DSC 152:
Applied Statistical Data Analysis

Lecture #8
Multiple Linear Regression

Thursday, April 23
Spring Quarter 2026
Peter Chi

Example #1: FEV

Researchers are studying the impact of smoking on lung function:

- Data were collected at a children's hospital in Boston in the 1970s
- Study subjects ranged in age from 3 to 19 years old
- The outcome variable of interest is "forced expiratory volume" (FEV)
- The predictor variable of interest is their smoking status, which was simply coded as:
  - 0 if the child does not smoke
  - 1 if the child self-reports that they "smoke regularly"

Example #1: FEV

FEV in children from ages 3-19

age	fev	height	smoke
9	1.708	57.0	0
8	1.724	67.5	0
7	1.720	54.5	0
9	1.558	53.0	0
9	1.895	57.0	0
8	2.336	61.0	0
6	1.919	58.0	0
6	1.415	56.0	0

- This continues on; the total sample size is 654
- `fev` is measured in liters, and represents the amount of air blown out in 1 second

Example #1: FEV

FEV in children from ages 3-19: The naive analysis

```
t.test(fev ~ smoke, data=FEV)
```

```
##
##  Welch Two Sample t-test
##
## data:  fev by smoke
## t = -7.1496, df = 83.273, p-value = 3.074e-10
## alternative hypothesis: true difference in means between group 0 and gr
## 95 percent confidence interval:
##  -0.9084253 -0.5130126
## sample estimates:
## mean in group 0 mean in group 1
##        2.566143        3.276862
```

Example #1: FEV

FEV in children from ages 3-19: The naive analysis

```
t.test(fev ~ smoke, data=FEV)
```

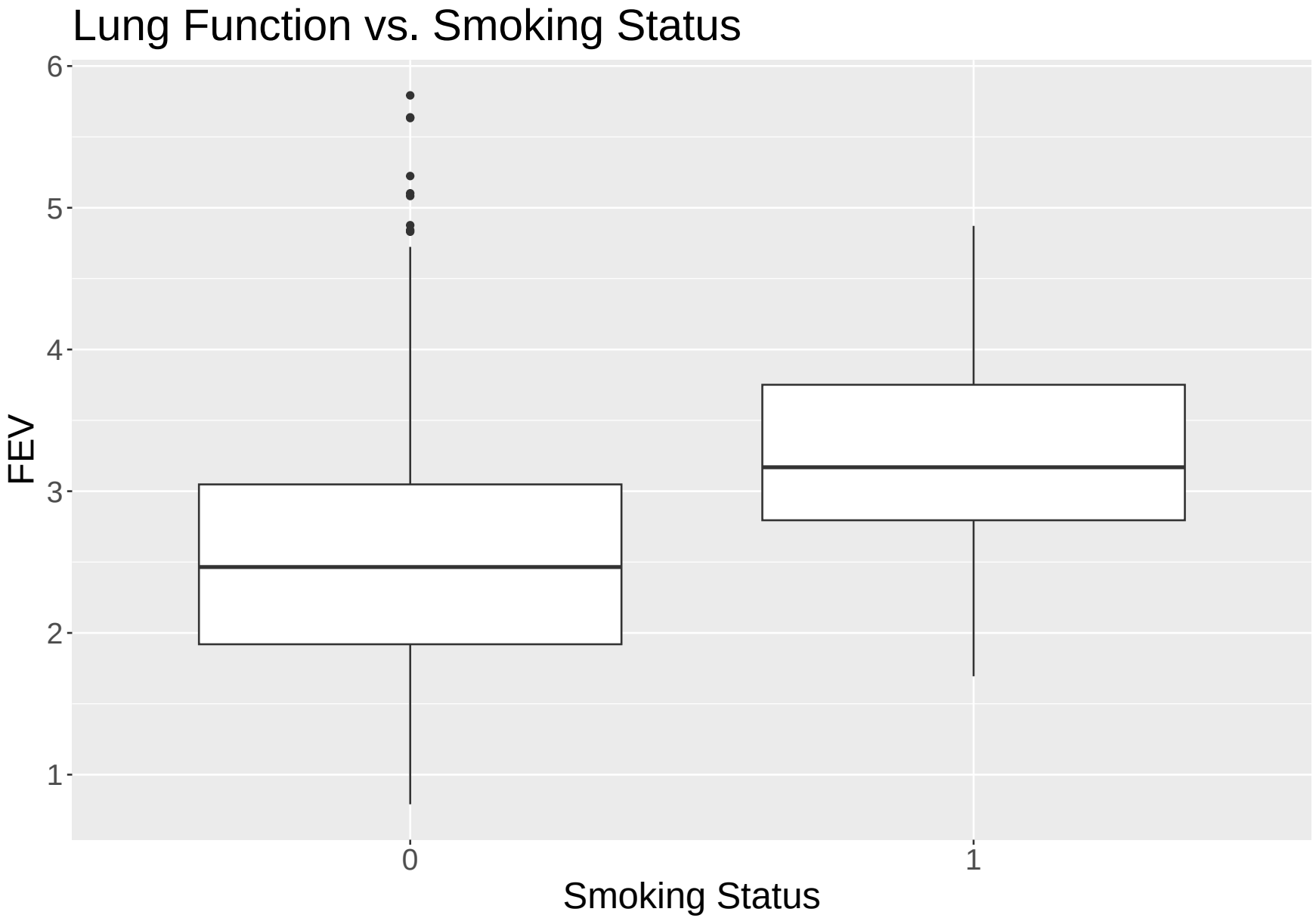
```
##
##  Welch Two Sample t-test
##
## data:  fev by smoke
## t = -7.1496, df = 83.273, p-value = 3.074e-10
## alternative hypothesis: true difference in means between group 0 and gr
## 95 percent confidence interval:
##  -0.9084253 -0.5130126
## sample estimates:
## mean in group 0 mean in group 1
##        2.566143        3.276862
```

What's wrong with this?

Why do the smokers have a higher FEV??

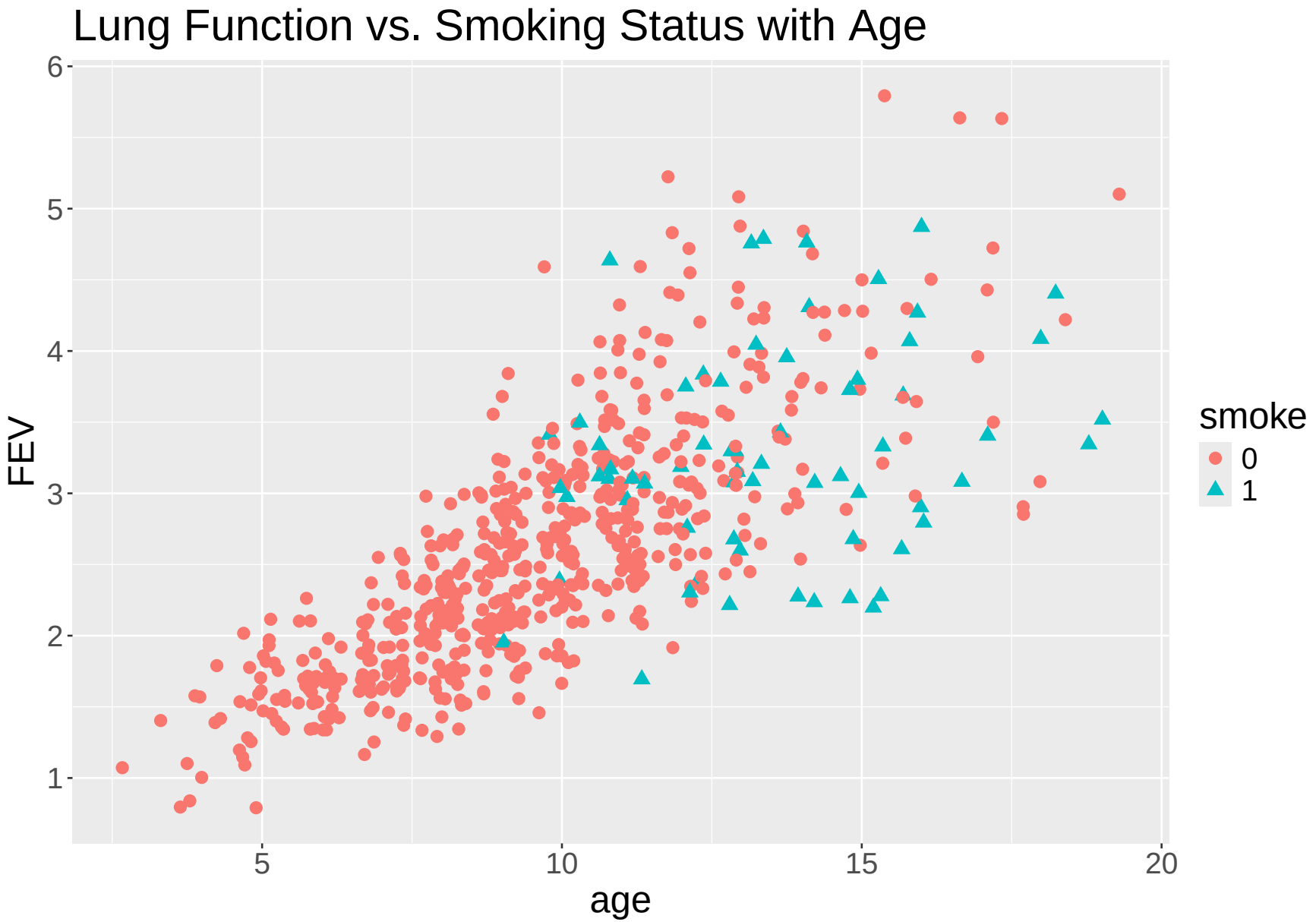
Example #1: FEV

FEV in children from ages 3-19: The naive analysis



Example #1: FEV

Let's do a little better



Example #1: FEV

Let's do a little better

For your reference, here is the code that made the plot on the previous slide:

```
ggplot(data=FEV, mapping = aes(x = age, y=fev)) +
  geom_jitter(aes(shape=factor(smoke),
                  color=factor(smoke)),
              size=3) +
  theme(text=element_text(size=20)) +
  labs(y="FEV", title="Lung Function vs. Smoking Status
       with Age", shape="smoke", color="smoke")
```

Example #1: FEV

So let's start to fix the analysis. The framework of linear models will be useful.

- First we note that the original t-Test could have been performed as a linear model:

$$\widehat{FEV} = \hat{\beta}_0 + \hat{\beta}_1 \cdot smoke$$

What would be the interpretations of $\hat{\beta}_0$ and $\hat{\beta}_1$ in this context?
https://pollev.com/chi

Example #1: FEV

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

```
##
## Call:
## lm(formula = fev ~ smoke, data = FEV)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.7751 -0.6339 -0.1021  0.4804  3.2269
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.56614    0.03466  74.037  < 2e-16 ***
## smoke        0.71072    0.10994   6.464 1.99e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8412 on 652 degrees of freedom
## Multiple R-squared:  0.06023,    Adjusted R-squared:  0.05879
## F-statistic: 41.79 on 1 and 652 DF,  p-value: 1.993e-10
```

Example #1: FEV

Sidenote: is it really identical?

Note that the p-value from the two-sample t-Test was:

```
t.test(fev ~ smoke, data=FEV)$p.value
```

```
## [1] 3.073813e-10
```

and from the linear model was:

```
summary(model1)$coefficients[2,4]
```

```
## [1] 1.992846e-10
```

So they aren't exactly equal. But...

Example #1: FEV

Sidenote: is it really identical?

Recall that linear regression has the condition of equal variances, whereas the version of the t-Test that we do doesn't. But what if we do the equal variance version of the t-Test?

```
t.test(fev ~ smoke, data=FEV, var.equal=TRUE)$p.value
```

```
## [1] 1.992846e-10
```

and from the linear model again:

```
summary(model1)$coefficients[2,4]
```

```
## [1] 1.992846e-10
```

So the point is that a t-Test can be formulated as a linear regression question, with a single binary predictor variable.

Example #1: FEV

Then, under this linear model formulation, we can add additional predictor variables!

$$\widehat{FEV} = \hat{\beta}_0 + \hat{\beta}_1 \cdot smoke$$

becomes

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

How does this help us?

Let's work backwards. First, running the analysis in R is easy:

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

Example #1: FEV

And instead of just printing the summary (like I did on Slide 9) we can easily clean up the presentation of the output using the `kable` function:

```
library(knitr)
kable(summary(model2)$coefficients)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.3673730	0.0814357	4.511203	0.0000076
smoke	-0.2089949	0.0807453	-2.588321	0.0098598
age	0.2306046	0.0081844	28.176209	0.0000000

A couple of questions. . .

https://pollev.com/chi

Example #1: FEV

Two Daily Check Questions:

- Which p-value(s) from this output do we care about? What is H_0 for each one that we do care about?

- What is the interpretation of the output value of -0.2089949, in the context of this scenario?

Example #1: FEV

(more space for notes if needed)

Example #1: FEV

Important note

The decision of whether to include `age` in the model should NOT be driven by whether its p-value is statistically significant.

- It is a common approach in data science (and even in applied statistics) to do that – that is, to build a model based on p-values (or any other metric such as MSE or R^2).
  - This is OK if the end goal from that dataset is simply to build that model, or to do prediction.
  - It is NOT ok if the end goal from that dataset is statistical inference.

Simply put, to do this is asking too much of your data. We will explain this concept in more detail in Lecture #14.

Example #1: FEV

Conclusions

We find that there is statistically significant evidence at the $\alpha = 0.05$ level that smoking status is associated with lung function, while adjusting for age (p=0.0099). Specifically, we estimate that smoking is associated with an average decrease in FEV of 0.209 liters when comparing children of the same age, with 95% CI: (-0.3675, -0.0504).

Example #1: FEV

Conclusions

We find that there is statistically significant evidence at the $\alpha = 0.05$ level that smoking status is associated with lung function, while adjusting for age (p=0.0099). Specifically, we estimate that smoking is associated with an average decrease in FEV of 0.209 liters when comparing children of the same age, with 95% CI: (-0.3675, -0.0504).

Note #1: the text above was written with inline R code

```
We find that there is statistically significant evidence at the $\alpha=0.05$
level that smoking status is associated with lung function, while adjusting
for age (p=`r round(summary(model2)$coefficients[2,4], 4)`). Specifically, we
estimate that smoking is associated with an average decrease in FEV of `r
abs(round(summary(model2)$coefficients[2,1], 4))` liters when comparing
children of the same age, with 95\% CI: (`r round(confint(model2,
'smoke')[1], 4)`, `r round(confint(model2, 'smoke')[2], 4)`).
```

Example #1: FEV

Conclusions

We find that there is statistically significant evidence at the $\alpha = 0.05$ level that smoking status is associated with lung function, while adjusting for age (p=0.0099). Specifically, we estimate that smoking is associated with an average decrease in FEV of 0.209 liters when comparing children of the same age, with 95% CI: (-0.3675, -0.0504).

Note #2: what are some key features of this written conclusion?

- "while adjusting for age"
- "smoking is associated with an average decrease in FEV of..."
  - Why wouldn't it be correct to say "smoking decreases FEV by..."?

Example #1: FEV

Daily Check Questions

- What is the role of age in these data?
- If we run an analysis that adjusts for age (as we just did), why can we still not infer a causal relationship between smoking and lung function?

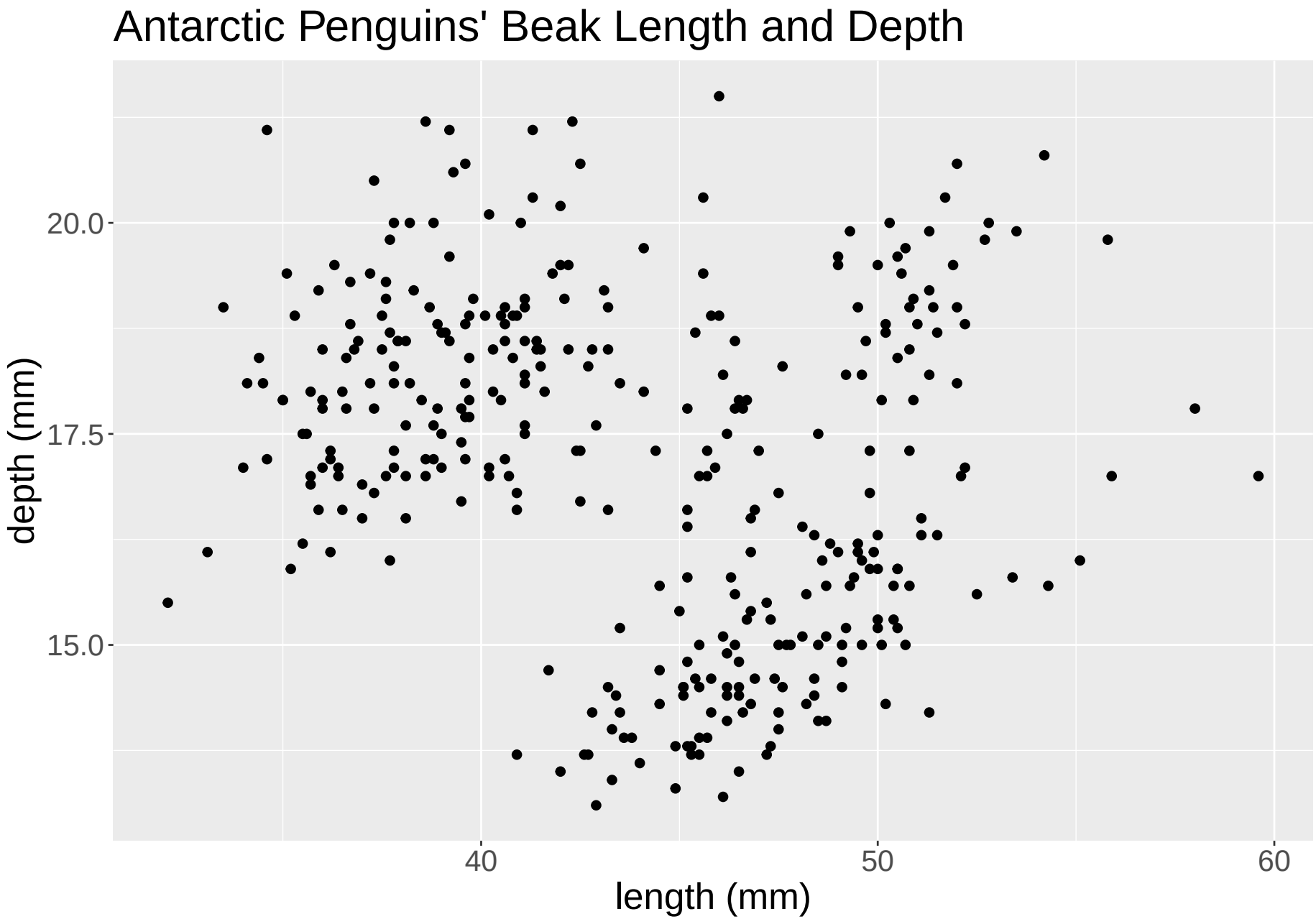
Example #2: Penguin Beak Dimensions

An ecologist is studying penguin beaks, and for reasons of conservation and understanding the penguins' dietary habits, is specifically interested in the relationship between:

- The depth of a penguin's beak
- The length of the penguin's beak

Data were collected at Palmer Station in Antarctica, with a total sample size of 344 penguins.

Example #2: Penguin Beak Dimensions

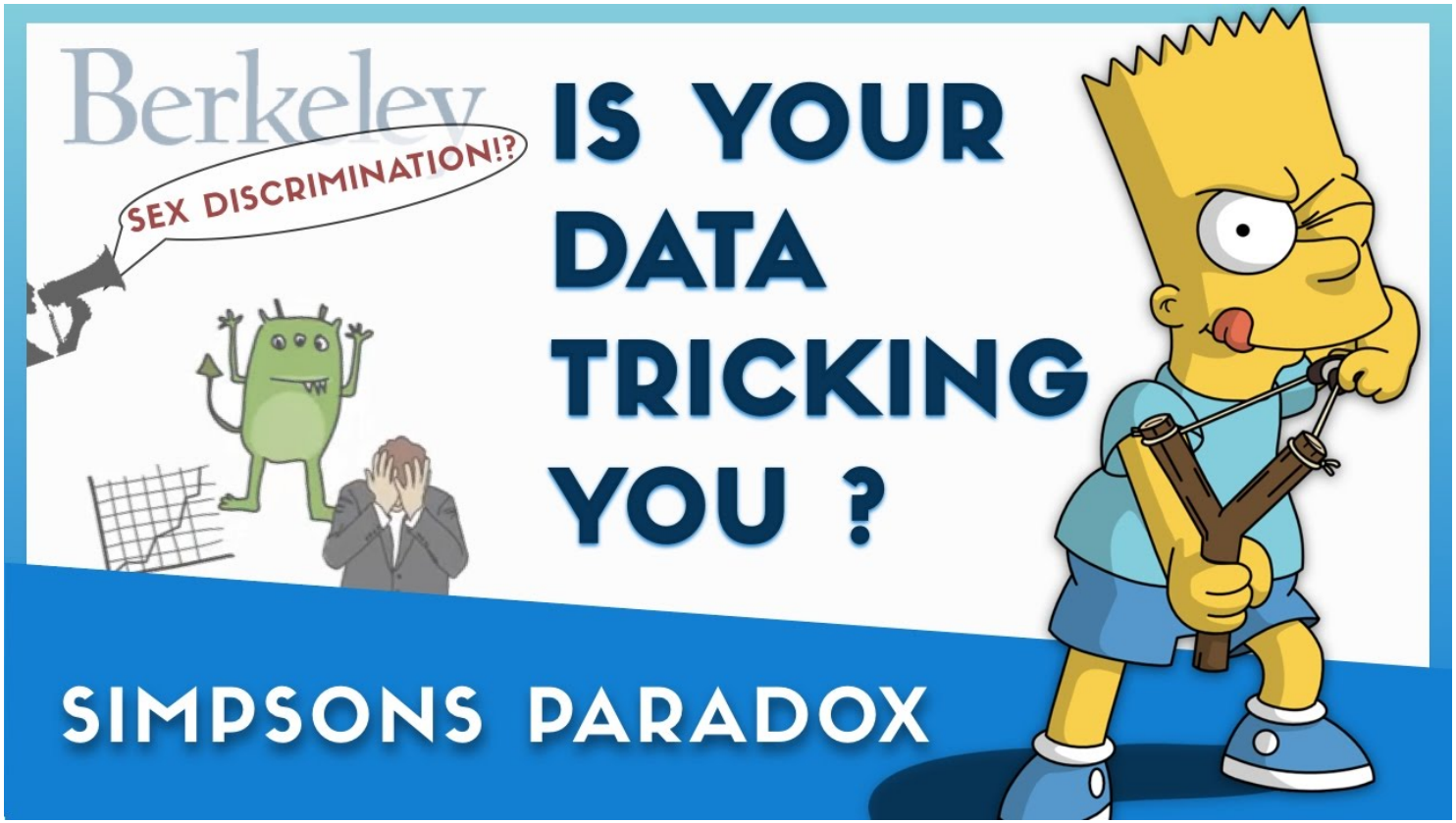


Example #2: Penguin Beak Dimensions

What do we observe?

It's kind of weird that there appears to be a negative relationship between beak length and beak depth

Is this another example of...



???

Example #2: Penguin Beak Dimensions

Yes it is! The confounding variable here is `species`:

```
table(penguins$species)
```

```
##
##    Adelie Chinstrap    Gentoo
##       152        68       124
```

(these are just the counts of how many of each species there are in the dataset)

Your Turn

- Load the dataset into your R Markdown document:
  - First, in the **R Console**, type `install.packages("palmerpenguins")` (do NOT put this in your Rmd file!!)
  - Then, **in your Rmd file**, put `library(palmerpenguins)` in a code chunk (you can also do this in your R Console if you want to view and work with the dataframe interactively)
  - The dataframe will then be loaded as `penguins`

- Create a scatterplot of the data as in Slide 21, but add the following:
  - Color the points by `species` (see code in Slide 7) – but note that `species` is already coded as a `factor` variable so you do not need to include the `factor()` part
  - Add regression lines for each group (you can do this by adding `+ geom_smooth(method="lm", se=FALSE)` to your plotting code)

Your Turn

- Run linear models:
  - One with just `y=depth`, `x=length`
  - Then add `species` as an additional covariate
  - Include explanations of the following:
    - Why the second model is probably the correct one
    - Which p-value(s) we care about from that output
    - Interpretation(s) of the regression coefficients that we care about
    - Your conclusions

Summary of today's Daily Check:

1. Answers to the questions on Slide 14
2. Answers to the questions on Slide 19
3. All of the Your Turn

Recap and Looking Ahead

Recap

- In the context of statistical inference, the purpose of multiple linear regression models is to adjust for potential confounders
- Here, we explored two examples:
  - Binary primary covariate with quantitative confounder
  - Quantitative primary covariate with categorical confounder
- We only care about the p-value for the primary covariate!

Looking Ahead

- What if our primary covariate is categorical (with more than two categories, so it's not just binary)? How do we get a proper p-value for inference in that situation?
- Model diagnostics for checking required conditions for valid inference