DSC 152:
Applied Statistical Data Analysis

Lecture #9
Inference with Categorical Predictor Variables

Tuesday, April 28
Spring Quarter 2026
Peter Chi

Last time

Example 1: Smoking and FEV

- The primary covariate of interest was binary (smoking status)
- The other covariate (the potential confounder) was quantitative (age)

Example 2: Penguin Beak Dimensions

- The primary covariate of interest was quantitative (beak length)
- The other covariate (the potential confounder) was categorical (species)

In neither case did we have to worry about how to do inference (that is, get a p-value) for a categorical variable (with more than 2 categories).

Example: Mental Health Dogs

ANTHROZOÖS
2022, VOL. 35, NO. 1, 1–22
https://doi.org/10.1080/08927936.2021.1944558

Routledge
Taylor & Francis Group

The Importance of Client–Canine Contact in Canine-Assisted Interventions: A Randomized Controlled Trial

John-Tyler Binfet[a], Freya L. L. Green[a], and Zakary A. Draper[b]

[a]Faculty of Education, University of British Columbia, Kelowna, BC, Canada; [b]Department of Psychology, University of British Columbia, Kelowna, BC, Canada

ABSTRACT

Researchers have claimed that canine-assisted interventions (CAIs) contribute significantly to bolstering participants' wellbeing, yet the mechanisms within interactions have received little empirical attention. The aim of this study was to assess the impact of client–canine contact on wellbeing outcomes in a sample of 284 undergraduate college students (77% female; 21% male, 2% non-binary). Participants self-selected to participate and were randomly assigned to one of two canine interaction treatment conditions (touch or no touch) or to a handler-only condition with no therapy dog present. To assess self-reports of wellbeing, measures of flourishing, positive and negative affect, social connectedness, happiness, integration into the campus

KEYWORDS

Canine-assisted intervention; human–animal interaction; stress reduction; university student; wellbeing

Example: Mental Health Dogs

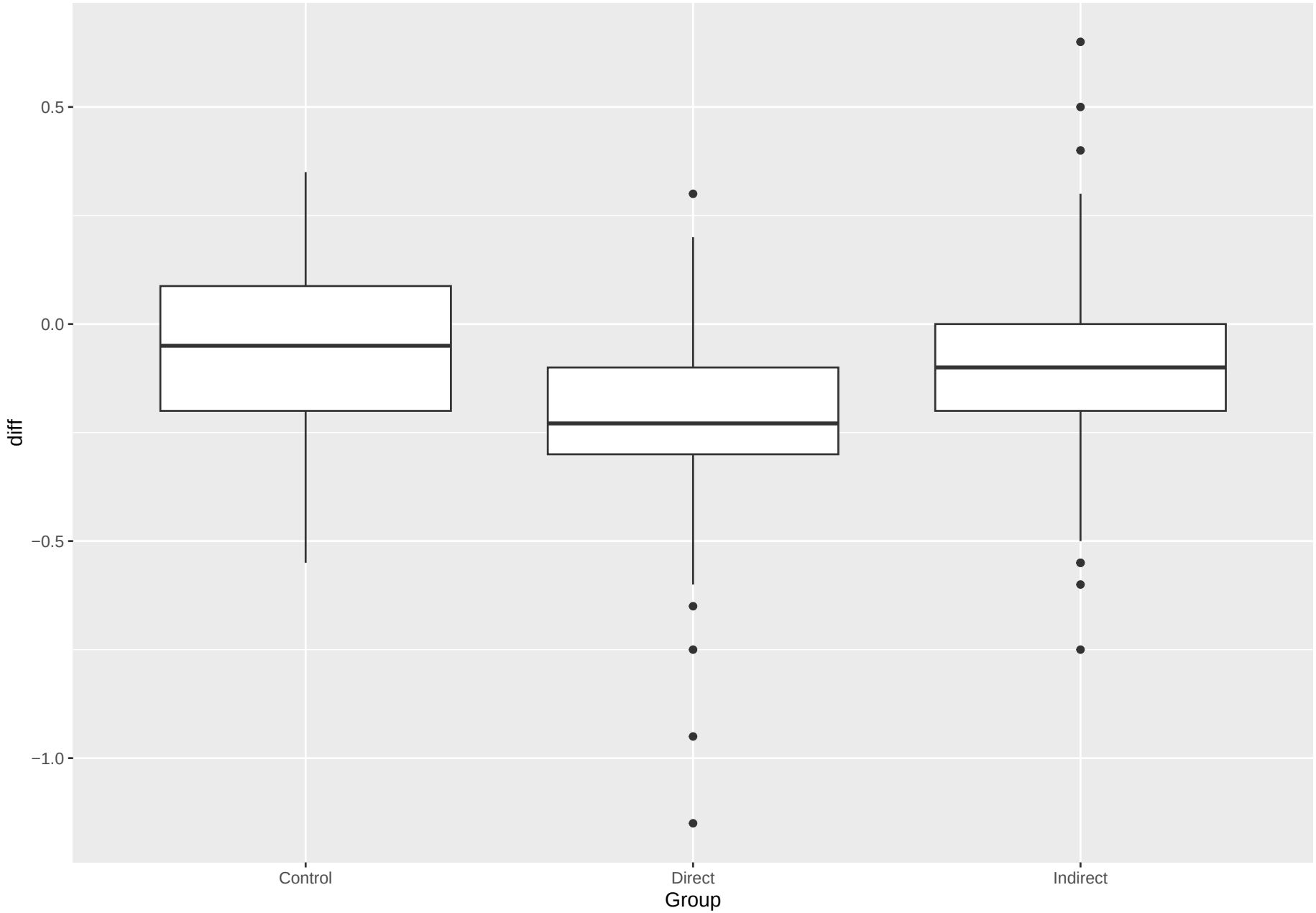
Study setup

There are three treatment groups:

- Control group
  - Go into treatment room, but there is no dog (only the handler)
- Indirect contact with therapy dog
  - Go into the treatment room, and there is a dog and its handler (but cannot pet the dog)
- Direct contact with therapy dog
  - Go into the treatment room, and there is a dog and its handler, can pet the dog

The outcome of interest is a post-pre difference in a loneliness score (negative values indicate that the person's loneliness went DOWN after the treatment, which is what we want!)

Example: Mental Health Dogs



Example: Mental Health Dogs

Now, how do we analyze these data statistically?

We have:

- A quantitative outcome variable (loneliness score)
- A categorical primary covariate (treatment group)

What's our null hypothesis?? And our alternative hypothesis??

Example: Mental Health Dogs

So, we can still run a linear model:

```
groups_model <- lm(diff ~ Group, data=lonely)
summary(groups_model)
```

```
##
## Call:
## lm(formula = diff ~ Group, data = lonely)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.92554 -0.12377  0.00991  0.12446  0.74260
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -0.05991    0.02253  -2.659  0.00828 **
## GroupDirect    -0.16455    0.03178  -5.178 4.28e-07 ***
## GroupIndirect -0.03269    0.03178  -1.029  0.30444
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2184 on 281 degrees of freedom
## Multiple R-squared:  0.09684,    Adjusted R-squared:  0.09041
## F-statistic: 15.06 on 2 and 281 DF,  p-value: 6.097e-07
```

Example: Mental Health Dogs

But wait, what model did this just fit, and what p-value(s) do we care about?

$$\widehat{\texttt{diff}} = \hat{\beta}_0 + \hat{\beta}_1 \cdot \texttt{Direct} + \hat{\beta}_2 \cdot \texttt{Indirect}$$

where

- `Direct` is equal to:
  - 1 if the person was in the Direct treatment group
  - 0 if they were not
- `Indirect` is equal to:
  - 1 if the person was in the Indirect treatment group
  - 0 if they were not

Example: Mental Health Dogs

Daily Check Question #1 (in an R Markdown document)

According to the model output on Slide 7, what are the estimated means of each treatment group?

First answer it at pollev.com, and then put in your Rmd file once you know that you are right.

Example: Mental Health Dogs

Now, as tempting as it might be to think otherwise, we do not ONLY care about the impact of being in the `Direct` group.

Recall what our null and alternative hypotheses were on Slide 6. What do those translate to in terms of the βs?

Example: Mental Health Dogs

```
null_model <- lm(diff ~ 1, data=lonely)
anova(null_model, groups_model)
```

```
## Analysis of Variance Table
##
## Model 1: diff ~ 1
## Model 2: diff ~ Group
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    283 14.844
## 2    281 13.407  2    1.4375 15.065 6.097e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

How does this work?

Example: Mental Health Dogs

RSS has the same definition that it did from Lecture 7:

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

(which, again, may have been called $R_{sq}(w_0, w_1)$ in DSC 40A)

The `null_model` from the previous slide, in this case, is the model with only the intercept: $\hat{y} = \hat{\beta}_0$.

...what exactly does this indicate $\hat{y}$ to be?

Example: Mental Health Dogs

Your Turn #1 (for Daily Check)

In an R Markdown file,

- Load in the data, `dog_data_lonely.csv`.
- Manually code a calculation of RSS for the null model, based on what we said $\hat{y}$ is from the previous slide
- Then, run the `groups_model` from Slide 7 and store the regression output
  - Use the coefficient estimates from the model to manually calculate RSS for this model
- Comment briefly on what you observe, and how it compares to the values in the `anova` function output

Example: Mental Health Dogs

Now, where exactly does the p-value come from?

If H_0 is actually true, what do we expect to be the case regarding RSS_{full} vs. RSS_{null}?

pollev.com

Daily Check Question

Explain each of these answers.

Example: Mental Health Dogs

More space for notes if needed

Example: Mental Health Dogs

```
## Analysis of Variance Table
##
## Model 1: diff ~ 1
## Model 2: diff ~ Group
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    283 14.844
## 2    281 13.407  2    1.4375 15.065 6.097e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Example: Mental Health Dogs

```
## Analysis of Variance Table
##
## Model 1: diff ~ 1
## Model 2: diff ~ Group
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    283 14.844
## 2    281 13.407  2    1.4375 15.065 6.097e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The test statistic for this output is:

$$\mathcal{F} = \frac{(RSS_{null} - RSS_{full})/p}{RSS_{full}/(n-k)}$$

where

- n is the total sample size
- k is the total number of coefficients in the full model
- p is the difference in the number of coefficients between the two models

Example: Mental Health Dogs

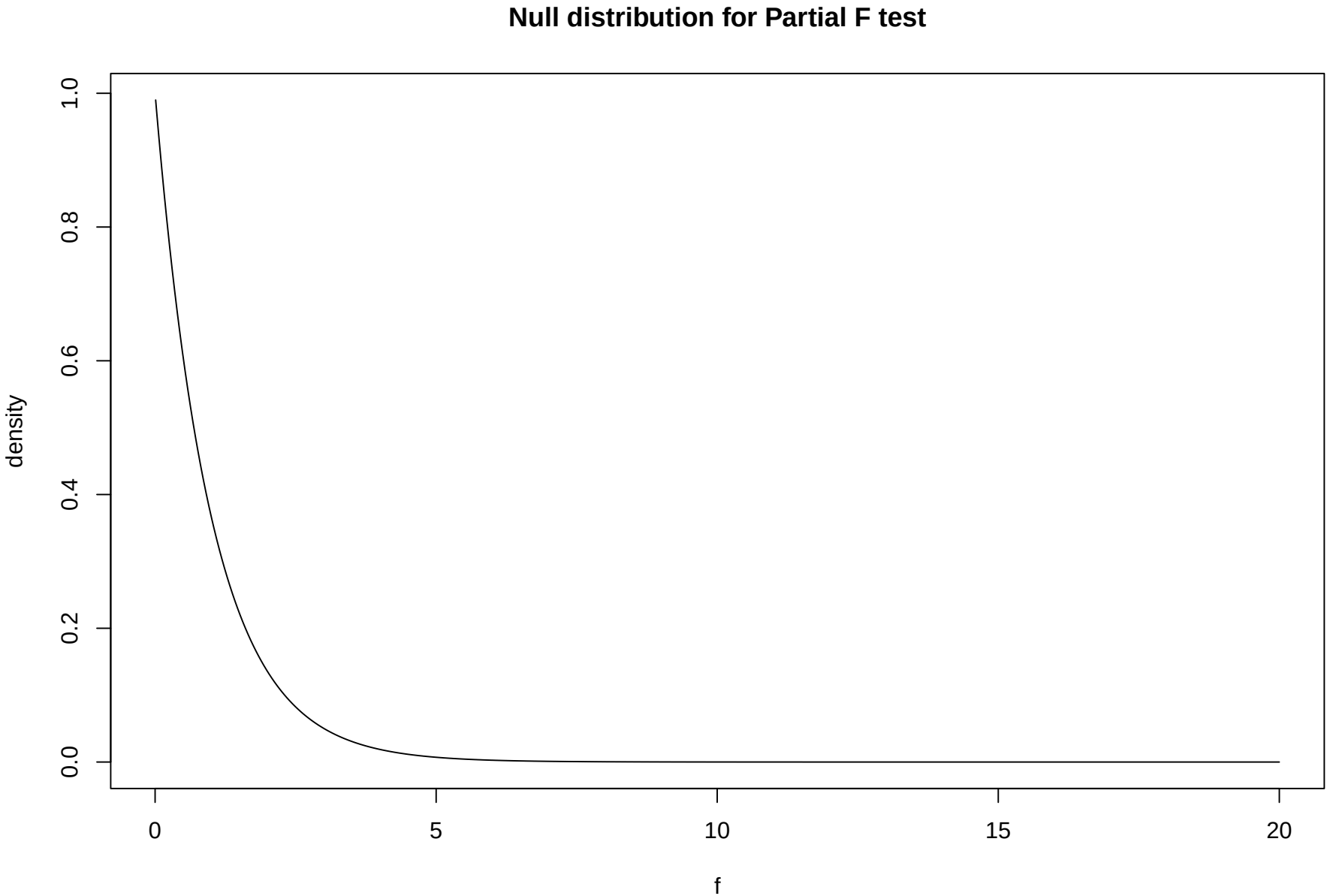
```
## Analysis of Variance Table
##
## Model 1: diff ~ 1
## Model 2: diff ~ Group
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    283 14.844
## 2    281 13.407  2    1.4375 15.065 6.097e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Now take the RSS values that we calculated in Your Turn #1:

Do the calculation of the F statistic with R code, and add it to the end of Your Turn #1.

This test is called a partial-$\mathcal{F}$ test, and its null distribution is the $\mathcal{F}$ distribution...

$\mathcal{F}$ distribution



Notice that 15.065 is quite extreme. . .

Example: Mental Health Dogs

So, we get a tiny p-value and we reject H_0.

Next Question: adding covariates

What if we want to investigate the impact of the treatment group, while adjusting for age?

That is, we just fit this model:

$$\widehat{\texttt{diff}} = \hat{\beta}_0 + \hat{\beta}_1 \cdot \texttt{Direct} + \hat{\beta}_2 \cdot \texttt{Indirect}$$

but now we want to fit this model:

$$\widehat{\texttt{diff}} = \hat{\beta}_0 + \hat{\beta}_1 \cdot \texttt{Direct} + \hat{\beta}_2 \cdot \texttt{Indirect} + \hat{\beta}_3 \cdot \texttt{age}$$

Fitting the model is easy...

Example: Mental Health Dogs

```r
groups_age_model <- lm(diff ~ Group + Age_Yrs, data=lonely)
summary(groups_age_model)
```

```
##
## Call:
## lm(formula = diff ~ Group + Age_Yrs, data = lonely)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.9265 -0.1225  0.0075  0.1278  0.7467
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -0.146810   0.111222  -1.320    0.188
## GroupDirect   -0.163772   0.031813  -5.148 4.97e-07 ***
## GroupIndirect -0.032696   0.031798  -1.028    0.305
## Age_Yrs        0.004357   0.005460   0.798    0.426
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2186 on 280 degrees of freedom
## Multiple R-squared:  0.09889,    Adjusted R-squared:  0.08923
## F-statistic: 10.24 on 3 and 280 DF,  p-value: 2.027e-06
```

Example: Mental Health Dogs

But now, again, how do we get the p-value that we actually care about?

Partial-$\mathcal{F}$ test again

We want to test the entire categorical variable, meaning that like before,

$$H_0 : \beta_1 = \beta_2 = 0$$

If this is our null hypothesis, then what is our null model?

Example: Mental Health Dogs

Your Turn #2

- Run the null model as described on the previous slide (and store it as something)
- Run the partial-$\mathcal{F}$ test using the `anova` function as we saw previously
- Report your p-value and conclusions, at $\alpha = 0.0$.

Recap and Looking Ahead

Recap

- If our primary covariate of interest is categorical, then the p-value of interest must be obtained via a partial-$\mathcal{F}$ test
  - This is because the null hypothesis is that all group means are equal to each other
- The null model for the partial-$\mathcal{F}$ test is the one without that categorical variable in it (but everything else from the full model remains in)

Recap and Looking Ahead

Summary of today's Daily Check

- Answer to the question on Slide 9
- Your Turn #1 on Slide 13
- Answer to the question on Slide 14
- Your Turn #2 on Slide 22

Looking Ahead

Model Diagnostics!