Problem set 6

Siwei Dai

November 5, 2021

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.

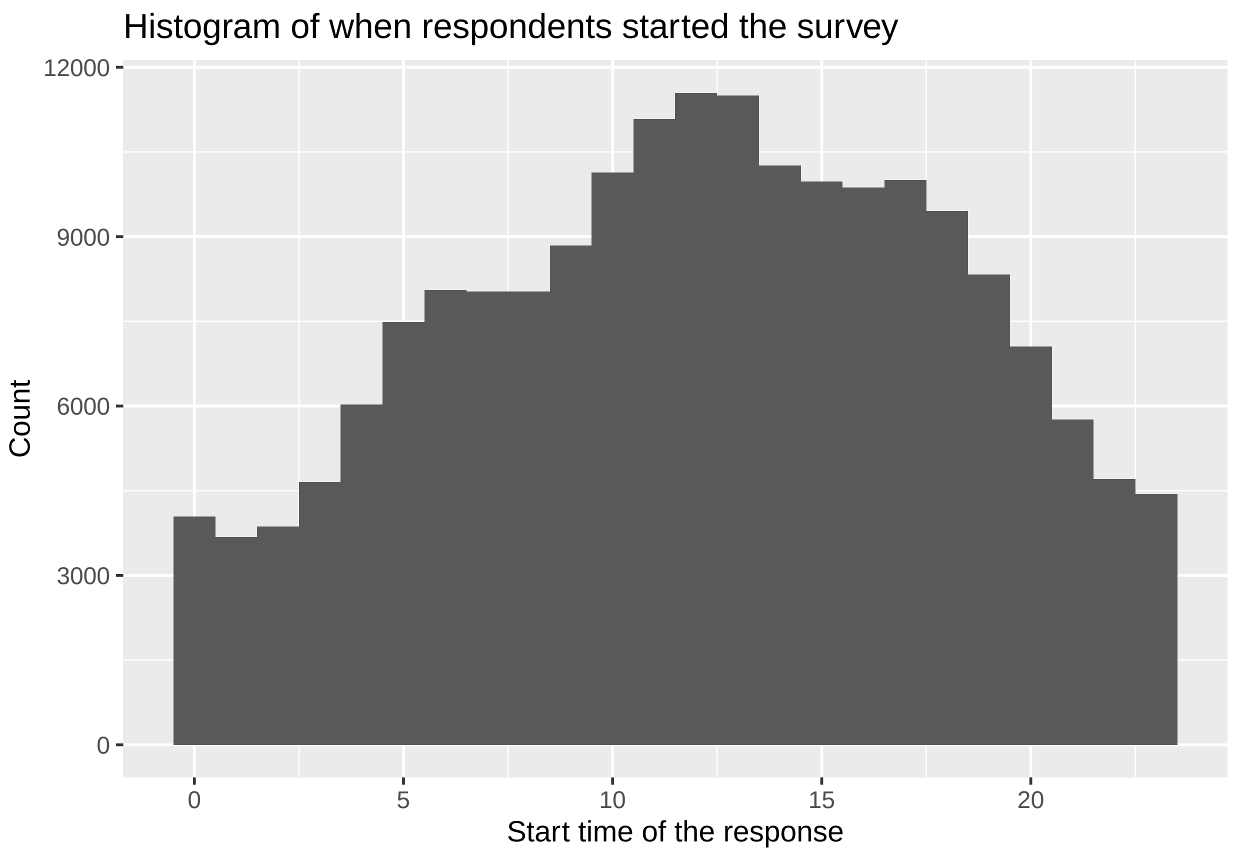
```
cces <- readRDS("./../data/cces/cumulative_2006-2020.rds")

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`.

```
plot1 <- dat %>%
    ggplot(aes(x = starthour)) +
    geom_histogram(binwidth = 1) +
    labs(x = "Start time of the response",
         y = "Count",
         title = "Histogram of when respondents started the survey")
plot1
```

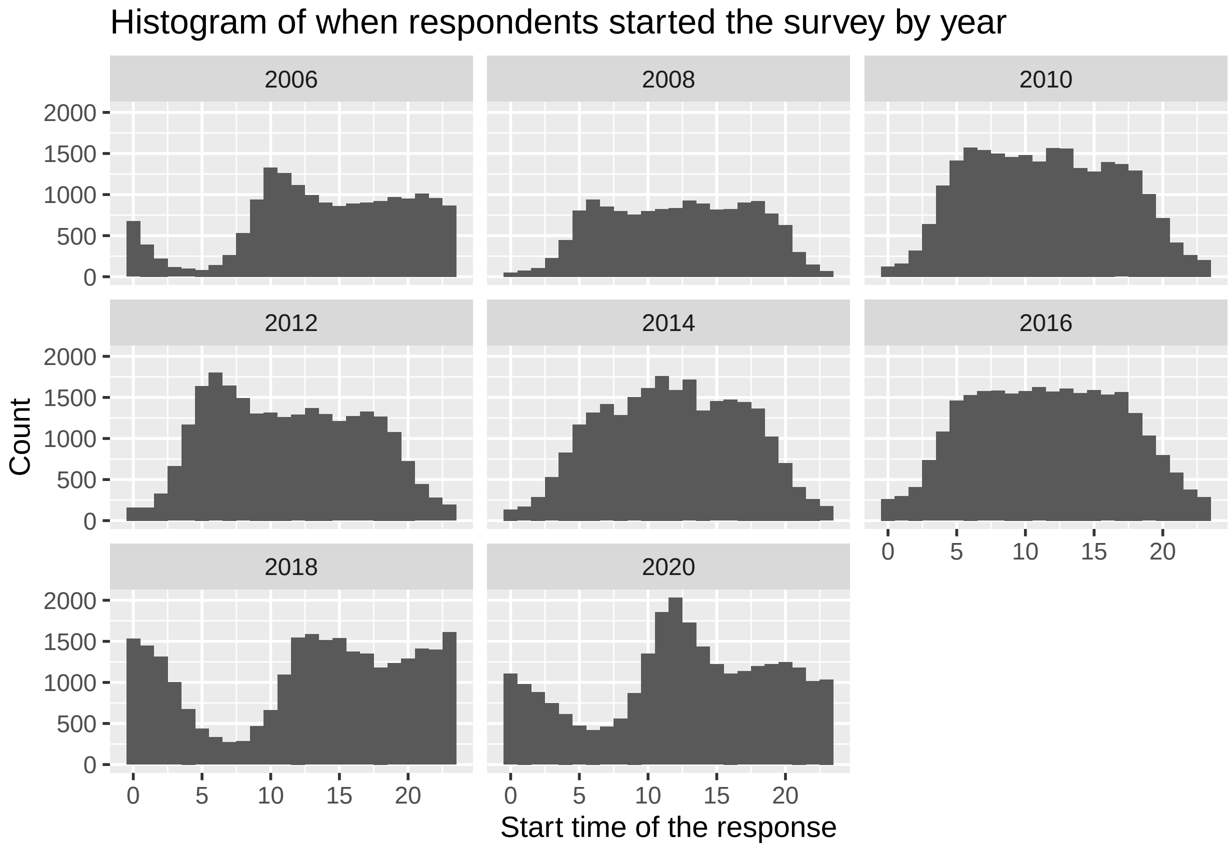
```
## Warning: Removed 196781 rows containing non-finite values (stat_bin).
```



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?

```
plot2 <- dat %>%
    ggplot(aes(x = starthour)) +
    geom_histogram(binwidth = 1) +
    facet_wrap(facets = vars(year)) +
    labs(x = "Start time of the response",
         y = "Count",
         title = "Histogram of when respondents started the survey by year")
plot2
```

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



The other pattern (excluding 2006, 2018, 2020) is less surprising as it reflects the bulk of respondents started survey response in the day, rather than at midnight. The number of respondents also decreases towards midnight and increases when it's approaching morning.

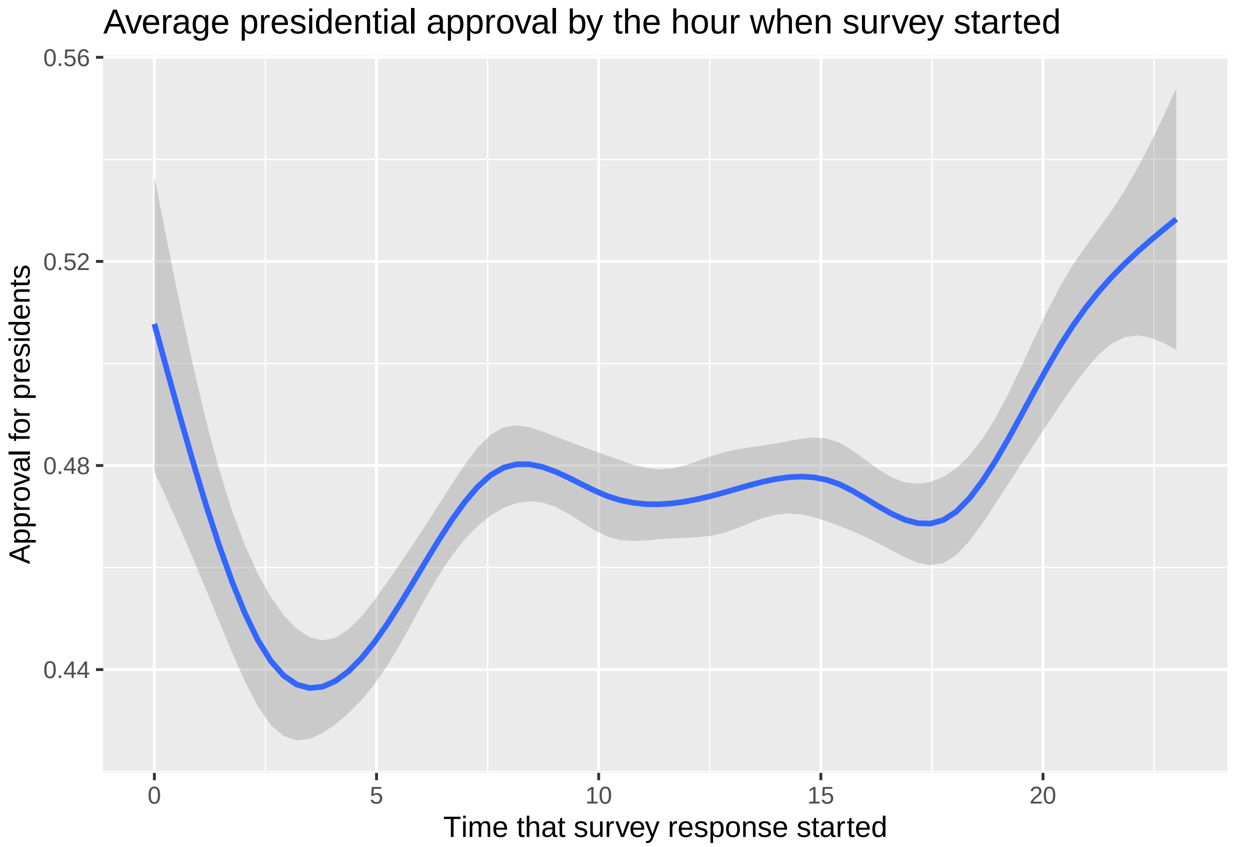
We're not sure why there are these two distinct patterns (it could be a difference in the manner of survey administration, or a coding error), but to be safe we'll focus on years with a similar pattern of response and the same president. In subsequent questions, restrict attention to 2010-2016.

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.

```r
dat <- dat %>%
    filter(year >= 2010 & year <= 2016)
plot3 <- dat %>%
    ggplot(aes(x = starthour,
               y = approve_pres)) +
    geom_smooth() +
    labs(x = "Time that survey response started",
         y = "Approval for presidents",
         title = "Average presidential approval by the hour when survey started")
plot3
```

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

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



The average presidential apporval from those who received survey between 7 AM to 17PM is around 0.47; The average presidential apporval increases in the observations by each increase in survey start time from 17PM to 23PM, to 0.52 for those who received survey at 23PM. The average presidential apporval decreases from the respondents by each increase in survey start time from 0AM to 4AM, to less than 0.44 around 4AM; finally, the average presidential approval in observations increases as survey start time is later around 0.47 by 7AM.

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.

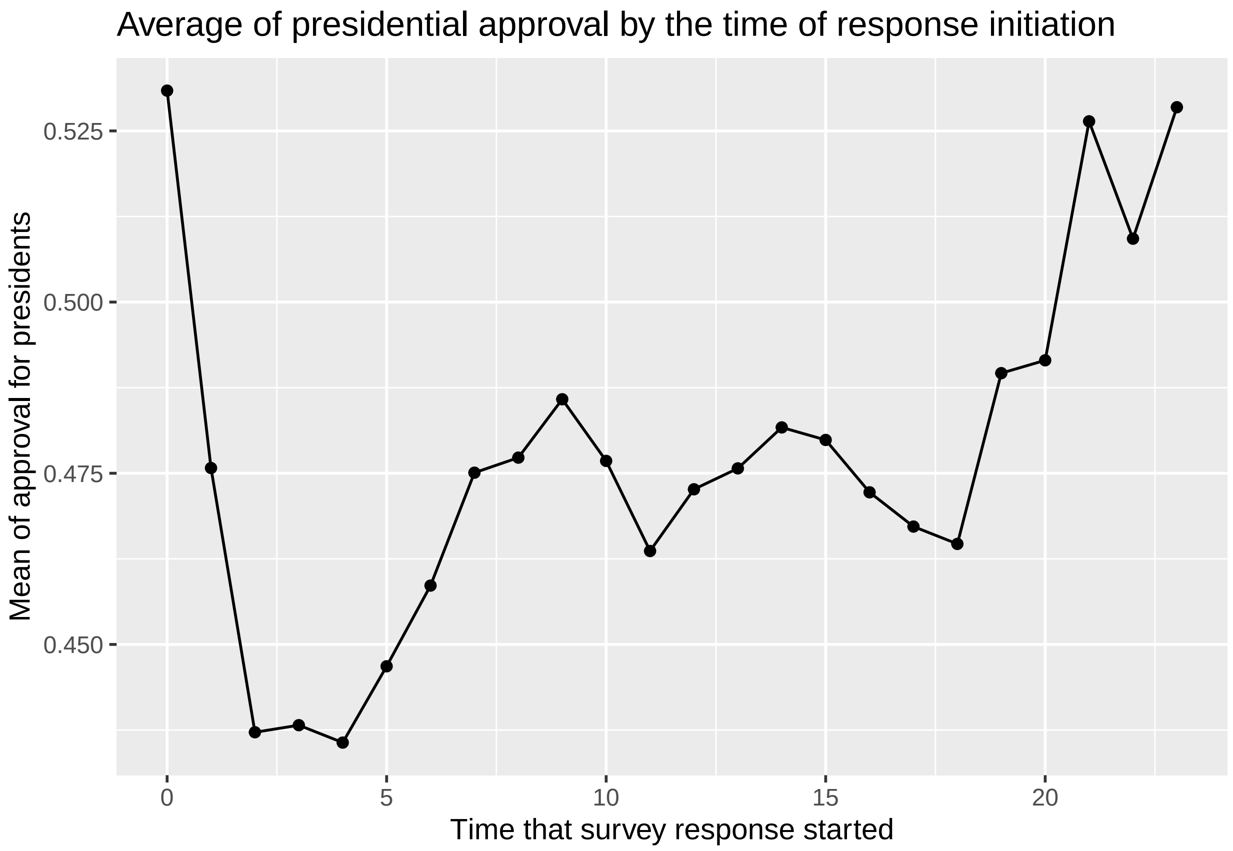
```
# dataset is restricted to 2010 - 2016 as requested in response 3
df4 <- dat %>%
    group_by(starthour) %>%
    summarize(avg_approve = mean(approve_pres, na.rm = TRUE))
plot4 <- df4 %>%
    ggplot(aes(x = starthour, y = avg_approve)) +
    geom_point() +
    geom_line() +
    labs(x = "Time that survey response started",
         y = "Mean of approval for presidents",
         title = "Average of presidential approval by the time of response initiation")
df4
```

```
## # A tibble: 25 x 2
##    starthour avg_approve
##        <dbl>       <dbl>
## 1          0       0.531
## 2          1       0.476
## 3          2       0.437
## 4          3       0.438
## 5          4       0.436
## 6          5       0.447
## 7          6       0.459
## 8          7       0.475
## 9          8       0.477
## 10         9       0.486
## # ... with 15 more rows
```

```
plot4
```

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

```
## Warning: Removed 1 row(s) containing missing values (geom_path).
```



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.

```
# dataset is restricted to 2010 - 2016 as requested in response 3
model5 <- lm(approve_pres ~ startcat, data = dat)
coef(model5)
```

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

```
df5 <- dat %>%
    group_by(startcat) %>%
    summarize(avg_approve = mean(approve_pres, na.rm = TRUE))
df5a <- df5 %>%
    mutate(diff_in_mean = avg_approve - avg_approve[startcat == '1) morning'][1])
df5a
```

```
## # A tibble: 5 x 3
##   startcat      avg_approve diff_in_mean
##   <chr>               <dbl>        <dbl>
## 1 1) morning          0.469      0
## 2 2) afternoon        0.476      0.00715
## 3 3) evening          0.485      0.0160
## 4 4) late night       0.440     -0.0290
## 5 <NA>                0.492      0.0229
```

The intercept means, for observations in which survey started in morning (as the baseline category in the model), the mean of presidential approval is 0.469172429 .The coefficient on "starcat2) afternoon" means that compared to the baseline category (ie people who started their survey in the morning), the mean of presidential approval for people who started the survey in the afternoon is higher by 0.007154341 In the same vein, the coefficient on "starcat3) evening" means that compared to the baseline category, the mean of presidential approval for people who started the survey in the evening is higher by 0.016032578; the coefficient on "starcat4) late night" means that compared to the baseline category, the mean of presidential approval for people who started the survey in late night is lower by 0.029005818

As shown in diff__in__mean column in the "df5a" table, the mean calculations correpond with the interpretations of regression coefficients.

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?

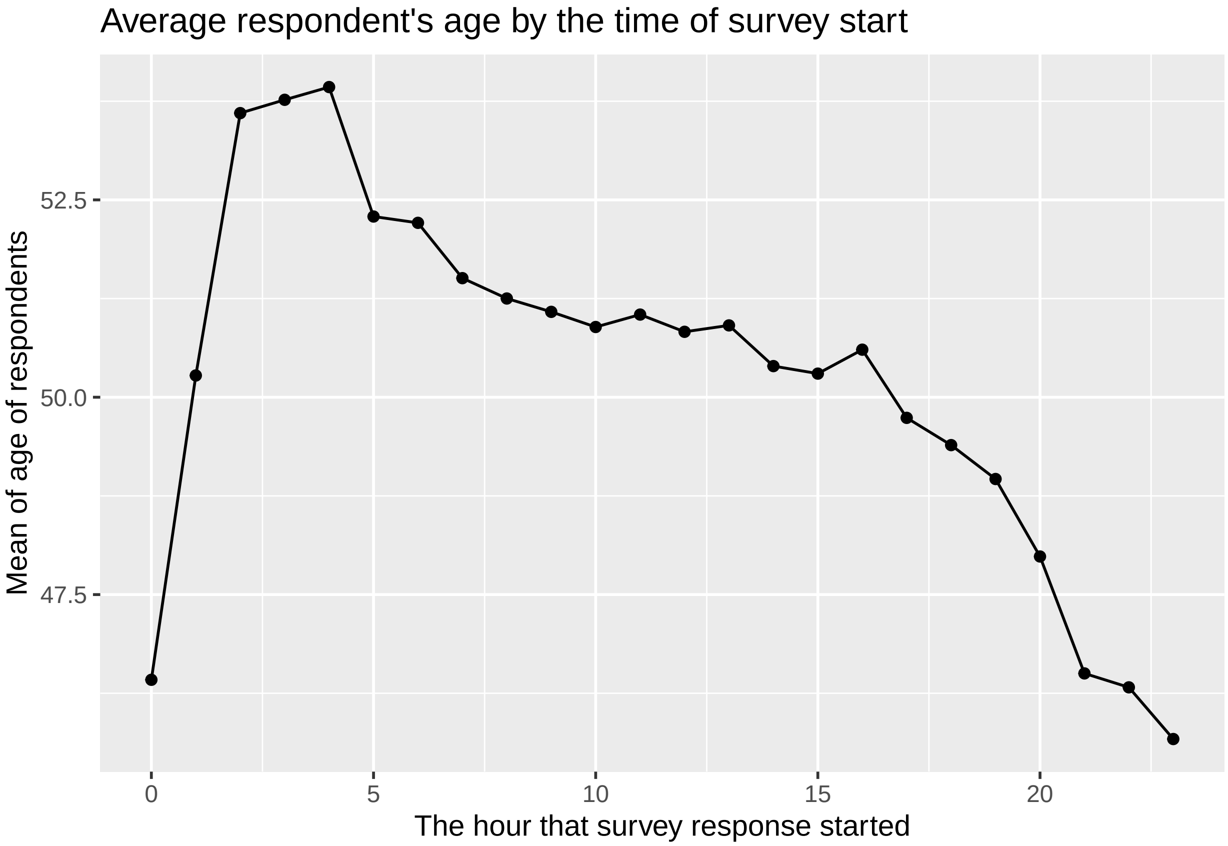
For a given sample, we randomly assign the treatment: the time each respondent start their survey. Then, we can use the difference-in-means estimator to estimate the effect of different response time on presidential approval.

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?

```
df7 <- dat %>%
    group_by(starthour) %>%
    summarize(avg_age = mean(age))
plot7 <- df7 %>%
    ggplot(aes(x = starthour, y = avg_age)) +
    geom_point() +
    geom_line() +
    labs(x = "The hour that survey response started",
         y = "Mean of age of respondents",
         title = "Average respondent's age by the time of survey start")
plot7
```

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

```
## Warning: Removed 1 row(s) containing missing values (geom_path).
```



The average age of people who started their response after 3 AM is larger than the average age of those who started in late night. This means that "starthour" is not randomly assigned; if it were randomly assigned, then the average age of respondents at each hour should be close to the expected value of age throughout the sample.

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`?

```r
model8 <- lm(approve_pres ~ age, data = dat)
coef(model8)
```

```
## (Intercept)          age 
## 0.690963258 -0.004123981
```

The model indicates that age is negatively associated with presidential approval. Consider the correlation we found in Question 7 between "starthour" and "age", it means that age simultaneously is related with "starthour" and "approve_res". This means age might be a confounder for the relationship between "starthour" and "approve_pres".

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.

```r
model9 <- lm(approve_pres ~ age + gender, data = dat)
coef(model9)
```

```
## (Intercept)          age       gender 
## 0.579350870 -0.003898562  0.065479521
```

```r
model9a <- lm(gender ~ age, data = dat)
coef(model9a)
```

```
## (Intercept)          age 
## 1.704538855 -0.003442585
```

$$
\begin{aligned}
\widehat{approve_pres} &= \hat{\beta}_0 + \hat{\beta}_1 age + \hat{\beta}_2 gender \\
\widehat{approve_pres} &= \hat{\gamma}_0 + \hat{\gamma}_1 age \\
\widehat{gender} &= \hat{\alpha}_0 + \hat{\alpha}_1 age \\
\hat{\beta}_1 &= -0.003898562 \\
\hat{\beta}_2 &= 0.065479521 \\
\hat{\alpha}_1 &= -0.003442585 \\
\hat{\gamma}_1 &= \hat{\beta}_1 + \hat{\beta}_2 * \hat{\alpha}_1 \\
&= -0.003898562 + 0.065479521 * (-0.003442585) = -0.004124 \\
OVB &= \hat{\beta}_2 * \hat{\alpha}_1 = -2.2541882 \times 10^{-4}
\end{aligned}
$$

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.)

```r
library(modelsummary)
model10 <- lm(approve_pres ~ startcat, data = dat)
model10a <- lm(approve_pres ~ startcat + age, data = dat)
```

	Age not included	1st-degree age	2nd-degree polynomial of age	3rd-degree polynomial of age
startcat2) afternoon	0.007+	0.004	0.004	0.004
	(0.004)	(0.004)	(0.004)	(0.004)
startcat3) evening	0.016***	0.005	0.005	0.004
	(0.004)	(0.004)	(0.004)	(0.004)
startcat4) late night	−0.029***	−0.021***	−0.021***	−0.021***
	(0.006)	(0.006)	(0.006)	(0.006)
Num.Obs.	102 288	102 288	102 288	102 288
R2	0.001	0.017	0.017	0.017
R2 Adj.	0.001	0.017	0.017	0.017
AIC	148 117.9	146 418.4	146 417.9	146 402.9
BIC	148 165.6	146 475.6	146 484.6	146 479.2
Log.Lik.	−74 053.962	−73 203.189	−73 201.933	−73 193.439
F	18.488	443.022	354.925	298.648

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

```r
model10b <- lm(approve_pres ~ startcat + poly(age, degree = 2), data = dat)
model10c <- lm(approve_pres ~ startcat + poly(age, degree = 3), data = dat)
models <- list("Age not included" = model10,
               "1st-degree age" = model10a,
               "2nd-degree polynomial of age" = model10b,
               "3rd-degree polynomial of age" = model10c)
tb10 <- modelsummary(models,
                     coef_omit = "Int|age",
                     stars = TRUE)
```

```
## Warning: In version 0.8.0 of the `modelsummary` package, the default significance markers produced b
## This warning is displayed once per session.
```

```r
tb10
```

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?

I would also control for "employ" (employment status). People employed in full-time jobs may not be able to start their survey responses during daytime. Meanwhile, employment status impact respondents' attitudes on taxation, welfare programs, which could impact their approval for a given president. Represented in a DAG, employment status is a non-collider as it is nor affected by the treatment.

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

I want to evaluate the effect of having one regional leader being demoted on changes in regional social spending in the following year. The dependent variable would be the difference between the social spending of the year in which bureaucrat removal occurred and the next year's social spending. The independent variables would include a binary variable indicating if there's a regional leader removed in a year, region fixed effects. But since the treatment is not raondomized, I should probably include more control variables after I'm more familiar with the literature.