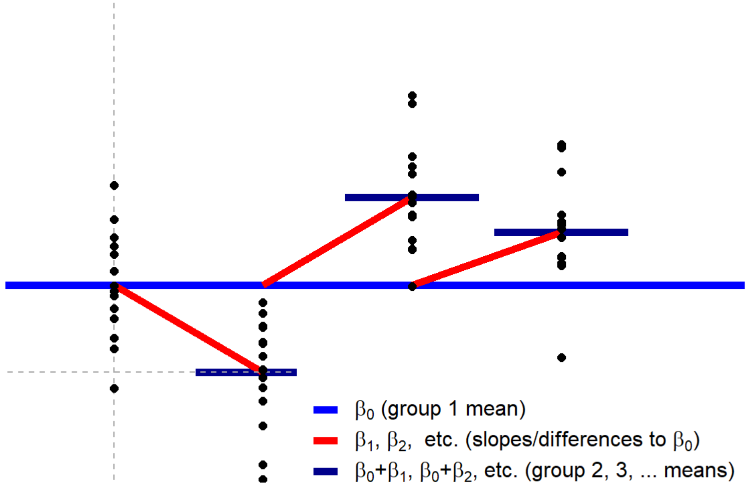
- n(independent variables)
  - one-way
  - two-way
  - three-way
- within or between subjects
  - between subjects: regular ANOVA
  - within-subjects: repeated measures ANOVA





one-way ANOVA

- predict the petal lengths using the full iris dataset




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

```
Call:
lm(formula = Petal.Length ~ Species, data = iris)

Residuals:
   Min     1Q Median     3Q    Max 
-1.260 -0.258  0.038  0.240  1.348 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)        1.46200    0.06086   24.02   <2e-16 ***
Speciesversicolor  2.79800    0.08607   32.51   <2e-16 ***
Speciesvirginica   4.09000    0.08607   47.52   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4303 on 147 degrees of freedom
Multiple R-squared:  0.9414,    Adjusted R-squared:  0.9406 
F-statistic:  1180 on 2 and 147 DF,  p-value: < 2.2e-16
```

```
full_iris_aov = aov(data = iris, Petal.Length ~ Species)
summary(full_iris_aov)
```

```
             Df Sum Sq Mean Sq F value Pr(>F)    
Species       2  437.1  218.55    1180 <2e-16 ***
Residuals   147   27.2    0.19                   
---
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
- compare to lm summary

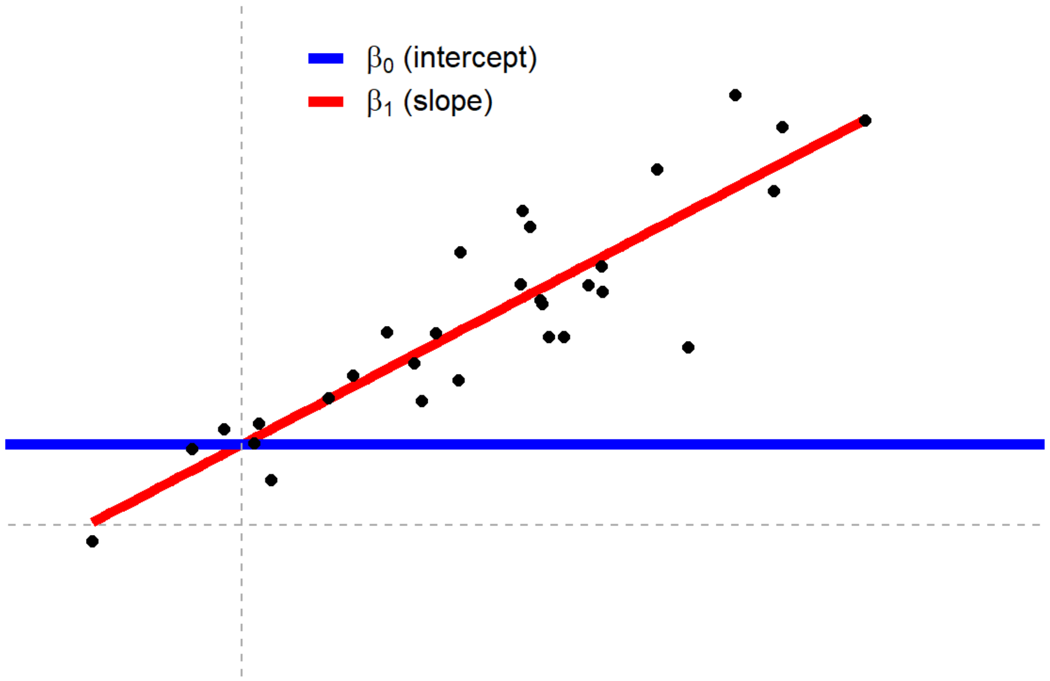
```
Call:
lm(formula = Petal.Length ~ Species, data = iris)

Residuals:
   Min     1Q Median     3Q    Max 
-1.260 -0.258  0.038  0.240  1.348 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)        1.46200    0.06086   24.02   <2e-16 ***
Speciesversicolor  2.79800    0.08607   32.51   <2e-16 ***
Speciesvirginica   4.09000    0.08607   47.52   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4303 on 147 degrees of freedom
Multiple R-squared:  0.9414,	Adjusted R-squared:  0.9406 
F-statistic:  1180 on 2 and 147 DF,  p-value: < 2.2e-16
```

```
#install.packages("emmeans")
emmeans::emmeans(full_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 
```

linear model: assumptions

- "all models are wrong, but some are useful" (Box, 1976)
- the model does not know where the data come from or whether they are appropriate for the model that is your responsibility as a researcher
  - linearity
  - normality of residuals
  - homoskedasticity
  - independence of observations

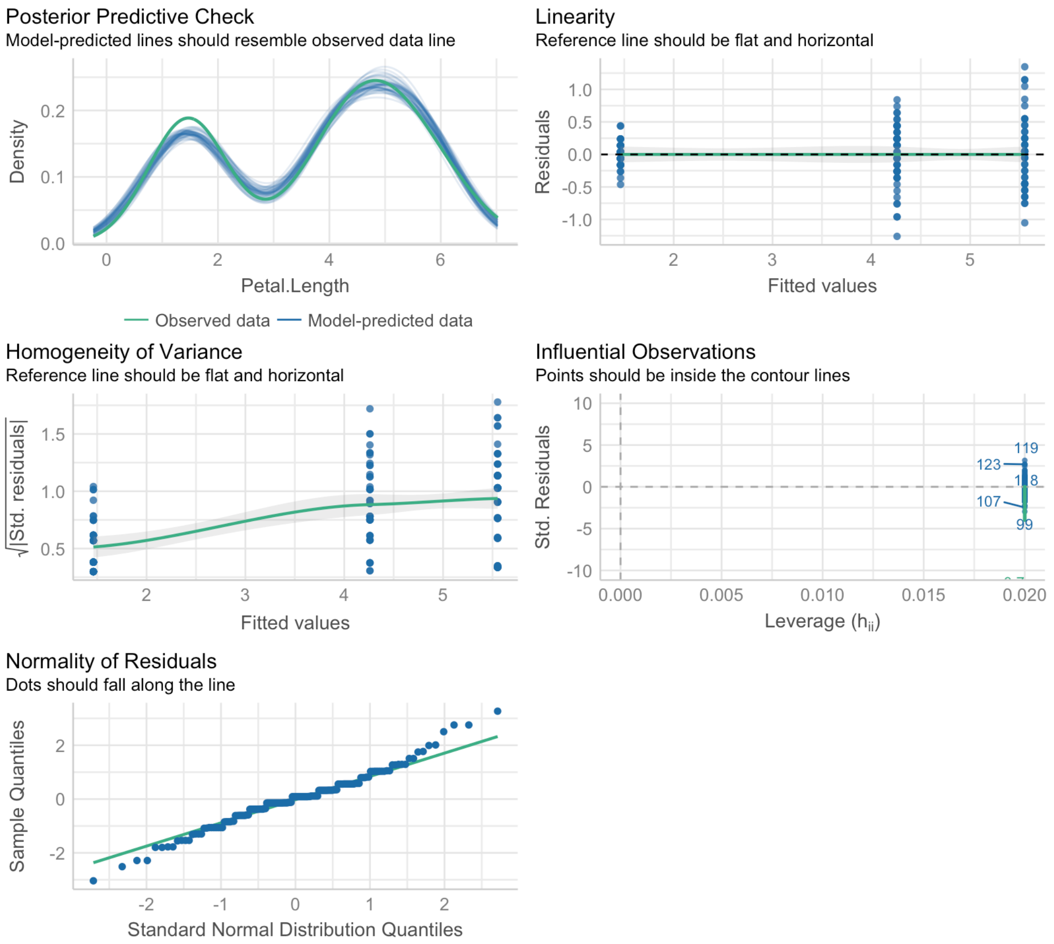


https://lindeloev.github.io/tests-as-linear/

inspecting the model

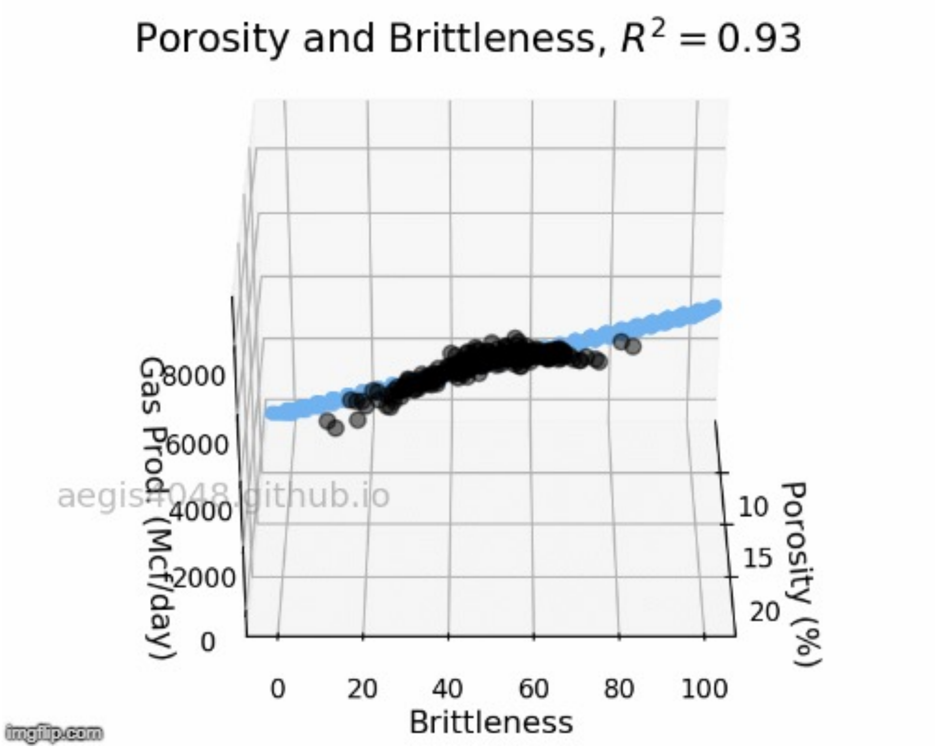
```r
## assumptions

#install.packages("performance")
#install.packages("see")

library(performance)
check_model(full_iris_model)
```

- first we install the performance and see packages
- load performance
- check the model
- minor variations are ok, major variations are warnings!



multiple linear regression

- often, we want to look at the influence of more than one variable on our response measures
- a multiple linear regression is a model that attempts to find the relationship between a dependent variable and more than one independent variable
  - $Y = aX_1 + bX_2 + c$
  - Y: dependent variable
  - $X_{1,2}$: independent variables



multiple linear regression: data

- we will use the jobsatisfaction dataset from the datarium package
- install the package datarium
- new heading (# multiple linear regression) & code chunk
- load and view the jobsatisfaction dataset

```
data("jobsatisfaction", package = "datarium")
View(jobsatisfaction)
```

id	gender	education_level	score
1	male	school	5.51
2	male	school	5.65
3	male	school	5.07
4	male	school	5.51
5	male	school	5.94
6	male	school	5.80
7	male	school	5.22
8	male	school	5.36
9	male	school	4.78
10	male	college	6.01
11	male	college	6.01
12	male	college	6.45

multiple linear regression: exploration

- let's explore the data:
  - find the mean and standard deviation of the score for each level of gender and education level

multiple linear regression: exploration

- let's explore the data:
  - find the mean and standard deviation of the score for each level of gender and education level

```
jobsatisfaction %>%
  group_by(gender, education_level) %>%
  summarize(mean = mean(score),
            sd = sd(score))
```

```
# A tibble: 6 x 4
# Groups:   gender [2]
  gender education_level  mean    sd
  <fct>  <fct>           <dbl> <dbl>
1 male   school           5.43 0.364
2 male   college          6.22 0.340
3 male   university       9.29 0.445
4 female school           5.74 0.474
5 female college          6.46 0.475
6 female university       8.41 0.938
```

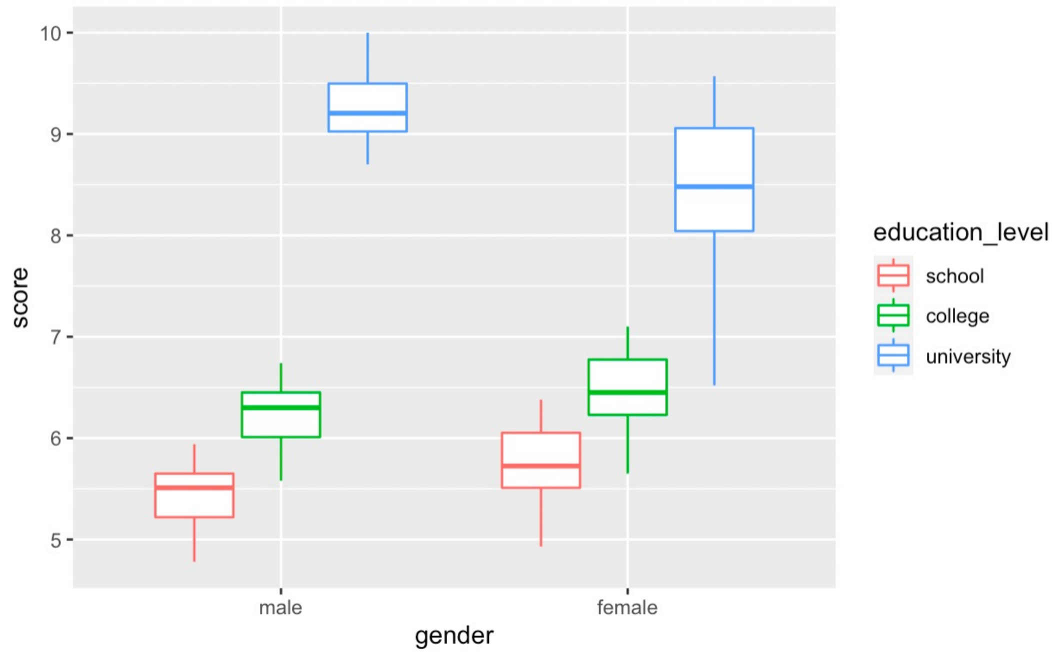
multiple linear regression: exploration

- let's explore the data:
  - visualize the pattern via a boxplot

multiple linear regression: exploration

- let's explore the data:
  - visualize the pattern via a boxplot
  - do you see differences in job satisfaction?

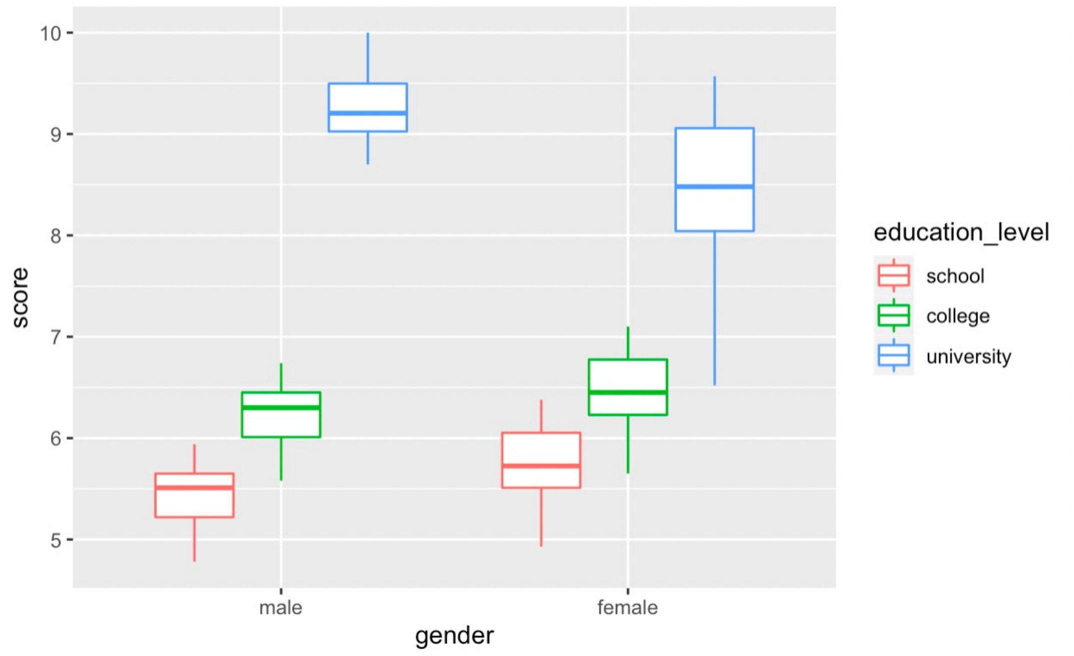
```
jobsatisfaction %>%
  ggplot()+
  geom_boxplot(aes(x = gender, y = score, color = education_level))
```



multiple linear regression: research question

- does job satisfaction vary as a function of gender and education level?
- dependent variable?
- independent variable?

```
jobsatisfaction %>%
  ggplot()+
  geom_boxplot(aes(x = gender, y = score, color = education_level))
```



main effects

- when you have multiple variables in your experiment design, there are few different possibilities for how the pattern of data might look
- you could have the dependent variable vary as a function of IV1 and/or IV2 (main effects), and these effects might interact with each other
- main effects refer to differences in means of levels of an independent variable
- what is an example of a main effect for the jobsatisfaction dataset?
- what would the plot of this main effect look like?

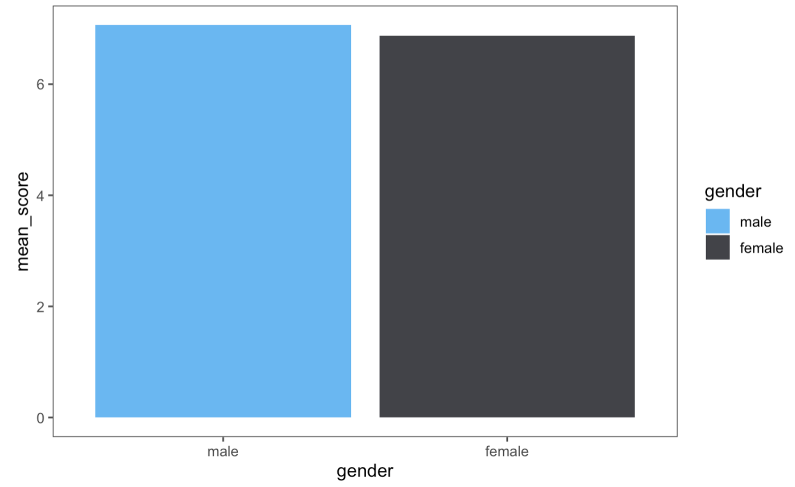
id	gender	education_level	score
1	male	school	5.51
2	male	school	5.65
3	male	school	5.07
4	male	school	5.51
5	male	school	5.94
6	male	school	5.80
7	male	school	5.22
8	male	school	5.36
9	male	school	4.78
10	male	college	6.01
11	male	college	6.01
12	male	college	6.45

interactions

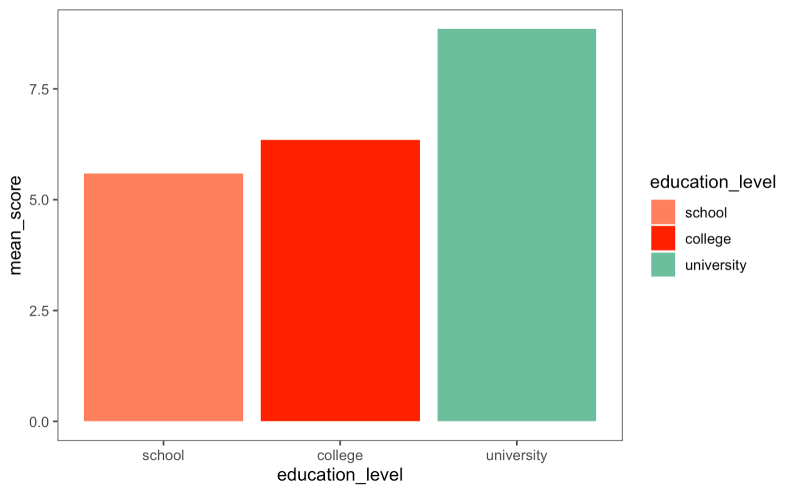
- interactions refer to situations when the difference in means between IV1's levels differs based on the levels of IV2, i.e., you cannot simply infer a difference in means
- what is an example of an interaction for the jobsatisfaction dataset?
- what would the plot of this main effect look like?

id	gender	education_level	score
1	male	school	5.51
2	male	school	5.65
3	male	school	5.07
4	male	school	5.51
5	male	school	5.94
6	male	school	5.80
7	male	school	5.22
8	male	school	5.36
9	male	school	4.78
10	male	college	6.01
11	male	college	6.01
12	male	college	6.45

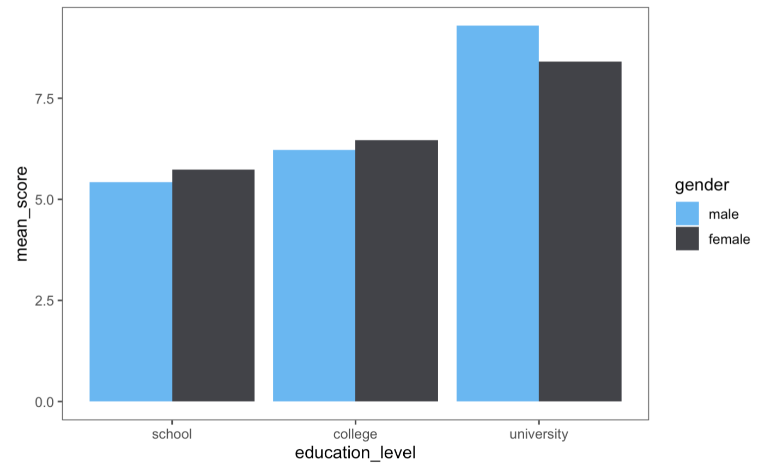
visually…



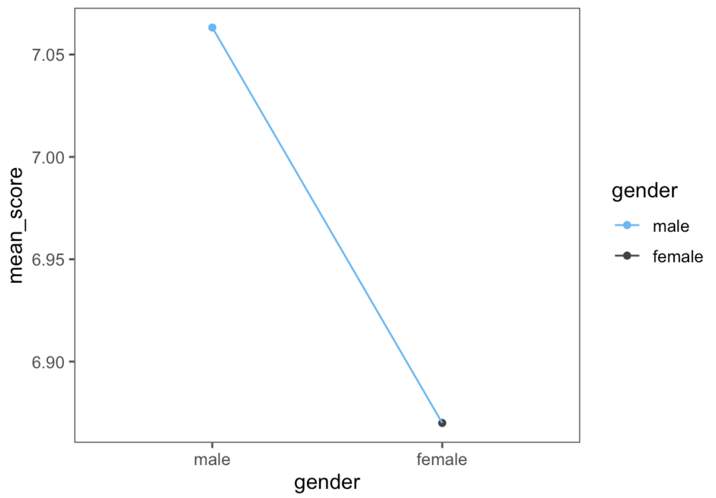
mean_score
0
2
4
6
male
female
gender
gender
male
female



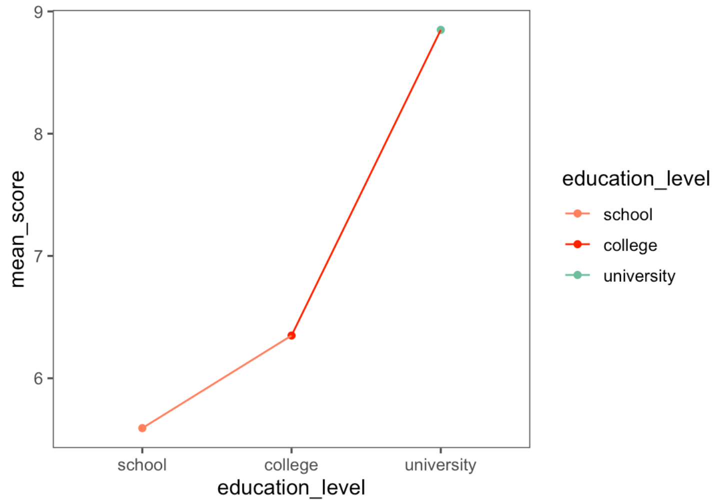
mean_score
0.0
2.5
5.0
7.5
school
college
university
education_level
education_level
school
college
university



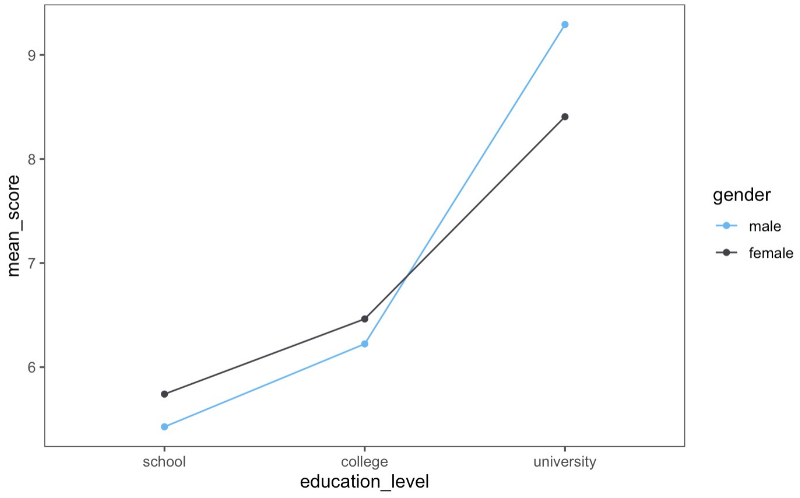
mean_score
0.0
2.5
5.0
7.5
school
college
university
education_level
gender
male
female



mean_score
6.90
6.95
7.00
7.05
male
female
gender
gender
male
female



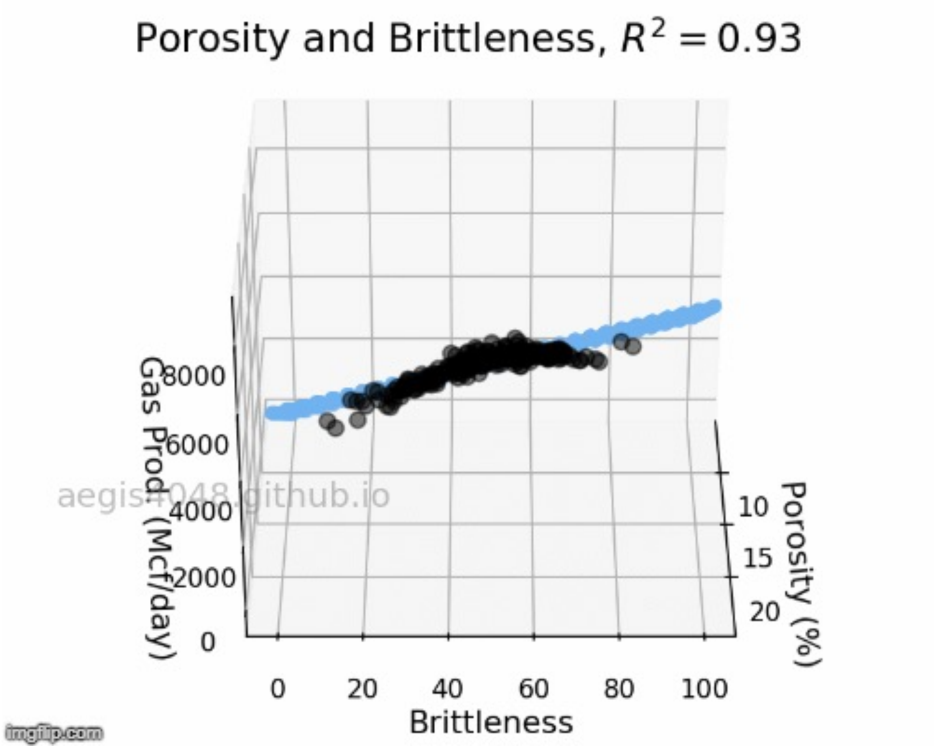
mean_score
6
7
8
9
school
college
university
education_level
education_level
school
college
university



mean_score
6
7
8
9
school
college
university
education_level
gender
male
female

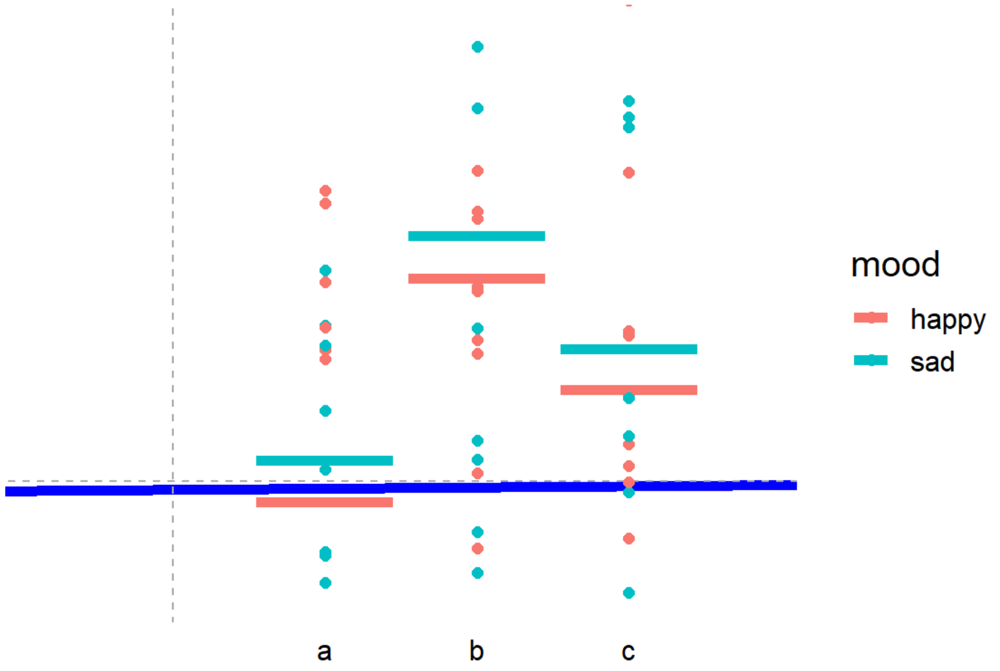
multiple linear regression

- often, we want to look at the influence of more than one variable on our response measures
- a multiple linear regression is a model that attempts to find the relationship between a dependent variable and more than one independent variable
  - $Y = aX_1 + bX_2 + c$
  - Y: dependent variable
  - $X_{1,2}$: independent variables



linear regression and ANOVAs

- ANOVAs are special cases of linear regression models, when the predictors are *categorical*
- two-way ANOVA equation
  - $y=\beta_0+\beta_1X_1+\beta_2X_2+\beta_3X_1X_2$
  - note that the X's here are different independent variables
  - H_0: $\beta_1=0$ (for X_1 main effect)
  - H_0: $\beta_2=0$ (for X_2 main effect)
  - H_0: $\beta_3=0$ (for interaction)

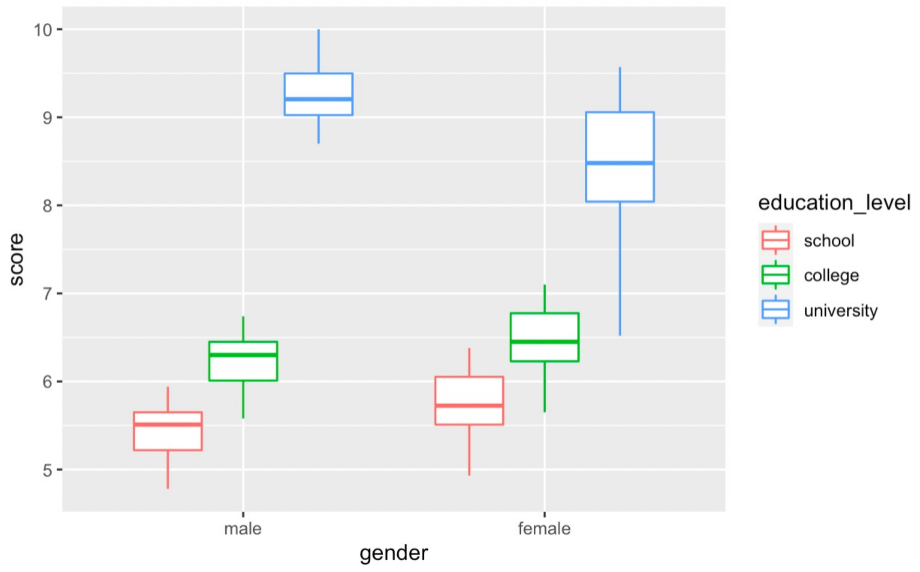


https://lindeloev.github.io/tests-as-linear/

mathematically…

- main effect of gender:
  - mean (male) – mean (female)
- main effect of education level
  - mean(school) – mean (college)
  - mean(college) – mean (university)
  - mean(university) – mean(school)
- interaction (difference of differences)
  - $\text{diff(male-female)}_{\text{school}} - \text{diff(male-female)}_{\text{college}}$
  - $\text{diff(male-female)}_{\text{university}} - \text{diff(male-female)}_{\text{college}}$
  - $\text{diff(male-female)}_{\text{school}} - \text{diff(male-female)}_{\text{university}}$

gender <fctr>	education_level <fctr>	mean <dbl>	sd <dbl>
male	school	5.426667	0.3638681
male	college	6.223333	0.3396322
male	university	9.292000	0.4445422
female	school	5.741000	0.4744225
female	college	6.463000	0.4746941
female	university	8.406000	0.9379078



next class

- **before** class
  - *resubmit:* formative assignment #2
  - *finalize:* experiment
  - submit: pre-registration
- **during** class
  - multiple regression in R
  - linear models for non-independent data