Problem Set 4

Due November 12th, 2020

1. (15 pt) Sampling distributions of coefficients

Download the `x.csv` dataset from `GitHub`. This data contains a set of fixed values for an independent variable X. Consider the following population regression model:

$$y_i = 3 + 5x_i + \epsilon$$
$$\epsilon_i \sim \mathcal{N}(0, 1)$$

In this situation we know the true population parameters $\beta_0 = 3$, $\beta_1 = 5$ and $\sigma_\epsilon^2 = 1$

a. Simulate the sampling distributions for β_0 and β_1 by doing the following steps $m = 1000$ times

1. Generate random errors ϵ from the $\mathcal{N}(0, 1)$ distribution.
2. Generate values for y using ϵ, the fixed x, and the true population parameters.
3. Run a regression of y on x
4. Record your OLS estimates for β_0 and β_1, as well as the standard errors for each coefficient.
5. Repeat

At the end of this process, you should have m draws of β_0 and β_1 which serve as draws from your sampling distributions. Plot the density of your two sampling distributions. Superimpose a line on each for the mean of the distributions. From your simulations, does the OLS estimator appear to be unbiased? Do the standard errors you get in the regressions match the standard deviation of the sampling distribution?

b. Repeat the previous part, except in this case, generate ϵ from an exponential distribution with $rate = 0.5$ (use `rexp`). You will need to subtract 2 from each ϵ to ensure it has expectation of zero. Do this for $n = 4$ (use only the first four observations for x) and $n = 1000$. What do you notice as n increases? Why is this happening?

c. BONUS (5 pt) Create a set of values for ϵ such that your errors will be heteroskedastic (but still meet the zero mean assumption). Plot your errors against the x values to demonstrate that your errors are indeed heteroskedastic. Repeat the steps from the previous part using this new ϵ. Describe your results. Why do you think you are getting results like this?

2. (15 pt) Voter turnout experiment

For the next **three** problems, we will work with the original version of the Gerber, Green and Larimer data introduced briefly in class. In this experiment, the authors attempted to uncover why people vote, despite the fact that the casting of a single vote is of no significance where there is a multitude of electors. One hypothesis is adherence to social norms; voting is widely regarded as a civic duty and citizens worry that others will think less of them if they fail to participate in elections. According to this theory, voters may receive two different types of utility from voting; (a) the intrinsic rewards that voters obtain from performing this duty and (b) the extrinsic rewards that voters receive when others observe them doing so.

To gauge the effects of priming intrinsic motives and applying varying degrees of extrinsic pressure on voting behavior, Gerber, Green, and Larimer conducted a field experiment in Michigan prior to the August 2006 primary election. The sample for the experiment was $344,084$ voters in the state of Michigan. Voters were randomly assigned to either the control group $(N_c = 191,243)$ or one of four treatment groups. The researchers sent one of four mailings to each voter in treatment group.

1. All four treatments carry the message "DO YOUR CIVIC DUTY - VOTE!" The first type of mailing (`Civic Duty`) provides a baseline for comparison with the other treatments.

2. The second mailing adds to this civic duty baseline a mild form of social pressure, in this case, observation by researchers. Households receiving the `Hawthorne` mailing were told "YOU ARE BEING STUDIED!" and informed that their voting behavior would be examined by means of public records.

3. The `Self` mailing exerts more social pressure by informing recipients that who votes is public information and listing the recent voting record of each registered voter in the household. The "Self" condition thus combines the external monitoring of the Hawthorne condition with actual disclosure of voting records.

4. The fourth mailing, `Neighbors`, ratchets up the social pressure even further by listing not only the household's voting records but also the voting records of those living nearby. By threatening to "publicize who does and does not vote," this treatment is designed to apply maximal social pressure.

The latter two treatments suggested that a follow-up mailing after the election would report to the household or the neighborhood the subject's turnout in the upcoming election.

For this problem set, we have provided you with the original data of Gerber et al. The data is available on the course GitHub (`gerber.csv`). Table 1 contains a list of the variable definitions in the dataset.

Table 1: Contents of gerber.csv

Variable	Content
female	gender (1 if female, 0 if male)
yob	year of birth
g2000	1 if Respondent voted in the 2000 General Election (0 otherwise)
g2002	1 if Respondent voted in the 2002 General Election (0 otherwise)
g2004	1 if Respondent voted in the 2004 General Election (0 otherwise)
p2000	1 if Respondent voted in the 2000 Primary Election (0 otherwise)
p2002	1 if Respondent voted in the 2002 Primary Election (0 otherwise)
p2004	1 if Respondent voted in the 2004 Primary Election (0 otherwise)
voting	1 if Respondent voted in the 2006 Primary Election (0 otherwise)
control	1 if Respondent is assigned to the control group (0 otherwise)
civicduty	1 if Respondent is assigned to the “Civic Duty” group (0 otherwise)
hawthorne	1 if Respondent is assigned to the “Hawthorne” group (0 otherwise)
self	1 if Respondent is assigned to the “Self” group (0 otherwise)
neighbors	1 if Respondent is assigned to the “Neighbors” group (0 otherwise)

a. Regress voting on the four treatment variables not adjusting for any of the other variables. Report the estimates, the standard errors and 95 percent confidence intervals (`confint()` might be helpful). Briefly interpret substantive and statistical significance of the estimates. Do you have a lot of confidence in these estimates? Why or why not? Discuss the plausibility of each of the regression assumptions required for causal statistical inference.

b. For the rest of the questions, drop all observations except the control and "Neighbors" groups. Run 5 regressions of `voting` on only `neighbors` and different combinations of the pre-treatment variables (feel free to use interactions or polynomials if you want). Provide a table with the specifications you used, your estimates of the treatment effect, and their associated standard error. Do your estimates of the effect of the `neighbors` treatment change across your five regressions? Why or why not?

3. (15 pt) Voter turnout and gender

We'll keep working with the Gerber data (again, drop all observations except the control and "Neighbors"groups). **To get full credit on this problem, you may NOT use packages beyond base R and the tidyverse!**

a. Regress `voting` on `neighbors` and an interaction term between `neighbors` and `female` (ie. two independent variables overall in the regression). Obtain and report four point estimates (no need to report standard errors) for the expected probabilities of voting for the four groups: unassigned males (ie. males in control group), unassigned females, assigned males (ie. males in treatment group), and assigned females in a 2 by 2 matrix. What is the average effect of treatment for males? For this model, what is the estimated average treatment effect for females?

b. Now, regress `voting` on `neighbors`, an interaction term between `neighbors` and `female` and `female` itself (ie. three independent variables overall in the regression). Obtain and report four point estimates (no need to report standard errors) for the expected probabilities of voting for the four groups: unassigned males, unassigned females, assigned males, and assigned females in a 2 by 2 matrix. What is the average effect of treatment for males? What is the average effect of treatment for females? Comparing b) and c), which model do you like better and why?

c. Based on the previous model, derive two analytic expressions (in terms of the various β's) for the marginal effect of treatment on women's probability of voting and for the effect of treatment on men's probability of voting and provide your estimates

d. Derive two analytic expressions for the variances of these estimates (again, these will be in terms of the various β's). Then, compute both variances, and report them in the form of standard errors. Hint: You may want to recall that if X and Z are two random variables, then $\mathbb{V}[X+Z] = \mathbb{V}[X] + \mathbb{V}[Z] + 2Cov[X,Z]$. You can obtain the covariances between each pair of the independent variables (in matrix form) using the `vcov()` command.

4. (15 pt) Voter turnout and age

We'll keep working with the Gerber data. **To get full credit on this problem, you may NOT use packages beyond base R and the tidyverse!**

a. Create the variables age and age squared using `yob` (ie. $2006 - yob = age$, $age * age = age_{squared}$). Imagine you were to regress `voting` on `neighbors`, `p2004`, `age`, and `age squared`. Without doing any calculation, briefly interpret the conceptual meaning of the coefficients on `age` and `age squared` in such a regression.

b. Now regress voting on `neighbors`, `p2004`, `age`, and `age squared`. Report the results in a standard regression table.

c. Based on the previous model, generate a plot that visualizes how expected probabilities of voting vary with age for a person who received the `neighbors` treatment and voted in the 2004 Primary Election. Use an age range that seems reasonable given your data. Briefly explain the results.

To do this:

1. Create an empty dataset.
2. Generate a variable with the hypothetical age range
3. Calculate probability of voting for each age.
4. Plot your results

d. Derive an analytic expression for the marginal "effect" of `age` on the probability of voting. Compute the marginal "effect" of `age` on the probability of voting at the following levels of age: 20 years, 30 years, 40 years, 50 years, 60 years, 70 years, 80 years. Report the results in a table. Briefly discuss the results.

e. Plot the marginal effect you found in the previous part and - for extra credit (5 pt) - add a 95% confidence interval around the line in your plot.