Problem set 5

Your name here

Due 10/29/2021 at 5pm

NOTE1*: Start with the file* `ps5_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 an experiment that measured the effect of different messages on Michigan residents' likelihood of voting in the August 2006 primary election. The published paper is

Gerber, Alan S., Donald P. Green, and Christopher W. Larimer. 2008. "Social Pressure and Voter Turnout: Evidence from a Large-Scale Experiment." *American Political Science Review* 102(1): 33-48.

The data file is `ggl.RData` and it is found in the `data` directory of the course github repository.

To load the data (you may need to change the path after you save a copy of the dataset locally):

```
load("./../data/ggl.RData")
```

The dataset will be loaded as an object called `ggl`.

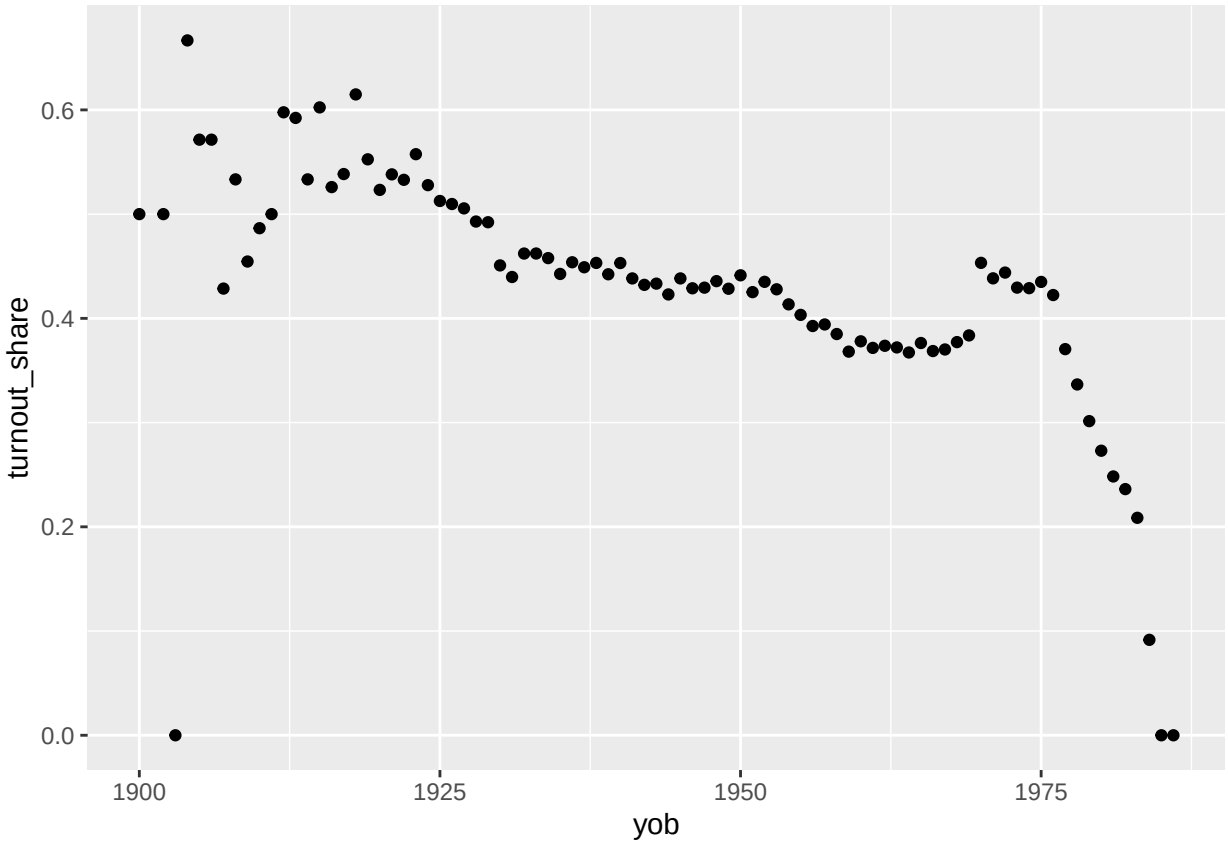
The variables in the dataset are as follows:

- `sex`: male or female
- `yob`: year of birth
- `g2000`, `g2002`, `g2004`: did this voter vote in the general elections in November of 2000, 2002, 2004? (binary)
- `p2000`, `p2002`, `p2004`: did this voter vote in the primary elections of August 2000, 2002, 2004? (binary)
- `treatment`: which of the five treatment did this voter's household receive?
  - "Control": No mailing
  - "CivicDuty": A mailing encouraging voting
  - "Hawthorne": A mailing encouraging voting and saying that the sender will 'be studying voter turnout in the August 8 primary election'
  - "Self": A mailing encouraging voting and showing the recipients' past turnout, saying 'We intend to mail you an updated chart when we have that information'
  - "Neighbors": Same thing, except including information on turnout by neighbors as well
- `cluster`: in what cluster of households was this voter's house located?
- `voted`: did the voter vote in the primary election of 2006?
- `hh_id`: what is the id number of this voter's household?
- `hh_size`: how many voters are in this household?

1) Use grouped summaries (`group_by()` and `summarize()`) to compute the proportion of subjects who voted in the 2002 primary election by year of birth. Make a plot showing the proportion voting (vertical axis) and year of birth (horizontal axis).

```
ggl %>%
  group_by(yob) %>%
  summarize(turnout_share = mean(p2002, na.rm = T)) %>%
```

```
  ggplot(aes(x = yob, y = turnout_share)) +
  geom_point()
```



2) Run a regression with voting in the 2002 primary as the dependent variable and year of birth as the independent variable. Provide the `R` output showing the intercept and slope coefficients. Explain what the slope coefficient means.

```
lm(p2002 ~ yob, data = ggl)
```

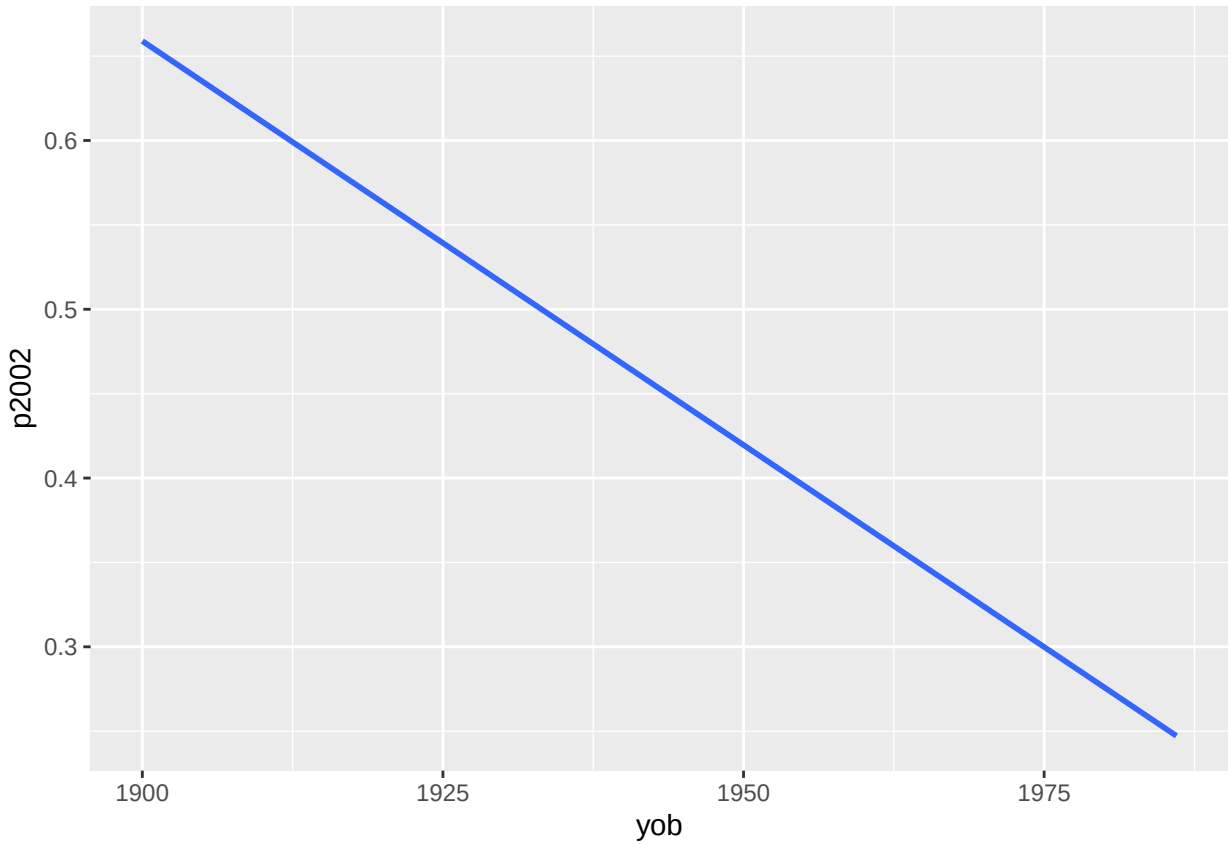
```
##
## Call:
## lm(formula = p2002 ~ yob, data = ggl)
##
## Coefficients:
## (Intercept)          yob
##    9.754675    -0.004787
```

The slope coefficient says that the predicted proportion of people who vote (or predicted probability of voting) decreases by .004787 with every increase in the person's year of birth. Put differently, with each additional year of age, the predicted probability of voting increases by about half a percentage point.

3) Use `geom_smooth()` to show the regression line from the same regression.

```
ggl %>%
  ggplot(aes(x = yob, y = p2002)) +
  geom_smooth(method = lm, se = F)
```

```
## `geom_smooth()` using formula 'y ~ x'
```



4) Using either syntax option shown in lecture (`poly()` or `I()`), run the same regression but now include a 4th-degree polynomial of `yob`. Show the coefficients from the output of `lm()`. Also, show the new regression line using `geom_smooth()` (specify `se = F`).

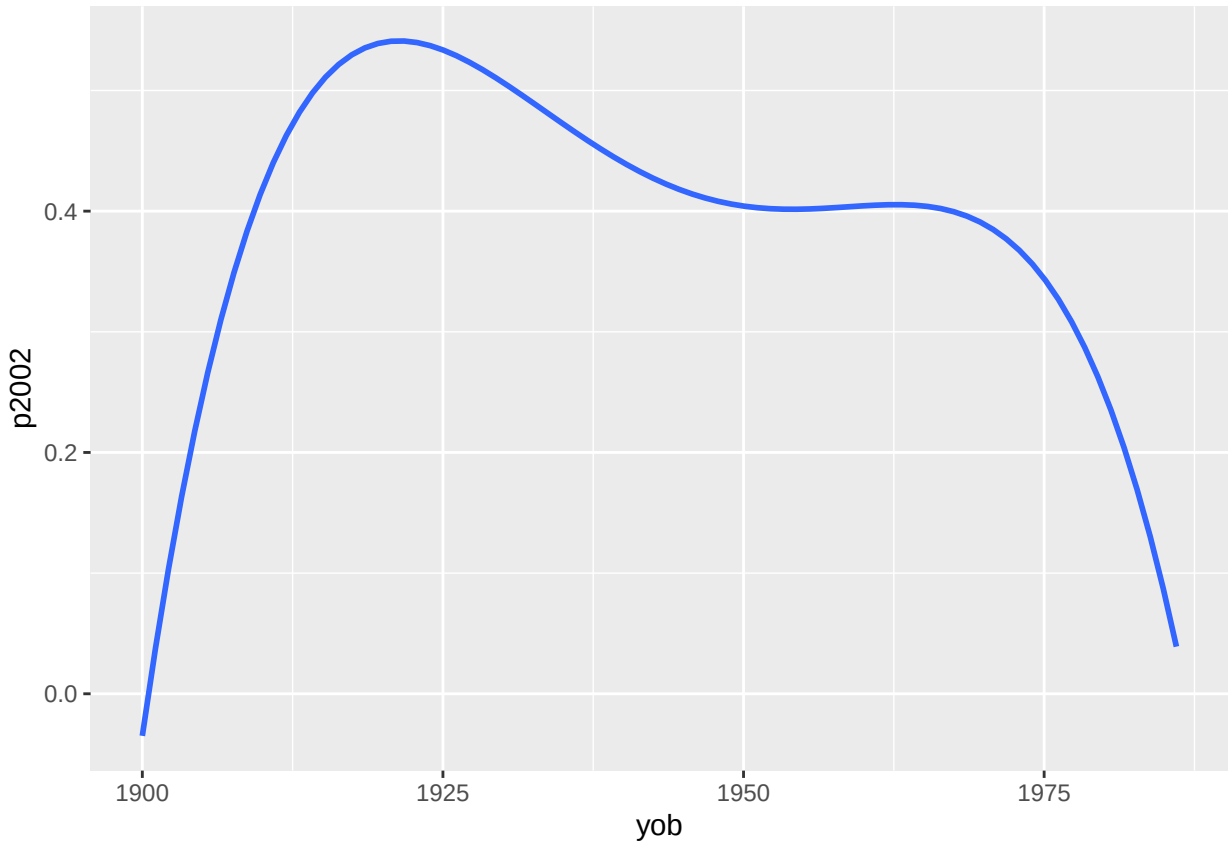
```
lm(p2002 ~ yob + I(yob^2) + I(yob^3) + I(yob^4), data = ggl)
```

```
##
## Call:
## lm(formula = p2002 ~ yob + I(yob^2) + I(yob^3) + I(yob^4), data = ggl)
##
## Coefficients:
## (Intercept)          yob     I(yob^2)     I(yob^3)     I(yob^4)
##  -3.438e+06    7.068e+03   -5.449e+00    1.867e-03   -2.398e-07
```

```
lm(p2002 ~ poly(yob, 4), data = ggl)
```

```
##
## Call:
## lm(formula = p2002 ~ poly(yob, 4), data = ggl)
##
## Coefficients:
##   (Intercept)  poly(yob, 4)1  poly(yob, 4)2  poly(yob, 4)3  poly(yob, 4)4
##        0.3898       -40.5748       -17.2646       -19.2764       -14.9286
```

```
ggl %>%
  ggplot(aes(x = yob, y = p2002)) +
  geom_smooth(method = lm, formula = y ~ poly(x, 4), se = F)
```




```
# equivalent:
#ggl %>%
#  ggplot(aes(x = yob, y = p2002)) +
#  geom_smooth(method = lm, formula = y ~ x + I(x^2) + I(x^3) + I(x^4), se = F)
```

5) Regress an p2002 on sex. Interpret the coefficients. What proportion of women in the sample voted in the 2002 primary?

```
lm(p2002 ~ sex, data = ggl) %>% summary()
```

```
##
## Call:
## lm(formula = p2002 ~ sex, data = ggl)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.3978 -0.3978 -0.3818  0.6021  0.6182
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.381804   0.001177  324.52   <2e-16 ***
## sexmale     0.016045   0.001663    9.65   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4876 on 344082 degrees of freedom
## Multiple R-squared:  0.0002706,  Adjusted R-squared:  0.0002677
## F-statistic: 93.13 on 1 and 344082 DF,  p-value: < 2.2e-16
```

The proportion of women who voted in the 2002 primary is about .382.

6) Regress `voted` on `sex`, `p2004`, and their interaction. Interpret the coefficient on `p2004`. Among men who voted in the 2004 primary, what is the proportion who voted in the 2006 primary? Show how to get that number from the regression coefficients.

```
lm(voted ~ sex*p2004, data = ggl) %>% summary()
```

```
##
## Call:
## lm(formula = voted ~ sex * p2004, data = ggl)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.4171 -0.2577 -0.2501  0.5829  0.7499
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.250083   0.001429 175.009  < 2e-16 ***
## sexmale       0.007661   0.002021   3.790 0.000151 ***
## p2004         0.149380   0.002258  66.147  < 2e-16 ***
## sexmale:p2004 0.009946   0.003190   3.118 0.001822 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4586 on 344080 degrees of freedom
## Multiple R-squared:  0.02669,    Adjusted R-squared:  0.02668
## F-statistic:  3145 on 3 and 344080 DF,  p-value: < 2.2e-16
```

The proportion of men voting in the 2006 primary, given that they voted in the 2004 primary, is the sum of all four cofficients:

```
sum(coef(lm(voted ~ sex * p2004, data = ggl)))
```

```
## [1] 0.4170691
```

7) Regress `p2004` on `treatment`. Show the `R` output. What does the coefficient on `treatmentCivicDuty` mean?

```
lm(p2004 ~ treatment, data = ggl)
```

```
##
## Call:
## lm(formula = p2004 ~ treatment, data = ggl)
##
## Coefficients:
##        (Intercept)  treatmentCivicDuty  treatmentHawthorne  treatmentNeighbors
##          0.4003388          -0.0008935           0.0028912           0.0063259
##      treatmentSelf
##          0.0021417
```

The coefficient on `treatmentCivicDuty` *tells us the difference in the proportion (or probability of) voting in the 2004 primary between people who in the "Civic Duty" treatment group and people in the control group. It should be small given that these groups are randomly determined and the 2004 primary happened before the treatment.*

8) Now regress `p2004` on `treatment` without an intercept (hint: add `-1` to the regression formula). Show the `R` output. Now what does the coefficient on `treatmentCivicDuty` mean?

```
lm(p2004 ~ treatment - 1, data = ggl)
```

```
##
## Call:
## lm(formula = p2004 ~ treatment - 1, data = ggl)
##
## Coefficients:
##   treatmentControl  treatmentCivicDuty  treatmentHawthorne  treatmentNeighbors
##             0.4003              0.3994              0.4032              0.4067
##      treatmentSelf
##             0.4025
```

The coefficient now means the proportion of people in the Civic Duty treatment group who voted in the 2004 primary.

9) Regress `voted` on `treatment`. What does the coefficient on `treatmentCivicDuty` mean?

```
lm(voted ~ treatment, data = ggl)
```

```
##
## Call:
## lm(formula = voted ~ treatment, data = ggl)
##
## Coefficients:
##        (Intercept)  treatmentCivicDuty  treatmentHawthorne  treatmentNeighbors
##            0.29664             0.01790             0.02574             0.08131
##      treatmentSelf
##            0.04851
```

This is the difference in the proportion (probability of) voting in 2006 between the Civic Duty group and the control group. Because this is an experiment, this is an unbiased estimate of the effect of the Civic Duty treatment compared to control: this treatment increased the probability of voting by about .018.

10) Add a fourth-degree polynomial of `yob` to the regression. Do any of the coefficients from the previous regression change much?

```
lm(voted ~ treatment + poly(yob, 4), data = ggl)
```

```
##
## Call:
## lm(formula = voted ~ treatment + poly(yob, 4), data = ggl)
##
## Coefficients:
##        (Intercept)  treatmentCivicDuty  treatmentHawthorne  treatmentNeighbors
##            0.29654             0.01865             0.02594             0.08141
##      treatmentSelf       poly(yob, 4)1       poly(yob, 4)2       poly(yob, 4)3
##            0.04835           -35.76682           -13.91142             5.22881
##      poly(yob, 4)4
##           -9.98646
```

No, and this shouldn't be surprising because the treatment should be uncorrelated with age. The omitted variable bias formula (discussed in week 6) will give a more technical perspective on this.