Problem set 5

Siwei Dai

October 29, 2021

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).

```r
# your code here
df <- ggl %>%
    group_by(yob) %>%
    summarize(turnout_yob = mean(p2002))

plot1 <- df %>%
    ggplot(aes(x = yob, y = turnout_yob)) +
    geom_point() +
    labs(x = 'Year of birth',
         y = 'Proportion of voting',
         title = 'Proportion of voting in 2002 primary by year of birth')
plot1
```

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.

```r
# your code here
model2 <- lm(formula = p2002 ~ yob, data = ggl) %>%
    summary()
model2
```

```
##
## Call:
```

```
## lm(formula = p2002 ~ yob, data = ggl)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.6590 -0.4004 -0.3191  0.5804  0.7432
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.755e+00  1.114e-01   87.54   <2e-16 ***
## yob         -4.787e-03  5.696e-05  -84.04   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4828 on 344082 degrees of freedom
## Multiple R-squared:  0.02011,    Adjusted R-squared:  0.02011
## F-statistic:  7063 on 1 and 344082 DF,  p-value: < 2.2e-16
```

The slope coefficient means the model predicts that the group mean of p2002 decreases -4.787e-03 for every year increase in birth year; in other words, every year increase in birth year is associated with -4.787e-03 decrease in the proportion of people born that year and voted in 2002 primary election.

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

```r
# your code here
plot3 <- ggl %>%
    ggplot(aes(x = yob, y = p2002)) +
    geom_point() +
    geom_smooth(method = 'lm', formula = y ~ x) +
    labs(x = 'Year of birth',
         y = 'Voting in 2002 primary',
         title = 'Proportion of people voted in 2002 primary election by their birth year')
plot3
```

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`).

```
# your code here
model4 <- lm(formula = p2002 ~ poly(yob, degree = 4), data = ggl) %>%
    summary()

plot4 <- ggl %>%
    ggplot(aes(x = yob, y = p2002)) +
    geom_point() +
    geom_smooth(method = 'lm',
                formula = y ~ poly(x, degree = 4),
                se = F) +
    labs(x = 'Year of birth',
         y = 'Voting in 2002 primary',
         title = 'Proportion of people voted in 2002 primary election by their birth year',
         subtitle = '4th-degree polynomial')

model4$coefficients
```

```
##                            Estimate   Std. Error   t value      Pr(>|t|)
## (Intercept)               0.3898379 0.0008184516 476.31152  0.000000e+00
## poly(yob, degree = 4)1 -40.5747753 0.4800928825 -84.51443  0.000000e+00
## poly(yob, degree = 4)2 -17.2646320 0.4800928825 -35.96102 1.146010e-282
```

```
## poly(yob, degree = 4)3 -19.2764290 0.4800928825 -40.15146  0.000000e+00
## poly(yob, degree = 4)4 -14.9285501 0.4800928825 -31.09513 5.529282e-212
```

```
plot4
```

The coefficient on first-power term is -40.5747753; coefficient for the second-power term is -17.2646320; coefficient for the third-power term is -19.2764290; the coefficient for the 4th-power term is -14.9285501.

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

```
# your code here
model5 <- lm(formula = p2002 ~ sex, data = ggl) %>%
    summary()
model5
```

```
##
## 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 intercept means the probability of women voting in 2002 primary is predicted to be 0.381804; in other words, the proportion of women voted in 2002 primary is 0.381804. The coefficient on 'sexmale' means the predicted probability of men voting in the 2002 primary is higher than the predicted probability of women voting in the 2002 primary by 0.016045.

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.

```r
# your code here
model6 <- lm(formula = voted ~ sex*p2004, data = ggl) %>%
    summary()
model6
```

```
##
## 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 coefficient on p2004 means if we control for the individual's gender and the interactive effect of whether voted in the 2004 primary on voting in 2006, conditional on the individual's gender, voting in the 2004 primary is associated with a 0.149380 increase in the predicted probability of voting in 2006.

The proportion of men who voted in 2004 voted in 2006 is the sum of all coefficient estimates, 0.250083 + 0.007661 + 0.149380 + 0.009946 = 0.41707, because it denotes the predicted value of 'voted' when $sexmale == 1$ **and** $p2004 == 1$ **and** $sexmale * p2004 == 1$ **.**

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

```
# your code here
model7 <- lm(formula = p2004 ~ treatment, data = ggl)
model7
```

```
##
## 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` means compared to the control group who didn't get mailing, the proportion of those received the CivicDuty mail treatment voted in 2004 primary election is lower by 0.0008935

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?

```
# your code here
model8 <- lm(formula = p2004 ~ treatment -1, data = ggl)
model8
```

```
##
## 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 means that the group mean of p2004 for those who received CivicDuty mailing is 0.3994; in other words, the proportion of those who received CivicDuty mailing treatment voted in 2004 primary election is 0.3994

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

```
# your code here
model9 <- lm(formula = voted ~ treatment, data = ggl)
model9
```

```
##
## Call:
## lm(formula = voted ~ treatment, data = ggl)
```

```
##
## Coefficients:
##        (Intercept)  treatmentCivicDuty  treatmentHawthorne  treatmentNeighbors
##            0.29664             0.01790             0.02574             0.08131
##      treatmentSelf
##            0.04851
```

The coefficient means compared to the control group who didn't receive any mailing, the proportion of those who received CivicDuty mailing voted in the 2006 election is higher by 0.01790

10) Add a fourth-degree polynomial of `yob` to the regression. You should find that none of the coefficients from the previous regression change much. Explain why this is not surprising.

```r
# your code here
model10 <- lm(voted ~ treatment + poly(yob, degree = 4), data = ggl)
model10
```

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

If the year of birth has no correlation with what treatment a particular individual receives, as in this case where the treatment assignment is random, it wouldn't interfere with the estimate on the treatment effect