Review Session 7

Sophie Hill

11/03/2022

Agenda

1. Feedback: Common mistakes on Quiz 2
2. Review: Interpreting regression coefficients

Quiz 2

Overall, great job!

The trickiest questions were:

- **Q3** (p = 0.06 means no chance of a relationship?)
- **Q4** (types of measurement error)
- **Q5** (attenuation bias)

Quiz 2: Q3

“Since the coefficient estimate does not have a p-value of p < 0.05, there is no chance that there is a meaningful relationship between these two variables.” Is this correct?

Quiz 2: Q3

“Since the coefficient estimate does not have a p-value of p < 0.05, there is no chance that there is a meaningful relationship between these two variables.” Is this correct?

Answer: This is **not** correct.

Why?

- 0.05 cutoff is arbitrary, so p = 0.06 is not substantively different from p = 0.04
- Relationship could be non-linear

Quiz 2: Q4

“According to the manufacturer of the test kits, whether or not a test kit returns a false negative does not depend on how sick the person is or what strain of Covid they might have or any other individual characteristic. Sometimes the kits simply fail.”

Quiz 2: Q4

"According to the manufacturer of the test kits, whether or not a test kit returns a false negative does not depend on how sick the person is or what strain of Covid they might have or any other individual characteristic. Sometimes the kits simply fail."

Answer: This is **non-systematic measurement error.**

Why?

The error is not correlated with the true values of any of our variables, so it is non-systematic.

Quiz 2: Q5

“Despite the inaccuracy of the test kits, the researcher has correctly concluded, for the data collected, that greater adherence to wearing masks correlates with lower COVID positivity rates afterwards.” True or false?

Quiz 2: Q5

“Despite the inaccuracy of the test kits, the researcher has correctly concluded, for the data collected, that greater adherence to wearing masks correlates with lower COVID positivity rates afterwards.” True or false?

Answer: This is **true**.

The presence of non-systematic measurement error weakens the correlation between two variables by adding “noise”. So the negative correlation between wearing masks and the true COVID positivity rates is even stronger than the observed relationship.

Interpreting regression coefficients

- Different types of **outcome** variable (continuous, log-scale, binary)
- Different types of **predictor** (continuous, log-scale, binary, categorical, squared)
- **Simple** regression vs **multiple** regression
- **Interaction** terms

Example

Let’s work with the CCES survey data we used in Problem Set 2.

For simplicity, let’s focus on the 2020 data, subset to Democratic and Republican voters, and create a new binary variable indicating whether the respondent voted Democratic (1) or Republican (0).

```
cces20 <- cces_pres %>%
  filter(year == 2020,
         voted_pres_party %in% c("Democratic", "Republican")) %>%
  mutate(vote_dem = case_when(voted_pres_party=="Democratic" ~ 1,
                              voted_pres_party=="Republican" ~ 0))
```

Intercept-only model

Let’s start by estimating the simplest model of all: a linear model including *just* the intercept term.

We can do this in R by including the number 1 in the RHS of the formula:

```
mod0 <- lm(vote_dem ~ 1, data = cces20)
tidy(mod0)
```

```
## # A tibble: 1 × 5
##   term        estimate std.error statistic p.value
##   <chr>          <dbl>     <dbl>     <dbl>   <dbl>
## 1 (Intercept)    0.597   0.00234      255.       0
```

Where does the intercept of 0.597 come from?

Intercept-only model

It’s just the mean of `vote_dem` in our sample!

```
cces20 %>%
  summarize(mean_vote_dem = mean(vote_dem))
```

```
## # A tibble: 1 × 1
##   mean_vote_dem
##           <dbl>
## 1         0.597
```

Adding a binary predictor

Now let’s add a binary predictor to the model: `gender`.

Note: we don’t need to keep using the number 1 to specify the intercept, R includes it by default.

```
mod1 <- lm(vote_dem ~ gender, data = cces20)
tidy(mod1)
```

```
## # A tibble: 2 × 5
##   term         estimate std.error statistic  p.value
##   <chr>           <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)    0.547    0.00351     156.  0
## 2 genderFemale   0.0889   0.00470      18.9 1.28e-79
```

The intercept has changed from 0.597 to 0.547. Why?

Adding a binary predictor

Before, the intercept was the proportion of *all respondents* who voted Democratic.

Now the intercept is the proportion of *males* who voted Democratic:

```
cces20 %>%
  filter(gender == "Male") %>%
  summarize(mean_vote_dem_male = mean(vote_dem))
```

```
## # A tibble: 1 × 1
##   mean_vote_dem_male
##                <dbl>
## 1              0.547
```

Why?

Predicted values

This is our sample regression model:

$$\hat{Y}_i = 0.547 + (0.0889 \cdot Female_i)$$

For men:

$$\hat{Y}_{men} = 0.547 + (0.0889 \cdot 0) = 0.547$$

For women:

$$\hat{Y}_{women} = 0.547 + (0.0889 \cdot 1) = 0.636$$

So the mean of `vote_dem` among women is 0.636, and the coefficient on *Female* represents the **difference** between the mean of `vote_dem` among women vs men.

Statistical significance

If we have the coefficient and standard error, we can compute the t-statistic and the p-value.

```
## # A tibble: 2 × 5
##   term         estimate std.error statistic  p.value
##   <chr>           <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)    0.547    0.00351     156.   0
## 2 genderFemale   0.0889   0.00470      18.9 1.28e-79
```

```
(0.08891878 - 0) / 0.004696622
```

```
## [1] 18.9325
```

```
2 * pt(-abs(18.9325), df = nrow(cces20)-1 )
```

```
## [1] 1.280106e-79
```

Key takeaways

- For a categorical predictor variable, one category is “omitted”. (This is also called the “reference category”.)
- The coefficients represent the effect of being in a given category *versus being in the reference category.*
- So including a categorical variable with n categories is equivalent to including $n - 1$ dummy variables.

Adding a continuous variable

Let's include age alongside `gender` as a predictor in the model:

```
mod2 <- lm(vote_dem ~ gender + age, data = cces20)
tidy(mod2)
```

```
## # A tibble: 3 × 5
##   term          estimate std.error statistic   p.value
##   <chr>            <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)    0.810    0.00846       95.7 0
## 2 genderFemale   0.0768   0.00465       16.5 3.61e- 61
## 3 age           -0.00481  0.000142     -34.0 2.20e-250
```

How do we interpret the coefficient on age?

A 1-unit increase in age is associated with a $100 \cdot 0.00481 = 0.481$ *decrease* in the predicted probability of voting Democratic.

Predicted values

What is the predicted probability of voting Democratic for a man aged 30? What about a woman aged 30?

This is our sample regression model:

$$\hat{Y}_i = 0.810 + (0.0768 \cdot Female_i) - (0.00481 \cdot Age_i)$$

Man aged 30:

$$\hat{Y}_{man30} = 0.810 + (0.0768 \cdot 0) - (0.00481 \cdot 30) = 0.666$$

Woman aged 30:

$$\hat{Y}_{woman30} = 0.810 + (0.0768 \cdot 1) - (0.00481 \cdot 30) = 0.743$$

Interactions

Perhaps the effect of age varies by gender. To allow for this, we need to include an *interaction term* in our model, where the values of the two variables are multiplied together.

In R, an interaction between x and y is written as x*y or, equivalently, `x + y + x:y`.

```
mod3 <- lm(vote_dem ~ gender*age, data = cces20)
tidy(mod3)
```

```
## # A tibble: 4 × 5
##   term             estimate std.error statistic   p.value
##   <chr>               <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)       0.882    0.0123       71.4  0        
## 2 genderFemale     -0.0457   0.0160       -2.86 4.25e-  3
## 3 age              -0.00614  0.000217    -28.2  8.79e-174
## 4 genderFemale:age  0.00229  0.000286      8.01 1.19e- 15
```

Interactions

```
## # A tibble: 4 × 5
##   term              estimate std.error statistic   p.value
##   <chr>                <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)         0.882    0.0123       71.4  0        
## 2 genderFemale       -0.0457   0.0160       -2.86 4.25e-  3
## 3 age                -0.00614  0.000217    -28.2  8.79e-174
## 4 genderFemale:age    0.00229  0.000286      8.01 1.19e- 15
```

How do we interpret the interaction term?

The effect of being a women vs being a man, aged 30, is: $-0.0457 + (0.0022 \cdot 1 \cdot 30) = 0.023$

The effect of being a women vs being a man, aged 60, is: $-0.0457 + (0.0022 \cdot 1 \cdot 60) = 0.092$

Among 30-year-olds, women are **2.3 ppts** more like to vote Democratic than men.

Among 60-year-olds, women are **9.2 ppts** more likely to vote Democratic than men.

Interactions

We can use the ggpredict() function from the ggeffects package to get a nice summary of the interaction effects:

```
library(ggeffects)
ggpredict(mod3,
          terms = c("gender", "age [30, 60]"))
```

```
## # Predicted values of vote_dem
##
## # age = 30
##
## gender | Predicted |       95% CI
## ---------------------------------
## Male   |      0.70 | [0.69, 0.71]
## Female |      0.72 | [0.71, 0.73]
##
## # age = 60
##
## gender | Predicted |       95% CI
## ---------------------------------
## Male   |      0.51 | [0.51, 0.52]
## Female |      0.61 | [0.60, 0.61]
```

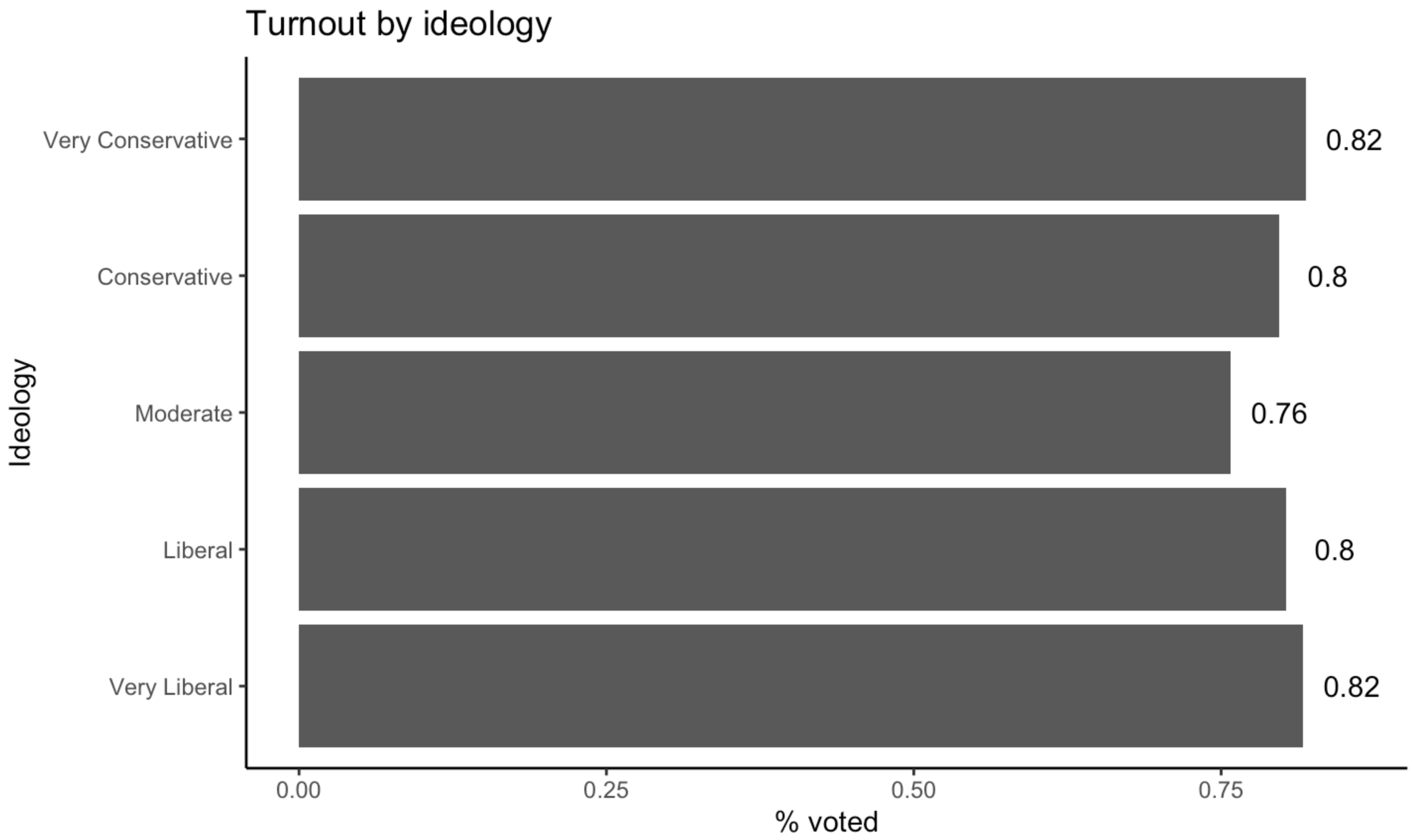
Non-linear relationships

Consider the relationship between political ideology and turnout.

We might imagine that people with a stronger ideology (in either direction) will be more engaged in politics and hence more likely to turnout.

```
cces20 %>%
  filter(ideo5!="Not Sure") %>%
  group_by(ideo5) %>%
  summarize(m = mean(voted, na.rm = TRUE)) %>%
  ggplot(aes(x = ideo5, y = m,
             label = round(m, 2))) +
  geom_col() +
  geom_text(nudge_y = 0.04) +
  labs(x = "Ideology", y = "prop voted",
       title = "Turnout by ideology") +
  coord_flip()
```

Non-linear relationships



Non-linear relationships

To examine this, let’s turn ideology into a numeric variable ranging from -2 (Very Liberal) to +2 (Very Conservative).

```
cces20 <- cces20 %>%
  mutate(ideo_num = case_when(ideo5 == "Very Liberal" ~ -2,
                              ideo5 == "Liberal" ~ -1,

                              ideo5 == "Moderate" ~ 0,
                              ideo5 == "Conservative" ~ 1,
                              ideo5 == "Very Conservative" ~ 2))
```

The `ideo_num` variable is not a good predictor of turnout:

```
lm(voted ~ ideo_num, data = cces20) %>% tidy()
```

```
## # A tibble: 2 × 5
##   term         estimate std.error statistic p.value
##   <chr>           <dbl>     <dbl>     <dbl>   <dbl>
## 1 (Intercept)  0.791      0.00197   402.      0
## 2 ideo_num    -0.000914   0.00160    -0.572   0.567
```

Non-linear relationships

However, `ideo_num^2` *is* a significant predictor of turnout, because the relationship is non-linear:

```
mod4 <- lm(voted ~ ideo_num + I(ideo_num^2),
           data = cces20)
tidy(mod4)
```

```
## # A tibble: 3 × 5
##   term           estimate std.error statistic  p.value
##   <chr>             <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)    0.772      0.00271   285.    0       
## 2 ideo_num      -0.000378   0.00160    -0.237 8.13e- 1
## 3 I(ideo_num^2)  0.0127     0.00123    10.3   6.61e-25
```

Note: It’s important to include the quadratic term like this, rather than by adding a new variable equal to `ideo_num^2`, because R needs to know that `ideo_num` and `ideo_num^2` cannot vary independently.

Predicted values

Here is our sample regression model:

$$\hat{Y}_i = 0.772 - (0.000378 \cdot ideo_i) + (0.0127 \cdot ideo_i^2)$$

For those who are Very Liberal (ideo = -2):

$$\hat{Y}_i = 0.772 - (0.000378 \cdot -2) + (0.0127 \cdot 4) = 0.823$$

For those who are Moderate (ideo = 0):

$$\hat{Y}_i = 0.772 - (0.000378 \cdot 0) + (0.0127 \cdot 0) = 0.772$$

For those who are Very Conservative (ideo = 2):

$$\hat{Y}_i = 0.772 - (0.000378 \cdot 2) + (0.0127 \cdot 4) = 0.822$$

Predicted values

Of course, we don't need to do this manually.

We can use `ggpredict()` to give us predicted values for each value of `ideo_num`:

```
ggpredict(mod4, terms = "ideo_num")
```

```
## # Predicted values of voted
##
## ideo_num | Predicted |       95% CI
## -----------------------------------
##       -2 |      0.82 | [0.81, 0.83]
##       -1 |      0.79 | [0.78, 0.79]
##        0 |      0.77 | [0.77, 0.78]
##        1 |      0.78 | [0.78, 0.79]
##        2 |      0.82 | [0.81, 0.83]
```

Practice problems

1. How many categories does the variable `race` have?
2. How many coefficients will this produce if we include `race` in our model?
3. Run a regression of Democratic voting on race.
4. What is the reference category?
5. Interpret each coefficient.
6. Create dummy variables for each category of `race`. Re-run the regression using all of these dummy variables except `White,` instead of using the variable `race`. How do the coefficients compare?
7. Re-run the regression, choosing the dummy variables so that `Hispanic` is the omitted category. Interpret the coefficients.

Practice problems (answers)

Q1. How many categories does the variable `race` have?

The variable `race` has 5 categories.

```
cces20 %>% count(race)
```

```
## # A tibble: 5 × 2
##   race          n
##   <fct>     <int>
## 1 White     32869
## 2 Black      4071
## 3 Hispanic   4081
## 4 Other      1724
## 5 Asian      1145
```

Q2. How many coefficients will this produce if we include `race` in our model?

Including `race` will produce $5 - 1 = 4$ coefficients.

Practice problems (answers)

Q3. Run a regression of Democratic voting on race.

```
lm(vote_dem ~ race, data = cces20) %>% tidy()
```

```
## # A tibble: 5 × 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)   0.544    0.00263    207.   0
## 2 raceBlack     0.389    0.00791     49.1  0
## 3 raceHispanic  0.124    0.00790     15.7  4.45e-55
## 4 raceOther    -0.0251   0.0118      -2.14 3.26e- 2
## 5 raceAsian     0.223    0.0143      15.5  2.32e-54
```

Q4. What is the reference category?

The reference category is `White`.

Practice problems (answers)

Q5. Interpret each coefficient.

```
## # A tibble: 5 × 5
##   term          estimate std.error statistic  p.value
##   <chr>            <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)     0.544    0.00263    207.   0       
## 2 raceBlack       0.389    0.00791     49.1  0       
## 3 raceHispanic    0.124    0.00790     15.7  4.45e-55
## 4 raceOther      -0.0251   0.0118      -2.14 3.26e- 2
## 5 raceAsian       0.223    0.0143      15.5  2.32e-54
```

Being Black (vs White) is associated with a **38.9** percentage point *increase* in the probability of voting Democrat.

Being Hispanic (vs White) is associated with a **12.4** percentage point *increase* in the probability of voting Democrat.

Being in the “Other” race category (vs White) is associated with a **2.5** percentage point *decrease* in the probability of voting Democrat.

Being Asian (vs White) is associated with a **22.3** percentage point *increase* in the probability of voting Democrat.

Practice problems (answers)

Q6. Create dummy variables for each category of `race`. Re-run the regression using all of these dummy variables except `White`, instead of using the variable `race`. How do the coefficients compare?

```r
cces20 <- cces20 %>%
  mutate(White = ifelse(race == "White", 1, 0),
         Black = ifelse(race == "Black", 1, 0),
         Hispanic = ifelse(race == "Hispanic", 1, 0),
         Other = ifelse(race == "Other", 1, 0),
         Asian = ifelse(race == "Asian", 1, 0))

lm(vote_dem ~ Black + Hispanic + Other + Asian, data = cces20) %>% tidy()
```

```
## # A tibble: 5 × 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)   0.544    0.00263    207.   0
## 2 Black         0.389    0.00791     49.1  0
## 3 Hispanic      0.124    0.00790     15.7  4.45e-55
## 4 Other        -0.0251   0.0118      -2.14 3.26e- 2
## 5 Asian         0.223    0.0143      15.5  2.32e-54
```

Exactly the same results!

Practice problems (answers)

Q7. Re-run the regression, choosing the dummy variables so that `Hispanic` is the omitted category. Interpret the coefficients.

```
lm(vote_dem ~ White + Black + Other + Asian, data = cces20) %>% tidy()
```

```
## # A tibble: 5 × 5
##   term        estimate std.error statistic   p.value
##   <chr>          <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)   0.668    0.00745     89.6  0
## 2 White        -0.124    0.00790    -15.7  4.45e- 55
## 3 Black         0.265    0.0105      25.1  2.59e-138
## 4 Other        -0.149    0.0137     -10.9  1.51e- 27
## 5 Asian         0.0988   0.0159       6.21 5.43e- 10
```

The coefficients should now be interpreted as the “effect” of being in a given racial category versus being Hispanic.