DSC 152:
Applied Statistical Data Analysis and Inference

Lecture #12
Interactions Terms Continued

Thursday, May 7
Spring Quarter 2026
Peter Chi

Mid-Quarter Survey Summary

How often do you attend lecture in-person?

132 responses

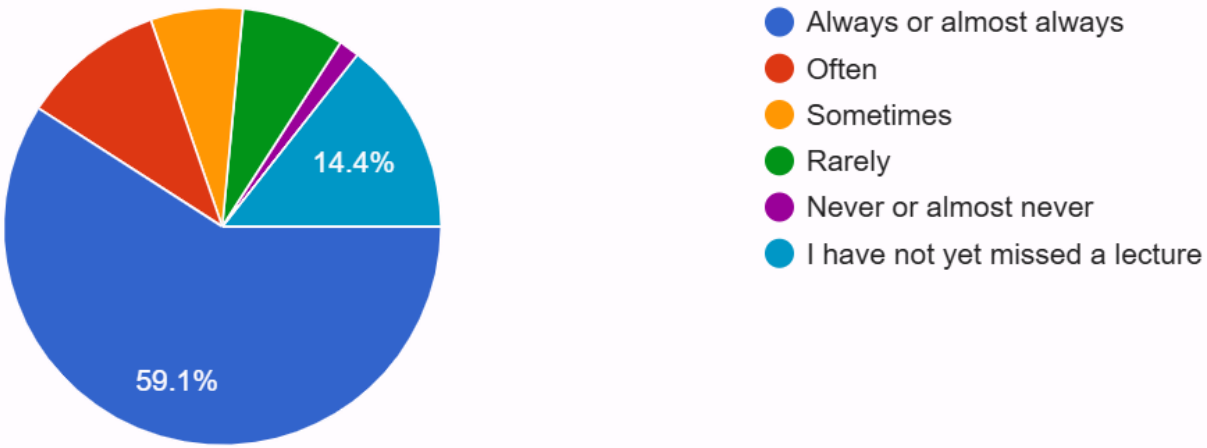
Copy chart

Mid-Quarter Survey Summary

When you miss lecture, how often do you watch the podcast?

132 responses

Copy chart



Mid-Quarter Survey Summary

Lecture Feedback

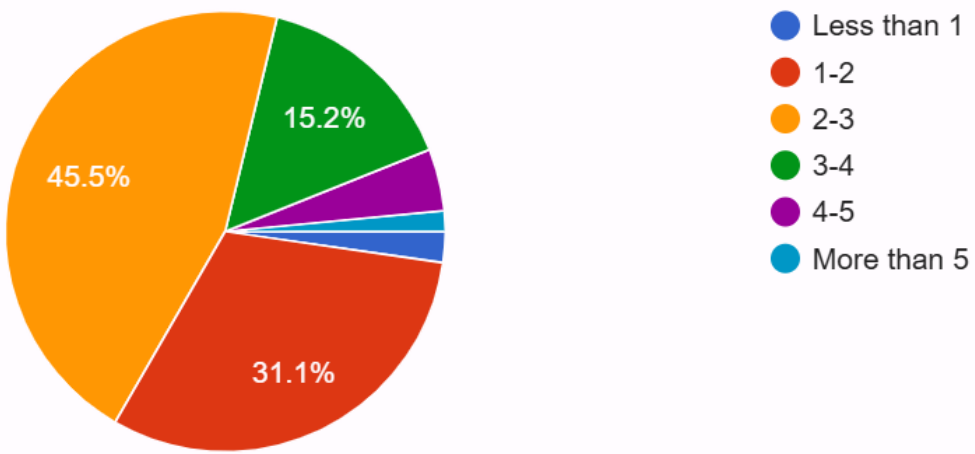
- Mostly positive things
- Many comments about the class being too early
- A few people said I go too fast sometimes

Mid-Quarter Survey Summary

How many hours do you typically spend on a lab assignment? Although the labs have been different in length, please just estimate your average time as best as you can.

Copy chart

132 responses

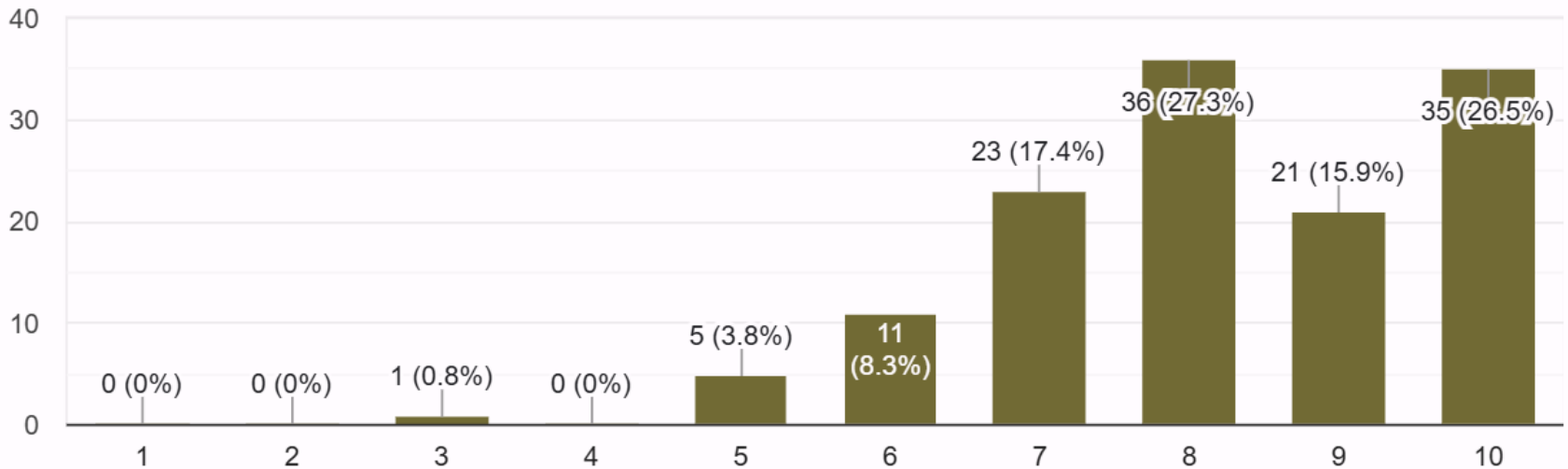


Mid-Quarter Survey Summary

On a scale of 1 to 10, how would you rate Lab 1 in terms of how it helped your understanding of its material?

Copy chart

132 responses

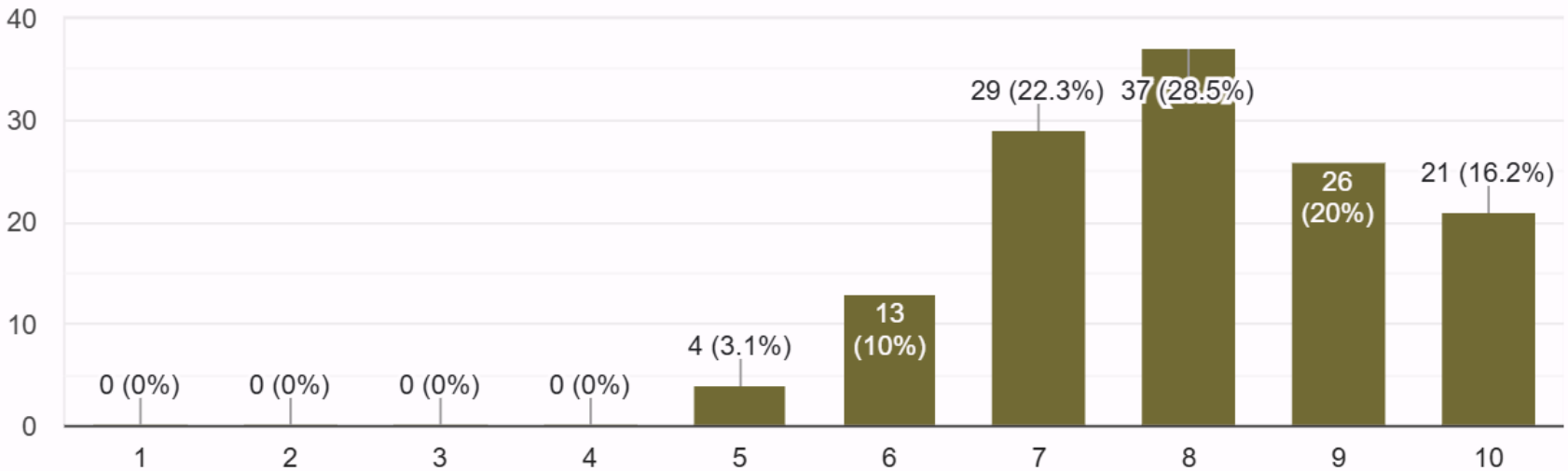


Mid-Quarter Survey Summary

On a scale of 1 to 10, how would you rate Lab 2 in terms of how it helped your understanding of its material?

Copy chart

130 responses

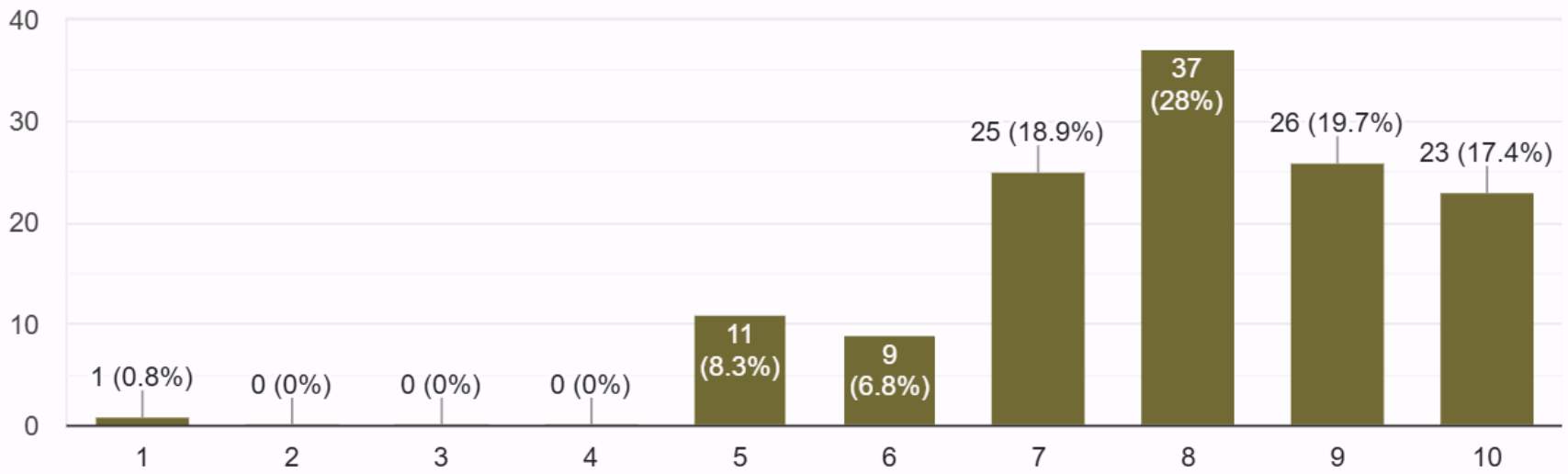


Mid-Quarter Survey Summary

On a scale of 1 to 10, how would you rate Lab 3 in terms of how it helped your understanding of its material?

Copy chart

132 responses

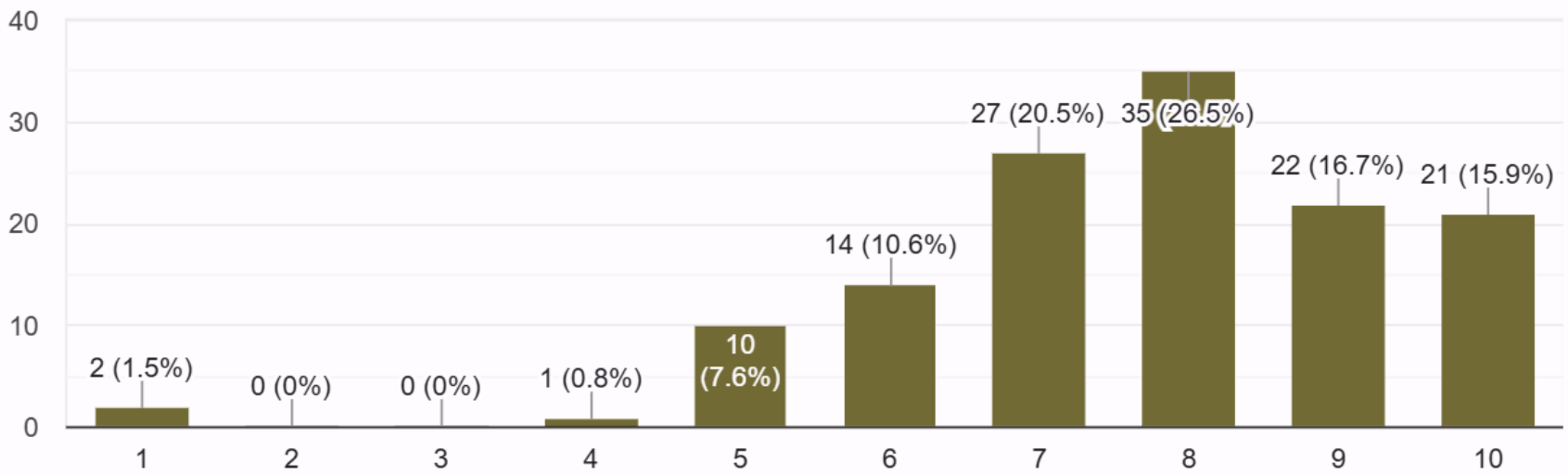


Mid-Quarter Survey Summary

On a scale of 1 to 10, how would you rate Lab 4 in terms of how it helped your understanding of its material?

Copy chart

132 responses

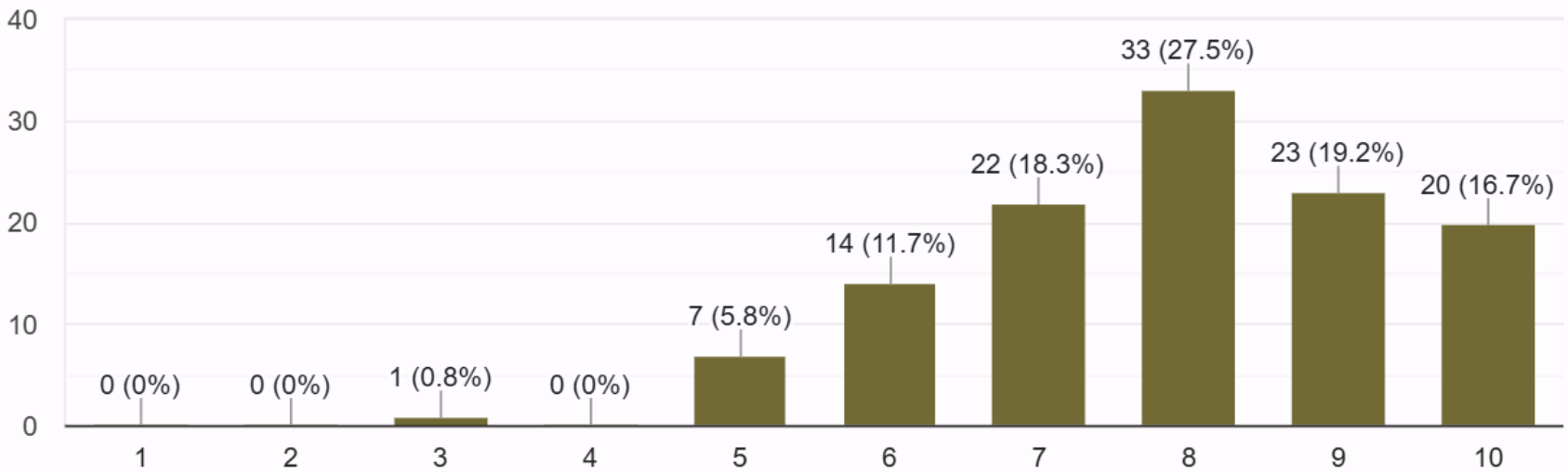


Mid-Quarter Survey Summary

On a scale of 1 to 10, if you have completed or even started Lab 5, how would you rate it in terms of how it helped your understanding of its material?

Copy chart

120 responses



Mid-Quarter Survey Summary

Lab Feedback

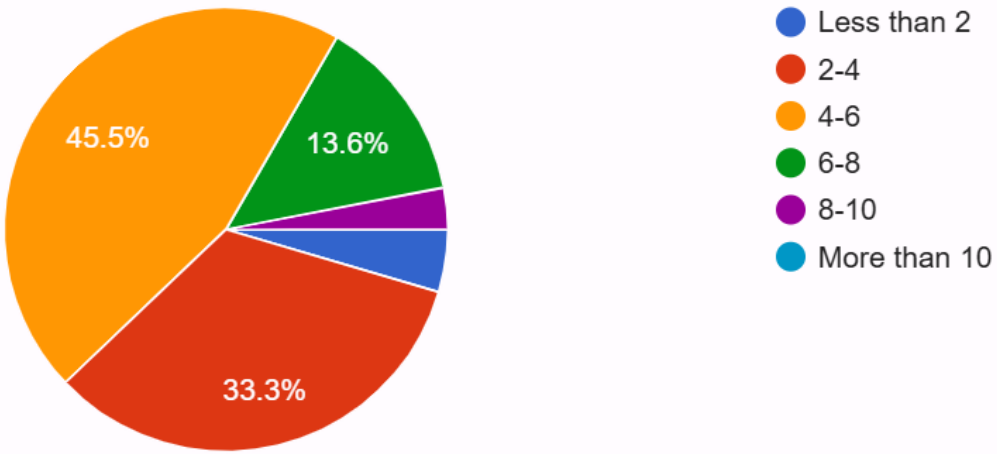
- Lab 1: Mostly good/fine/easy
- Lab 2: Maybe too easy?
- Lab 3: "The sheer tedium of this lab single handedly drilled the impact of distributions and stuff on power, etc into my head... I didn't love working on it, but it's my favorite/most effective lab so far."
- Lab 4: Long
- Lab 5: "I thought it was very clearly presented and easy to follow! Very easy to finish and relearn about the violations/conditions for linear regression & on full/reduced models."
  - Also there was some longer constructive feedback, thank you!

Mid-Quarter Survey Summary

How many hours did you spend on Homework 1?

132 responses

Copy chart

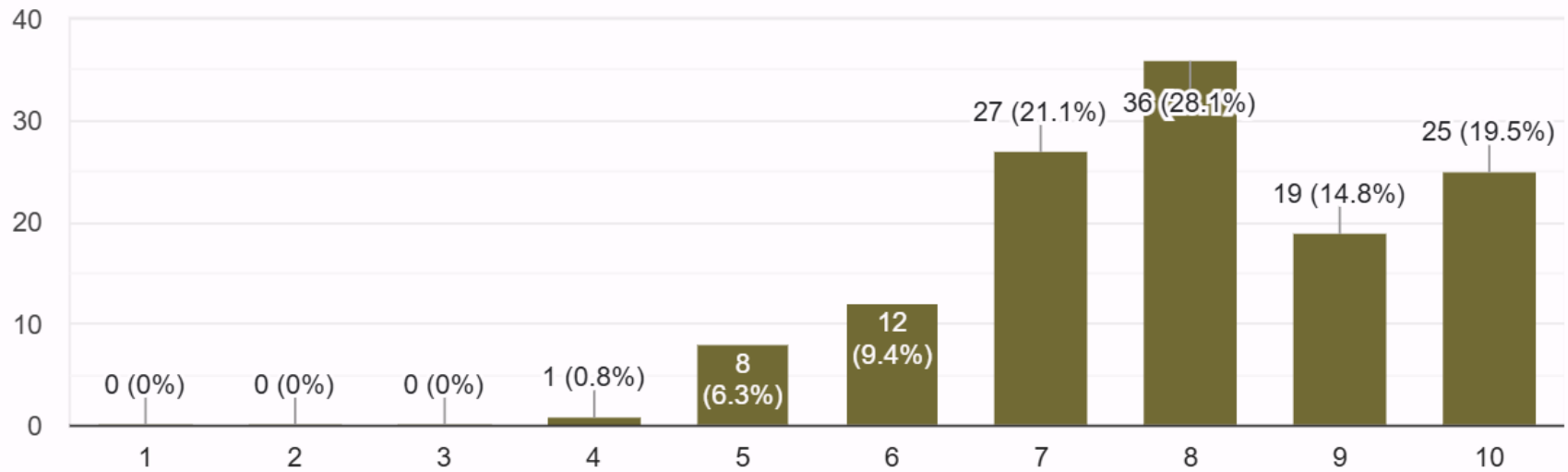


Mid-Quarter Survey Summary

On a scale of 1 to 10, how would you rate Homework 1 in terms of how it helped your understanding of its material?

Copy chart

128 responses



Mid-Quarter Survey Summary

HW1 Feedback

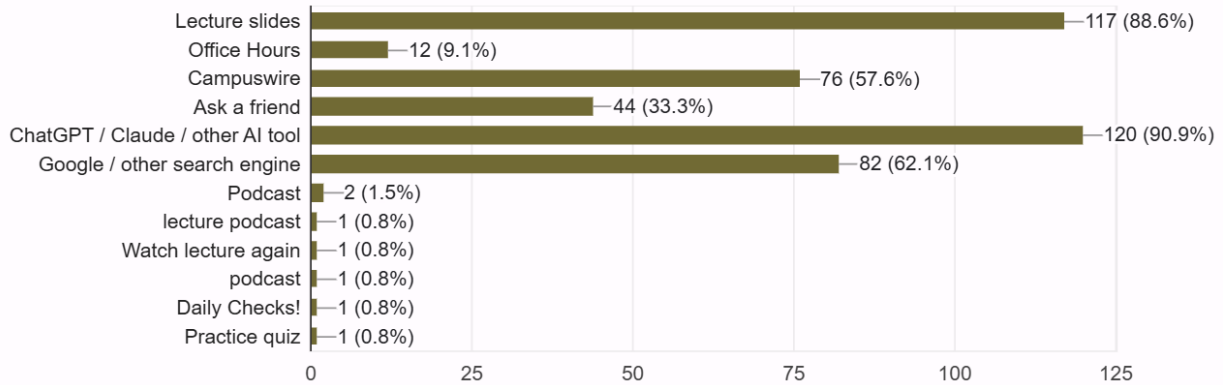
- Good preparation for quiz
- Clarity or too vague sometimes

Mid-Quarter Survey Summary

When you don't understand something in this class, which of the following have you ever turned to thus far? Select all that apply. Note that any of these are perfectly acceptable; there is no judgement being made here.

Copy chart

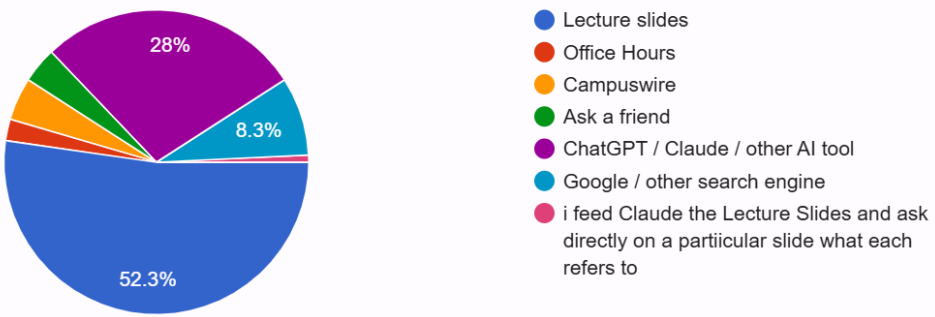
132 responses



When you don't understand something in this class, which of these do you turn to FIRST for help?

Copy chart

132 responses



Mid-Quarter Survey Summary

Other comments

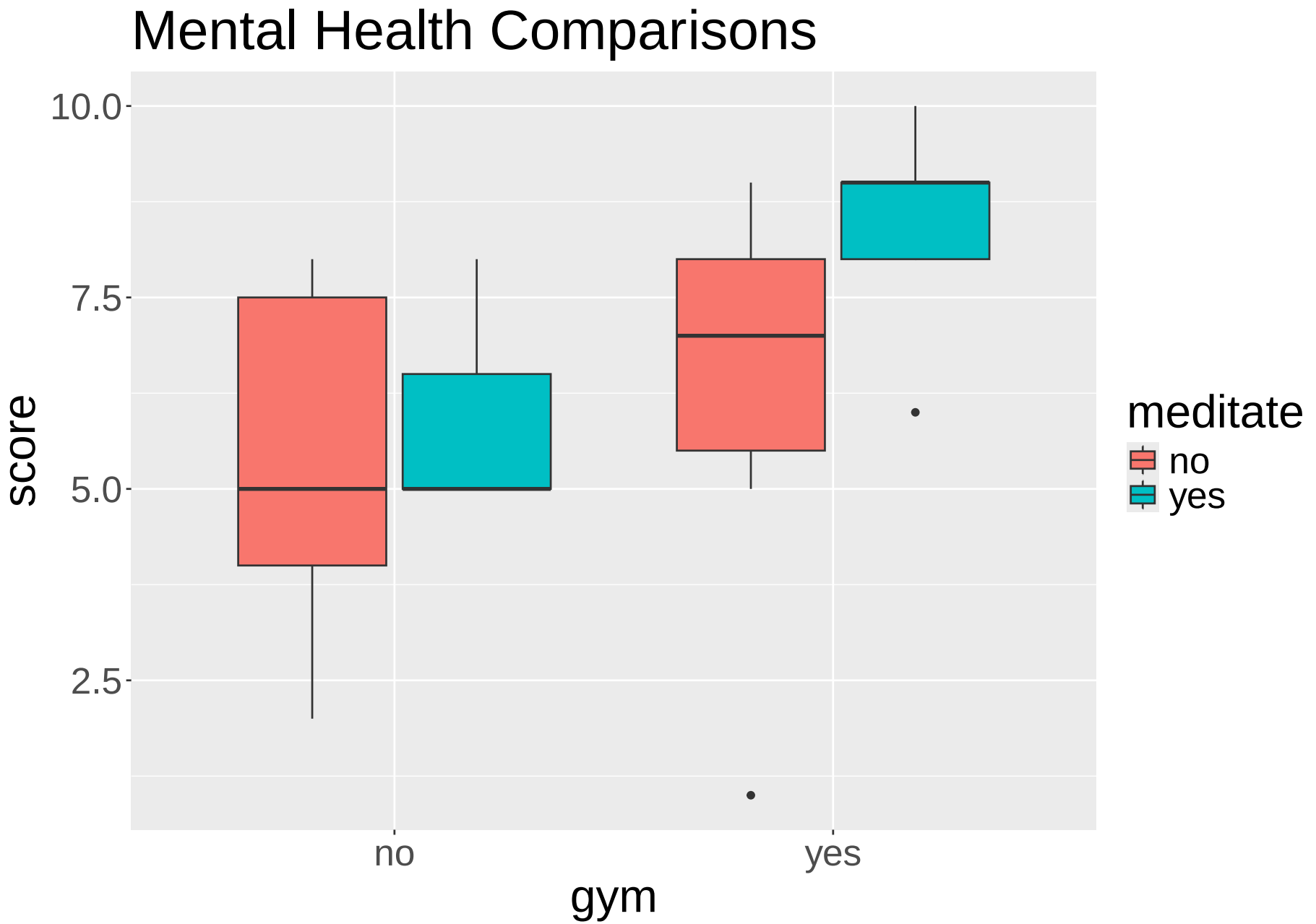
- TONS of good feedback for Quiz 1
  - Most people said it was very fair, good coverage of content
  - Some comments about clarity of questions
- Lots of positive comments about the poker lecture, with feedback saying that memorable examples help in terms of retention of the topics
- Positive comments on course organization
- Some issues with chalk visibility
- "Professor Chi is really great, one of my favorite DSC professors at UCSD so far, right after Eldridge. Great job man"

Mid-Quarter Survey Summary

THANK YOU

We appreciate the time you took to fill out the survey! Everything stated, even if not mentioned here, will be carefully considered for the sake of improving the course both for the remainder of this quarter and in the future.

Recall: Mental Health Self-Experiment



Recall: Mental Health Self-Experiment

The linear model was:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_1 x_2$$

where

- x_1 is `gym` (treated as a 0/1 variable)
- x_2 is `meditate` (treated as a 0/1 variable)
- $x_1 x_2$ is the interaction between `gym` and `meditate`

Last time, we investigated the following null hypotheses:

- $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$
  - (Does any combination of the two factors and their interaction have an impact on the mental health score?)
- $H_0 : \beta_3 = 0$
  - (Is there an interaction between going to the gym and meditation?)

Recall: Mental Health Self-Experiment

The linear model was:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_1 x_2$$

where

- x_1 is `gym` (treated as a 0/1 variable)
- x_2 is `meditate` (treated as a 0/1 variable)
- $x_1 x_2$ is the interaction between `gym` and `meditate`

Last time, we investigated the following null hypotheses:

and finally,

- $H_0 : \beta_2 = \beta_3 = 0$
  - (Is meditation in combination with its interaction statistically significant?)

Recall: Mental Health Self-Experiment

The linear model was:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_1 x_2$$

where

- x_1 is `gym` (treated as a 0/1 variable)
- x_2 is `meditate` (treated as a 0/1 variable)
- $x_1 x_2$ is the interaction between `gym` and `meditate`

What if we actually wanted to ask a different question?

Suppose our interest is in whether going to the gym and meditation have a different impact *from each other*. What would be the H_0 to test for this??

pollev.com

Recall: Mental Health Self-Experiment

What if we actually wanted to ask a different question?

Suppose our interest is in whether going to the gym and meditation have a different impact *from each other*. How do we test for it?

If H_0 is true, then

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_1 x_2$$

would be equivalent to

$$\hat{y} = \hat{\beta}_0^* + \hat{\beta}_1^* (x_1 + x_2) + \hat{\beta}_3^* x_1 x_2$$

Daily Check Question: Why??

pollev.com

Recall: Mental Health Self-Experiment

We want to compare two models, so this calls for a partial $\mathcal{F}$-test!

```
full_model <- lm(score ~ gym + meditate + gym:meditate, data=mh_df)
null_model <- lm(score ~ I(gym + meditate) + gym:meditate, data=mh_df)
anova(null_model, full_model)
```

```
## Analysis of Variance Table
##
## Model 1: score ~ I(gym + meditate) + gym:meditate
## Model 2: score ~ gym + meditate + gym:meditate
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     25 94.357
## 2     24 93.714  1   0.64286 0.1646 0.6885
```

Your Turn #1

Recall again the data from Winter's data collection of my probability theory research project:

score	handwritten	coding
48.2	No	No
NA	No	Yes
46.2	Yes	Yes
NA	Yes	Yes
NA	Yes	No
NA	Yes	Yes
31.3	No	Yes
NA	No	Yes
NA	Yes	Yes

Your Turn #1

In actuality, we wanted to know if coding exercises had a *different* impact than handwritten exercises.

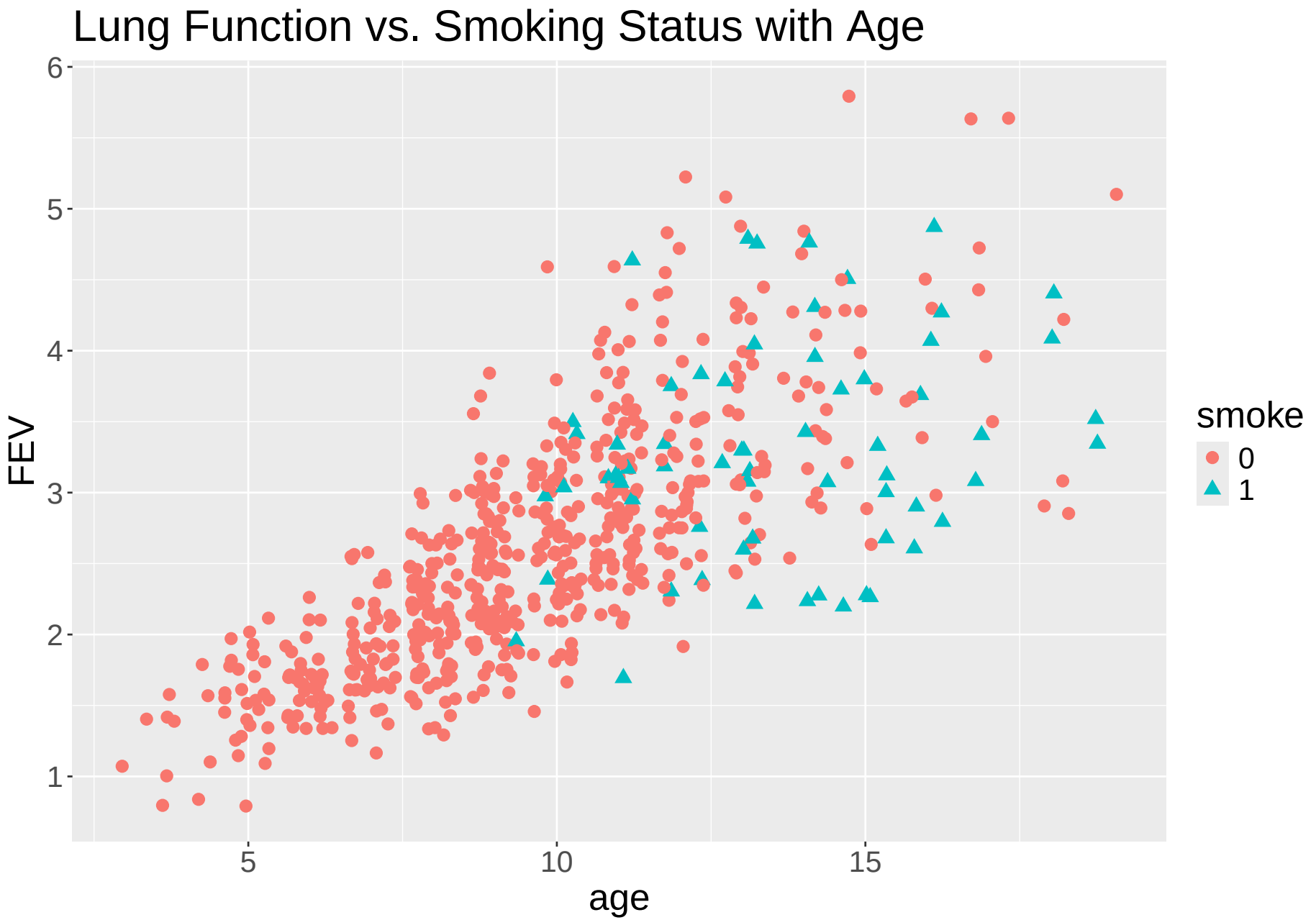
In an R Markdown file:

Load the `learn_prob.csv` dataset into R Markdown again, and then:

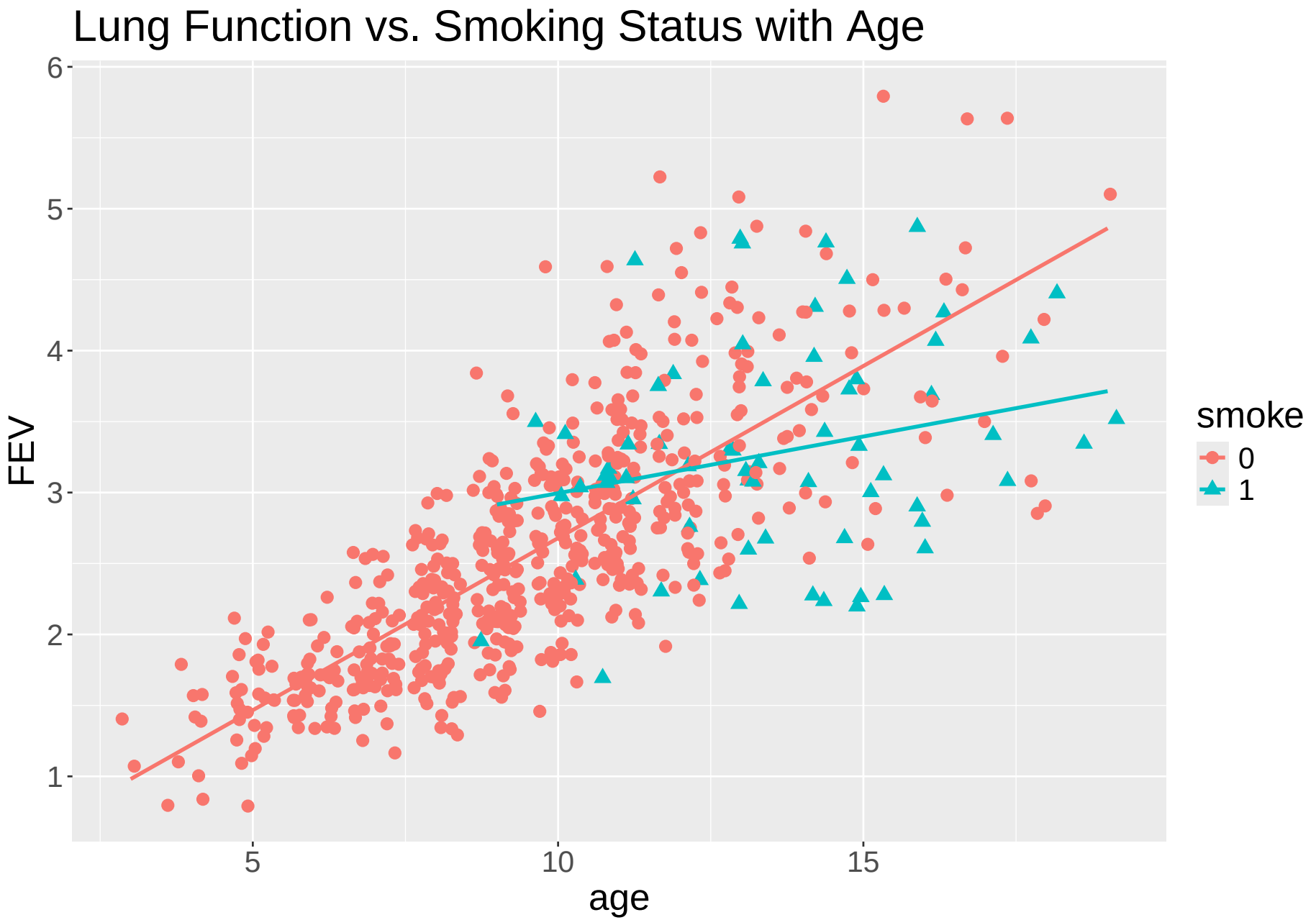
- Run the full model
- Run the appropriate null model
- Get a p-value for this question from a partial $\mathcal{F}$-test

Reminder: in this dataframe, the outcome variable is `score`; the relevant covariates are `handwritten` and `coding`.

Recall: FEV and Smoking



Recall: FEV and Smoking



Recall: FEV and Smoking

Question: what does the fact that the regression lines for each group are not parallel suggest??

Recall: FEV and Smoking

Here was the original model we ran in Lecture #8:

```
model1 <- lm(fev ~ smoke + age, data=FEV)
summary(model1)
```

```
##
## Call:
## lm(formula = fev ~ smoke + age, data = FEV)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.6653 -0.3564 -0.0508  0.3494  2.0894
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.367373   0.081436   4.511 7.65e-06 ***
## smoke       -0.208995   0.080745  -2.588  0.00986 **
## age          0.230605   0.008184  28.176  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5651 on 651 degrees of freedom
## Multiple R-squared:  0.5766, Adjusted R-squared:  0.5753
## F-statistic: 443.3 on 2 and 651 DF,  p-value: < 2.2e-16
```

Recall: FEV and Smoking

Adjusting for age is NOT the same thing as including an *interaction* with age.

In other words, what again is the interpretation of -0.2089949 on the previous slide?

Recall: FEV and Smoking

Interaction model

```
model2 <- lm(fev ~ smoke*age, data=FEV)
summary(model2)
```

```
##
## Call:
## lm(formula = fev ~ smoke * age, data = FEV)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.76645 -0.34947 -0.03364  0.33679  2.05990
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.253396   0.082651   3.066  0.00226 **
## smoke        1.943571   0.414285   4.691 3.31e-06 ***
## age          0.242558   0.008332  29.113  < 2e-16 ***
## smoke:age   -0.162703   0.030738  -5.293 1.65e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5537 on 650 degrees of freedom
## Multiple R-squared:  0.5941, Adjusted R-squared:  0.5922
## F-statistic: 317.1 on 3 and 650 DF,  p-value: < 2.2e-16
```

Recall: FEV and Smoking

Interaction model

Now what are the interpretations of each of the coefficient estimates from this output?

pollev.com

Recall: FEV and Smoking

```
null_model <- lm(fev ~ age, data=FEV)
anova(null_model, model2)
```

```
## Analysis of Variance Table
##
## Model 1: fev ~ age
## Model 2: fev ~ smoke * age
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    652 210.00
## 2    650 199.27  2    10.729 17.498 3.964e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

If the full model is:

$$\widehat{FEV} = \hat{\beta}_0 + \hat{\beta}_1 \cdot smoke + \hat{\beta}_2 \cdot age + \hat{\beta}_3 \cdot smoke \times age,$$

then what is the H_0 for the test shown here? pollev.com

Your Turn #2

Recall the data from HW1 on UCSD women's basketball player Rosa Smith.

- In HW1, we investigated a variety of Smith's statistics, each as a single variable of interest (3 point shooting, assists, minutes played).

- Suppose now that Coach VanDerveer wants you to investigate the potential relationship between the number of minutes that Smith plays in a game, and her field goal shooting percentage in that game
  - Specifically, VanDerveer wants to know if Smith tends to do better on average when she is in the game for longer.

The data are in the tab-delimited file `smith.txt` under this lecture on the course website (slightly modified from before)...

Your Turn #2

Date	Opponent	GS	MIN	FGM/A	fgp	TFG/A	tfgp
11/07/25	Denver	NA	16	2-3	0.667	1-2	0.500
11/12/25	Sacramento St.	NA	10	1-1	1.000	1-1	1.000
11/16/25	at San Francisco	NA	30	4-7	0.571	1-2	0.500
11/22/25	Air Force	*	30	2-5	0.400	2-4	0.500
11/24/25	Occidental	*	23	3-7	0.429	1-4	0.250
11/28/25	at Washington	*	40	8-14	0.571	2-6	0.333
11/30/25	at Portland St.	*	34	7-15	0.467	2-5	0.400
12/06/25	Long Beach St.	*	32	11-18	0.611	4-10	0.400

There are more rows and columns that this. Specifically,

- Each row represents a single game
- `fgp` is her 3-point shooting percentage in that game
- `MIN` is the number of minutes she played in that game
- `AST` is the number of assists she had in that game
- `TO` is the number of turnovers she had in that game

Your Turn #2

After discussion with Coach, you decide that the following full model is appropriate:

$$\widehat{\texttt{fgp}} = \hat{\beta}_0 + \hat{\beta}_1 \texttt{MIN} + \hat{\beta}_2 \texttt{AST} + \hat{\beta}_3 \texttt{TO} + \hat{\beta}_4 \texttt{MIN} \times \texttt{AST}$$

where, again, our primary covariate of interest is `MIN`.

Answer/do the following:

- What would be the rationale for choosing this as the model for inference, in terms of what we think the role of each of these variables is?
- Run the full model
- Provide brief interpretations of $\hat{\beta}_1$ and $\hat{\beta}_4$ using the output
- Then run the null model for the question of interest, and do the appropriate partial $\mathcal{F}$-test, reporting your p-value.

Recap and Looking Ahead

Recap

- Different statistical comparisons (e.g. $H_0 : \beta_1 = \beta_2$) can be made by properly specifying the null model and performing a partial $\mathcal{F}$-test
- Interactions can be modeled between binary, categorical, and quantitative variables

Looking Ahead

- Transformations
- Caution with Model Selection

Today's Daily Check

- Answer to the question on Slide 22
- The two Your Turns