Problem set 5

Your name here

Due 11/5/2021 at 5pm

NOTE: Start with the file **`ps6_2021.Rmd`** *(available from the github repository at https://github.com/UChicago-pol-methods/IntroQSS-F21/tree/main/assignments). Modify that file to include your answers. Make sure you can "knit" the file (e.g. in RStudio by clicking on the* **`Knit`** *button). Submit both the Rmd file and the knitted PDF via Canvas*

In this assignment we will examine data from the Cumulative CCES Common Content dataset assembled by Shiro Kuriwaki from the 2006-2020 Cooperative Congressional Election Studies. You can find the dataset and the codebook for the dataset at https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/II2DB6. They are also on the course repository.

The code chunk below loads the data and creates a few variables to get you started. (You may have to change the path to get the code to run, depending on where you saved the dataset.) In subsequent code you should work with `dat`, which is created by this code chunk.

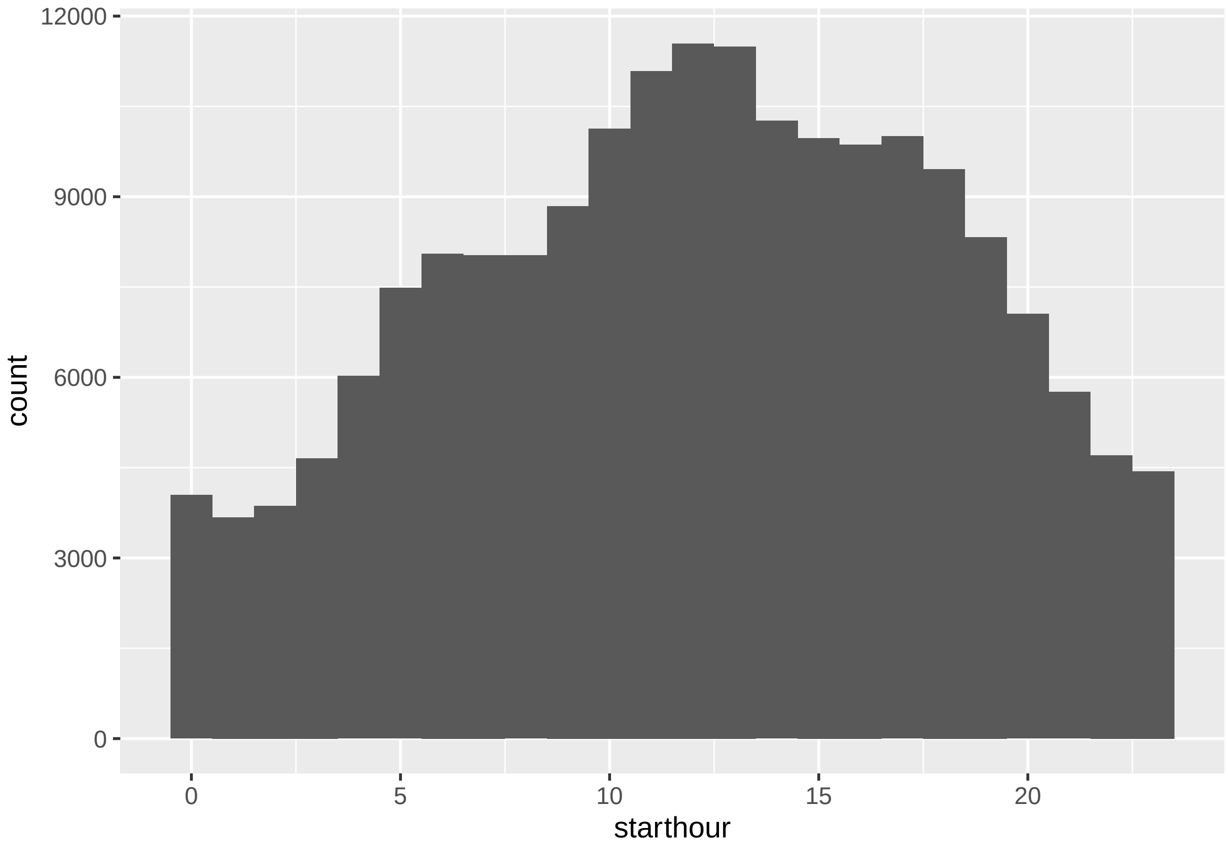
```r
cces <- readRDS("./../data/cces/cumulative_2006-2020.rds")

dat <- cces %>%
  filter(!st %in% c("IN", "KY", "TN", "NE", "KS", "SD", "ND", "ID", "HI", "AK") & year %% 2 == 0) %>%
  mutate(starthour = lubridate::hour(starttime),
         starthour = case_when(st %in% c("WA", "OR", "NV", "CA") ~ starthour - 3,
                               st %in% c("MT", "WY", "CO", "UT", "NM", "AZ") ~ starthour - 2,
                               st %in% c("OK", "TX", "MN", "IA", "MO", "AR", "LA", "WI",
                                         "IL", "MS", "AL") ~ starthour - 1),
         starthour = ifelse(starthour < 0, 24 + starthour, starthour),
         approve_pres = as.integer(approval_pres %in% c(1,2)),
         startcat = case_when(starthour >= 5 & starthour < 12 ~ "1) morning",
                             starthour >=12 & starthour < 17 ~ "2) afternoon",
                             starthour >= 17 | starthour < 1 ~ "3) evening",
                             starthour >= 1 & starthour < 5 ~ "4) late night"))
```

1) Create a histogram of `starthour`, which indicates what time (on the 24 hour clock) the respondent started the survey. Specify `binwidth = 1`.

```r
dat %>%
  ggplot(aes(x = starthour)) +
  geom_histogram(binwidth = 1)
```

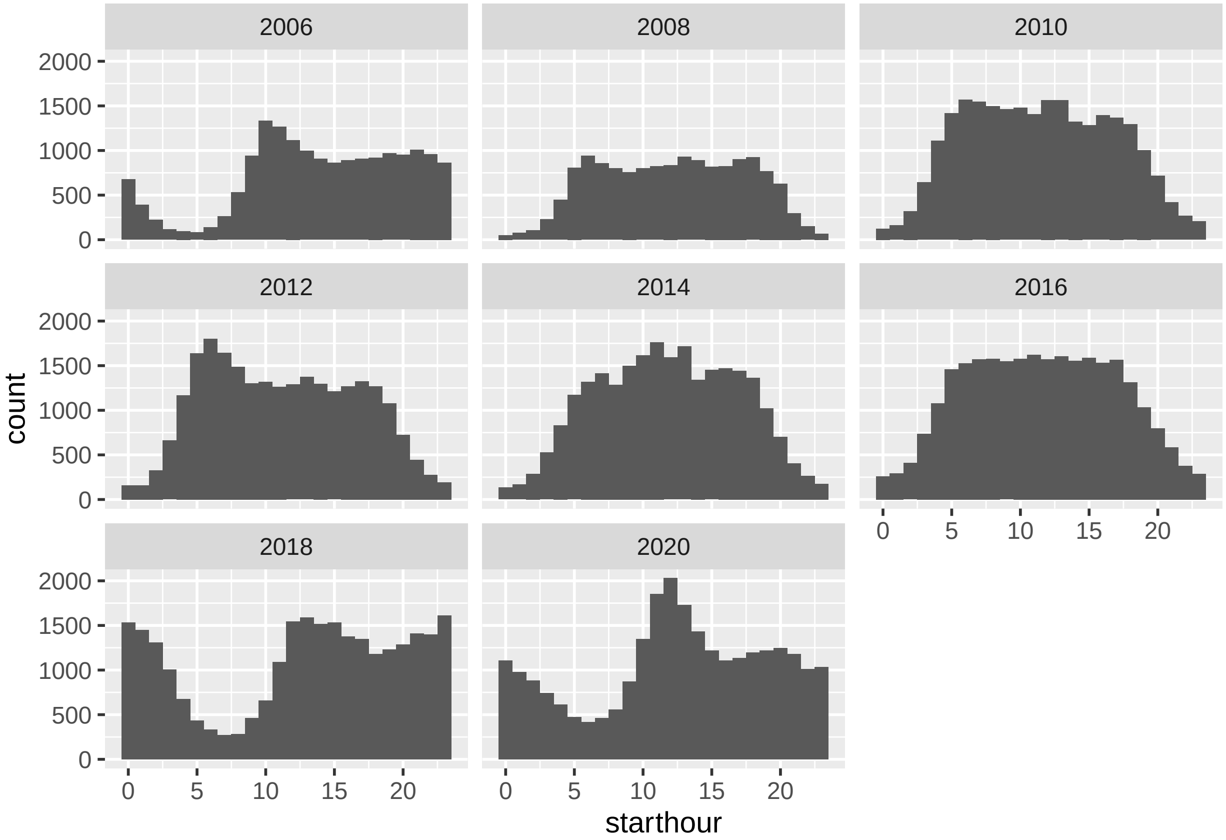
```
## Warning: Removed 196781 rows containing non-finite values (stat_bin).
```



2) Make another figure that shows the same histogram separately by year. You should see that there seems to be one pattern of survey timing for 2006, 2018, and 2020, and another for the other years. Which one is less surprising to you?

```
dat %>%
  ggplot(aes(x = starthour)) +
  geom_histogram(binwidth = 1) +
  facet_wrap(vars(year))
```

```
## Warning: Removed 196781 rows containing non-finite values (stat_bin).
```



Both patterns are a little surprising. In the 2008-2016 pattern, the least common time to start the survey is about midnight, which seems a bit early. (Do more people really take it at 3am?) In the 2006/2018/2020 pattern, the least common time to take the survey is between 5 and 7, which is also a little surprising.

We're not sure why there are these two distinct patterns (it could be a difference in the manner of survey administration, or a coding error), but to be safe we'll focus on years with a similar pattern of response and the same president. In subsequent questions, restrict attention to 2010-2016.

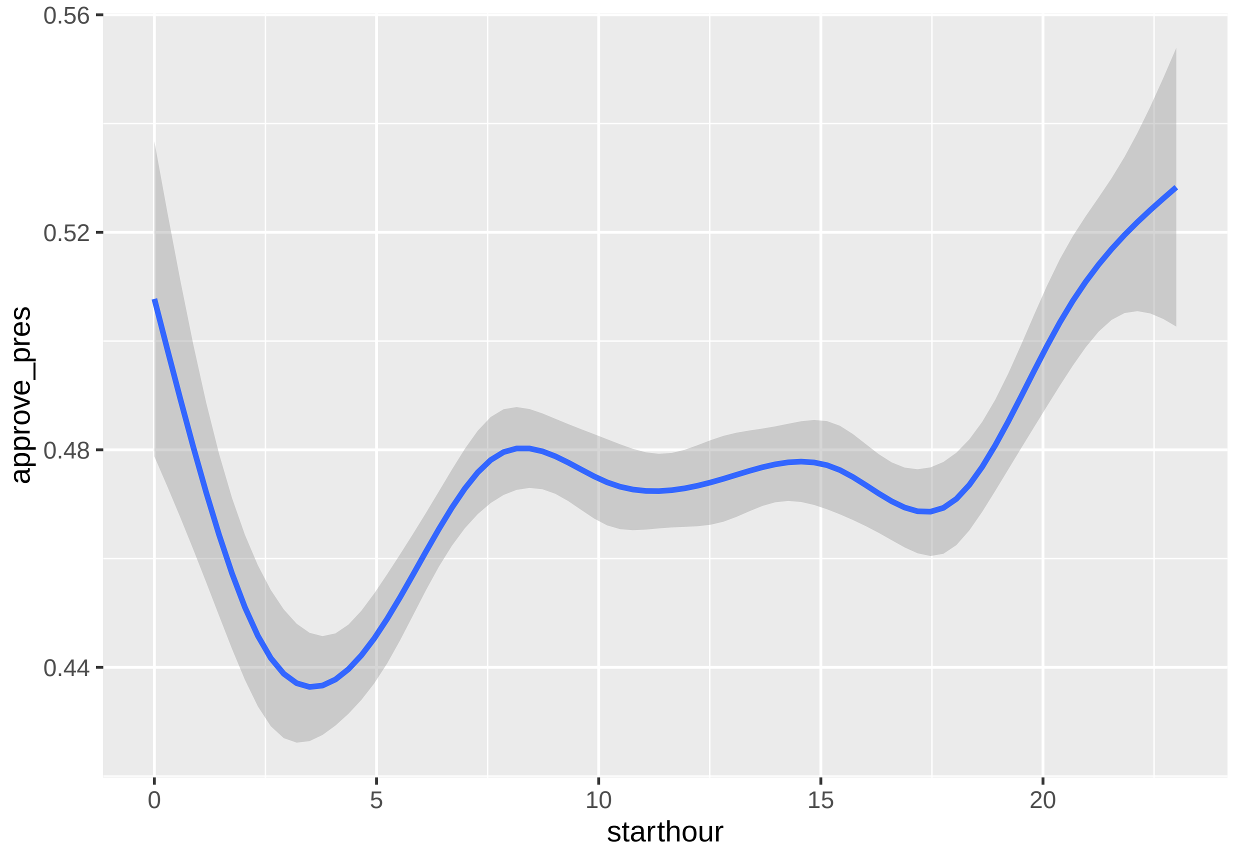
3) Use `geom_smooth()` to show how the average `approve_pres` (a variable that was created above) changes over the course of the day. Interpret the result.

```
dat %>%
  filter(year %in% c(2010:2016)) -> dat2

dat2 %>%
  ggplot(aes(x = starthour, y = approve_pres)) +
  geom_smooth()
```

```
## 'geom_smooth()' using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```

```
## Warning: Removed 108669 rows containing non-finite values (stat_smooth).
```

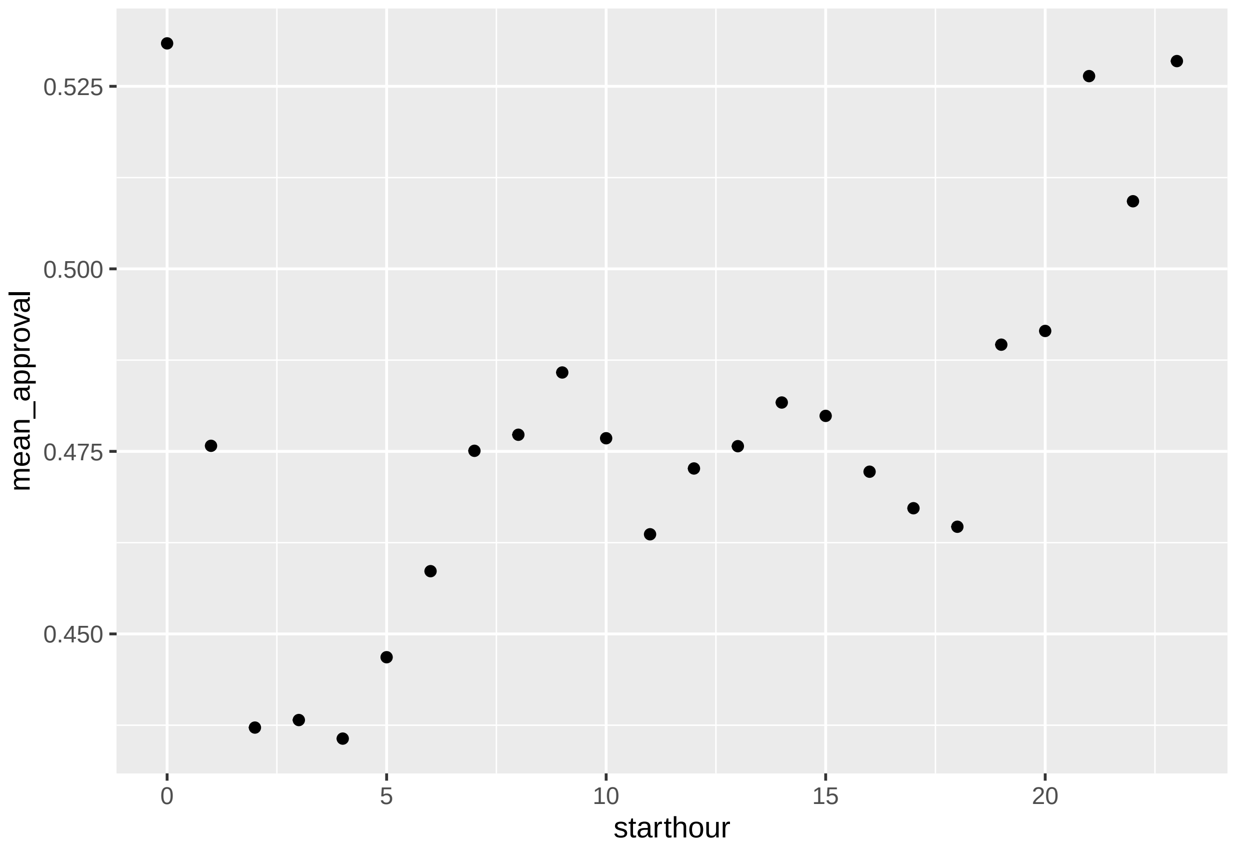


The figure suggests that average approval of the present is lowest in the early morning hours and highest in the late evening.

4) Use `group_by()` and `summarize()` to compute the average of `approve_pres` by `starthour` (another variable that was created above) and plot the result.

```
dat2 %>%
  group_by(starthour) %>%
  summarize(mean_approval = mean(approve_pres, na.rm = T)) %>%
  ggplot(aes(x = starthour, y = mean_approval)) +
  geom_point()
```

```
## Warning: Removed 1 rows containing missing values (geom_point).
```



5) Regress `approve_pres` on `startcat` (another variable created above) and interpret the coefficients. Calculate the mean of `approve_pres` by `startcat` using `group_by()` and `summarize()` and compare this to the regression coefficients.

```
lm(approve_pres ~ startcat, data = dat2) %>% coef()
```

```
##          (Intercept)  startcat2) afternoon    startcat3) evening
##          0.469172429           0.007154341           0.016032578
## startcat4) late night
##         -0.029005818
```

```
dat2 %>%
  group_by(startcat) %>%
  summarize(mean_approval = mean(approve_pres, na.rm = T))
```

```
## # A tibble: 5 x 2
##   startcat      mean_approval
##   <chr>                 <dbl>
## 1 1) morning            0.469
## 2 2) afternoon          0.476
## 3 3) evening            0.485
## 4 4) late night         0.440
## 5 <NA>                  0.492
```

In the regression, the omitted category is "morning". The intercept thus reports the average approval rate among respondents who start the survey in the morning. The other coefficients indicate the difference in the average approval rate for other groups of respondents relative to the group who started in the morning. If you combine those differences with the intercept, you get the same rates of approval as in the grouped means.

Looking at the results so far, one might wonder whether starting a survey in the evening causes respondents to give a higher rating to the president, while starting the survey in the late night/early morning causes respondents to give a lower rating to the president. There is a small literature suggesting that people's survey responses depend on the time of day when they take the survey.

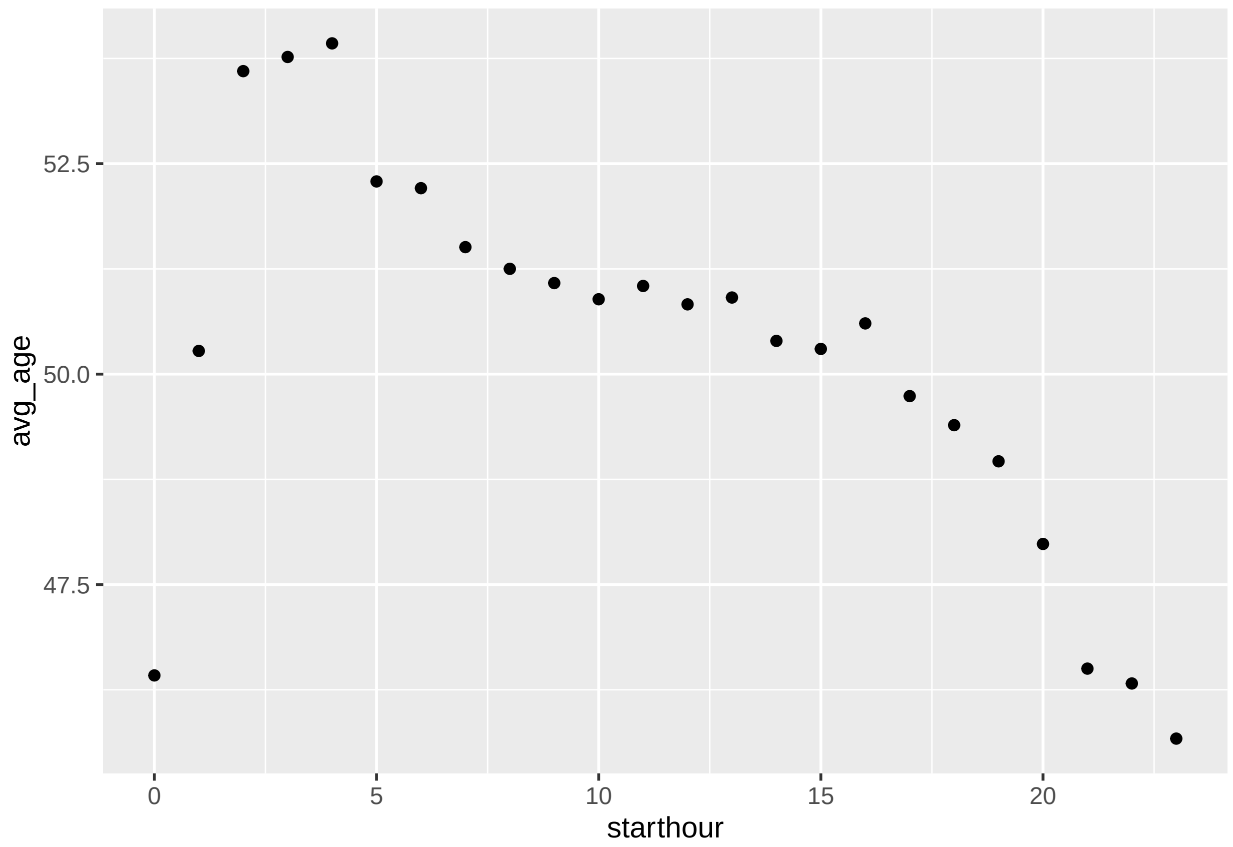
6) In brief, how could you design a randomized experiment to evaluate this hypothesis?

You could deliver surveys to potential respondents at randomly selected different times of day. A tricky challenge here is what to do with people who do not complete the survey right away. If you distribute the surveys at random times, but people start the survey when it's convenient for them, and then you compare surveys based on when the respondents start the survey, you don't have a valid comparison. If you instead compare surveys based on when they were distributed, you do have a valid comparison (and we call the associated estimand the "Intent to Treat", or ITT), but any differences will under reasonable assumptions be smaller than they would be if people took the survey when they were supposed to. This is why we do instrumental variables analysis, which you will see in causal inference.

7) Use `group_by()` and `summarize()` to compute the average age of respondents by `starthour` and plot the result. Interpret what you find. Does it appear that `starthour` is randomly assigned?

```r
dat2 %>%
  group_by(starthour) %>%
  summarize(avg_age = mean(age, na.rm = T)) %>%
  ggplot(aes(x = starthour, y = avg_age)) +
  geom_point()
```

```
## Warning: Removed 1 rows containing missing values (geom_point).
```



No, students who start the survey at different times appear to differ in age as well. People who start the survey at 3am are over 6 years older on average than people who start the survey at midnight.

8) Regress `approve_pres` on age and interpret the coefficients. Does this support the idea that age might be a confounder for the relationship between `starthour` and `approve_pres`?

```
short <- lm(approve_pres ~ age, data = dat2)
short
```

```
##
## Call:
## lm(formula = approve_pres ~ age, data = dat2)
##
## Coefficients:
## (Intercept)          age
##    0.690963    -0.004124
```

Yes, for each additional year of age, the proportion of respondents who approve of the president goes down by .004.

9) Regress `approve_pres` on age and gender and note how the coefficient on `age` differs from the previous regression. Show how to use the omitted variable bias formula to account for this difference.

```
long <- lm(approve_pres ~ age + gender, data = dat2)
long
```

```
##
## Call:
## lm(formula = approve_pres ~ age + gender, data = dat2)
##
## Coefficients:
## (Intercept)          age       gender
##    0.579351    -0.003899     0.065480
```

When we add gender to the regression, the coefficient decreases in magnitude, from -.0041 to -.0039.

To illustrate the omitted variable bias formula, we first regress `gender` *on* `age`*.*

```
auxiliary <- lm(gender ~ age, data = dat2)
auxiliary
```

```
##
## Call:
## lm(formula = gender ~ age, data = dat2)
##
## Coefficients:
## (Intercept)          age
##    1.704539    -0.003443
```

We can then show that the difference in the coefficient on `age` *when we control for* `gender` *is the product of the coefficient on* `gender` *in the* `long` *regression and the coefficient on* `age` *in the auxiliary regression ("impact" vs. "imbalance").*

```
coef(short)["age"] - coef(long)["age"]
```

```
##           age
## -0.0002254188
```

```
coef(long)["gender"]*coef(auxiliary)["age"]  # impact times imbalance
```

```
##        gender
## -0.0002254188
```

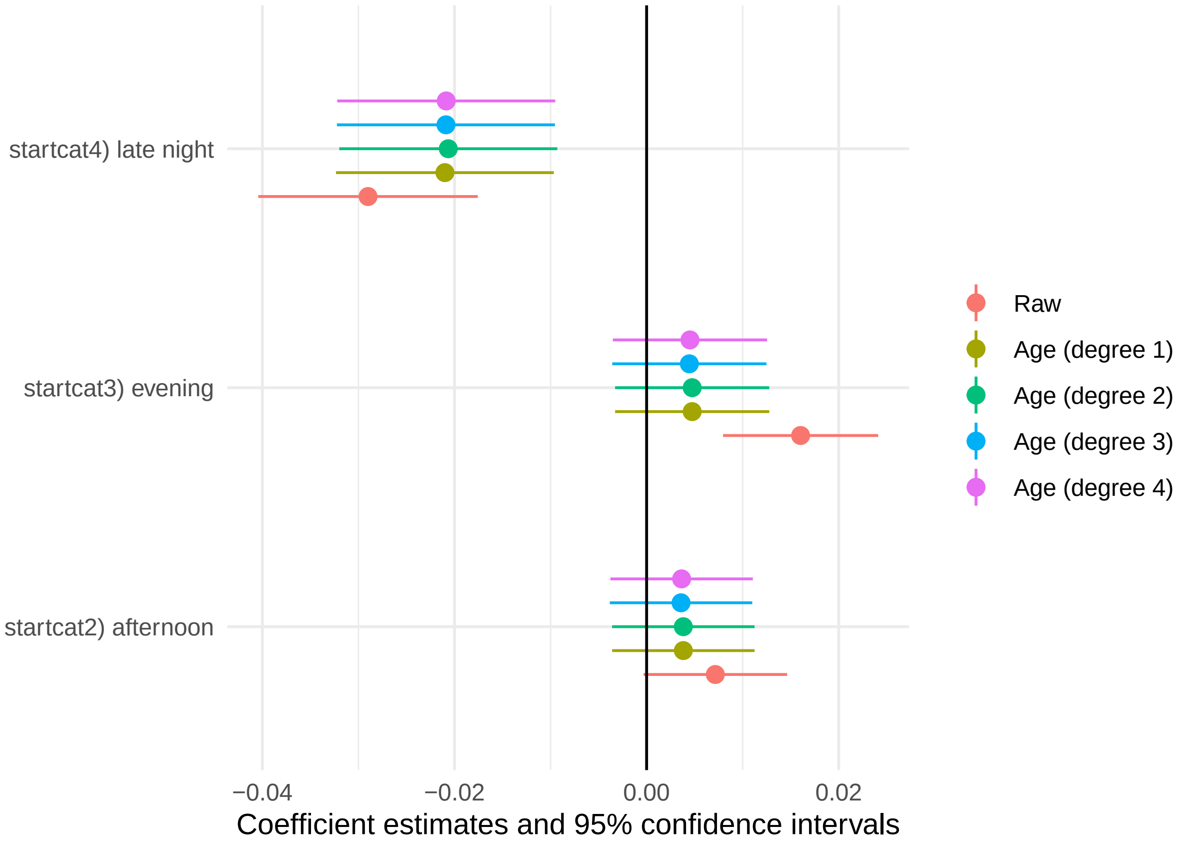
10) Regress `approve_pres` on `startcat` again but now controlling for age. Using the `modelsummary` package (which will be discussed in lab), make a figure or table comparing the key coefficients (i.e. those relating to the the time of day) when you don't control for age and when you control for different polynomials of age. (Hint: You might want to specify `coef_omit = "Int|age"`, which leaves out the intercept and the age coefficients.)

```
mods <- list("Raw" = lm(approve_pres ~ startcat, data = dat2),
             "Age (degree 1)" = lm(approve_pres ~ startcat + age, data = dat2),
             "Age (degree 2)" = lm(approve_pres ~ startcat + poly(age, 2), data = dat2),
             "Age (degree 3)" = lm(approve_pres ~ startcat + poly(age, 3), data = dat2),
             "Age (degree 4)" = lm(approve_pres ~ startcat + poly(age, 4), data = dat2))

modelsummary::modelsummary(mods, coef_omit = "Int|age")
```

	Raw	Age (degree 1)	Age (degree 2)	Age (degree 3)	Age (degree 4)
startcat2) afternoon	0.007	0.004	0.004	0.004	0.004
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
startcat3) evening	0.016	0.005	0.005	0.004	0.005
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
startcat4) late night	−0.029	−0.021	−0.021	−0.021	−0.021
	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)
Num.Obs.	102 288	102 288	102 288	102 288	102 288
R2	0.001	0.017	0.017	0.017	0.017
R2 Adj.	0.001	0.017	0.017	0.017	0.017
AIC	148 117.9	146 418.4	146 417.9	146 402.9	146 403.8
BIC	148 165.6	146 475.6	146 484.6	146 479.2	146 489.6
Log.Lik.	−74 053.962	−73 203.189	−73 201.933	−73 193.439	−73 192.910
F	18.488	443.022	354.925	298.648	256.136

```r
modelsummary::modelplot(mods, coef_omit = "Int|age") +
  geom_vline(xintercept = 0)
```



11) What other variables in the dataset would you control for if you wanted to assess the causal impact of the time of day on approval of the president?

You could choose a lot of options here, but I would look for factors that could affect the time someone takes the survey and might be related to political preferences – things like whether someone works (and what kind *of work), whether someone has children, marital status.*

12) If there is a regression you plan to run as part of your final project, describe it here. If not, explain why not.