POL SCI 231b (Spring 2017):

Problem Set 7

Spring 2017

University of California, Berkeley

Prof. Thad Dunning/GSI Natalia Garbiras-Díaz

Due Monday, April 3rd, at 9:00 AM (before lecture)

Each group should turn in its problem set solution by email to Natalia Garbiras-Díaz (nataliagarbirasdiaz+231b@berkeley.edu) by Monday at 9 AM. Please work out the problems on your own, before you meet with your group to agree on solutions.

1. **Cluster randomization.** A researcher runs a clustered randomized experiment with 297 units divided into 15 clusters of unequal size. The code below generates potential outcomes that are associated with cluster membership (note that average potential outcomes differ for each cluster). It then simulates one realization of the experiment. The data created by this code are in `cluster.experiment.Rda` on bcourses.

   (a) What is the average causal effect $\frac{1}{297}[\sum_{i=1}^{297} Y_i(1) - \sum_{i=1}^{297} Y_i(0)]$?

   (b) Use data from the realization of the experiment to estimate the average causal effect, using the difference of cluster averages.

   Thus, following Lecture 7, let the average outcome in the $A = 8$ clusters assigned to treatment be $Y^T = \frac{1}{A}\sum_{k \in A} \overline{Y}_k(1)$, where $\overline{Y}_k(1)$ is the average in the kth cluster, and let the average outcome in the $B = 7$ clusters assigned to control be $Y^C = \frac{1}{B}\sum_{k \in B} \overline{Y}_k(0)$. The difference of the cluster averages is $\widehat{\text{ATE}}_{\text{cluster}} = Y^T - Y^C$.

   It may be helpful to use the `ddply` command in the `plyr` package to calculate cluster means.

(c) Calculate the standard error of $\widehat{\text{ATE}}_{cluster}$, using the approach in Lecture 7. That is, use the conservative Neyman estimator of the variance, where the elements of the formula are the cluster averages defined in part (b): $\text{Var}(Y^T - Y^C) = \text{Var}(Y^T) + \text{Var}(Y^C)$, where Var refers to the variance of the sampling distribution of the estimators.

Now, use the ratio of $\widehat{\text{ATE}}_{\text{cluster}}$ and the estimated standard error to conduct a t-test and calculate a p-value, that is, the probability that you would observe the data from the realized experiment under the null hypothesis that $\text{ATE} = 0$

(d) Suppose you erroneously assumed that the data were produced by individual rather than cluster randomization. What standard error, t ratio, and p value would you calculate?

(e) Now, use a simulation to create 10,000 realizations of the experiment and construct a plot of the sampling distribution of $\widehat{\text{ATE}}_{cluster}$. In addition,

- For each of the replicates, calculate the nominal standard error, t-ratio, and p-value as in part (c). Also calculate the nominal standard error, t-ratio, and p-value that you would calculate if you erroneously assumed that the data were produced by individual rather than cluster randomization, as in part (d). Compare the standard deviation of the 10,000 $\widehat{\text{ATE}}_{cluster}$s to the averages of the nominal standard errors constructed as per parts (c) and (d). What do you conclude?
- Compare the mean of the bootstrap $\widehat{\text{ATE}}_{cluster}$s to the ATE from (a). What do you conclude?

See the data generating process below for details on the sampling procedure.

```
set.seed(94705)
cluster <- c(rep(1:4, each=5), rep(5:8, each=18),
             rep(9:12, each=25), rep(13:15, each=35))
y0 <- rnorm(length(cluster), sqrt(cluster), .1)
y1 <- y0 + rnorm(length(cluster), sqrt(cluster)/4, 1)
treat_cluster <- sample(1:15, 8)
treat <- as.numeric(cluster %in% treat_cluster)
yobs <- ifelse(treat==1, y1, y0)
data <- as.data.frame(cbind(y1, y0, cluster, treat, yobs))
```

(f) Double the size of the experiment by creating a new dataset that combines two copies of the original data (but changes the cluster names such that there are now 30 clusters). Repeat the exercise in (e) with this new dataset. Sample code to duplicate the dataset:

```
data2 <- data
data2$cluster <- data2$cluster + 15
data2 <- rbind(data, data2)
```

(g) Repeat (f), now using a data set duplicating the dataset with 30 clusters. Sample code to triplicate the dataset:

```
data3 <- data
data3$cluster <- data3$cluster + 30
data3 <- rbind(data2, data3)
```

(h) What do you conclude from your results in this exercise?

2. **Maximum Likelihood Estimation**

(a) **Likelihood Function.** Suppose that we observe the following independent data:

```
y <- c(10,4,5,3,9,2,7,3,6,4)
```

Plot the log-likelihood function using the Poisson distribution (hint: how do we get the likelihood function of independent data?). Recall that a Poisson distribution describes a discrete variable X using a parameter λ such that for $k = 0, 1, 2, ...$

$$Pr(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

The positive number λ is equal to the expected value of X and also its variance.

$$\lambda = E(X) = var(X)$$

Looking at the plot, what does the maximum appear to be? Here is some code that might help you.

```
# 1. write a function for the log-likelihood
# with lambda as the input
log.lik <- function(lambda) { LOG LIKELIHOOD FUNCTION }

# 2. generate a vector with the possible values of lambda
p.l <- seq(0,30,0.1)

# 3. plot the log likelihood function
plot(p.l, lapply(p.l, log.lik))
```

Now use R to solve for the MLE. There are a variety of commands you can use (you can check optim or optimize, for example). Below is some example code using optim.

```
mle <- optim(
  c(1), # starting value to look for max

  log.lik, # the function we are optimizing

  control=list(fnscale=-1), # this tells optim to maximize (vs minimize)
  lower=0, # lower bound for optimization
  upper=100, # upper bound

  method="Brent" # one dimensional optimization method
)

mle
```

Now compare the MLE estimate of λ to the mean of y. Interpret your results.

(b) **Logistic regression.** A scholar is interested in evaluating the effect of $X \in \mathbb{R}$ (X is continuous from $-\infty$ to ∞) on the probability that $Y = 1$. (Y is dichotomous). The scholar estimates two different logistic regression models M_1 and M_2.

M_1:

$$\text{Prob}(Y_i = 1) = \Lambda(\alpha_1 + \beta_{X1} X_i + \gamma_1 Z_{i,M_1}) \tag{1}$$

and

M_2:

$$\text{Prob}(Y_i = 1) = \Lambda(\alpha_2 + \beta_{X2} X_i + \gamma_2 Z_{i,M_2}). \tag{2}$$

Here, Λ is the logistic distribution function. There are possibly different covariates in each model: Z_{i,M_1} in the first model and Z_{i,M_2} in the second model. Also, γ_j and Z_{i,M_j} may be vectors. X is the same variable in models $M1$ and $M2$.

Now, suppose the scholar first estimates logistic regression model M1; this yields an estimated coefficient for X of $\hat{\beta}_{X1} = 0.3$. The scholar then estimates the second model M2; this yields an estimated coefficient for X of $\hat{\beta}_{X2} = 0.8$.

For the following questions, we are setting aside the more vexing problems for causal inference—like, both models can't be simultaneously true. Just assume the models in each case and answer the questions, taking each model as given:

i. For each model $M1$ and $M2$, express the marginal effect of a one-unit increase in X on the probability that $Y_i = 1$ in terms of the models in equations (1) and (2), respectively.

ii. The M2 estimate implies that the estimated effect on $Pr(Y = 1)$ of increasing X by one unit is greater than is implied by the M1 estimate. True or false? Explain your answer.

iii. Holding the baseline probability of Y constant, the M2 estimate implies that the estimated effect on $Pr(Y = 1)$ of increasing X by one unit is greater than is implied by the M1 estimate. True or false? Explain your answer.

iv. For any particular unit (say India-Pakistan-1986), the M2 estimate implies that the estimated effect on $Pr(Y = 1)$ of increasing X by one unit is greater than is implied by the M1 estimate. True or false? Explain your answer.

(c) **Probit Model and Interpretation.** For this exercise you will need the `camp1.dta` dataset available on bcourses.[1] Consider the probit model with

[1] You will need to use the `foreign` library to open the dataset.

democratic win (DWIN) as the dependent variable, and the following independent variables:

- JULYECQ2: 2nd Quarter GNP Growth
- PRESINC: Elected Incumbent Seeking Reelection (1=Democrat, 0=no incumbent, 1-=Republican)
- ADAACA: State Liberalism Index (ADA & ACA)
- An interaction term between PRESINC and JULYECQ2.

We have the following hypotheses:
H1: *Economic growth in the months prior to the election increases the chances that the Democrat will win.*
H2: *The Democrat has a better chance of winning if he/she is the incumbent President.*
H3: *The more liberal the state, the more likely the Democrat will win.*
H4: *Growth prior to the election only helps the Democrat if he or she is the incumbent.*

Run the analysis in R including an intercept in the model. You can write your own function or use *glm* in R.

```
glm(y ~ x, family = binomial(link = "probit"))
```

i. Suppose you could manipulate $ADAACA$, the state liberalism index. (Just suppose). What is the estimated effect of that variable on a Democratic win, holding all other variables at their median values in the dataset?
ii. Arbitrarily define two levels (low and high) for the variable measuring growth, JULYECQ2, so that 1.5 is “high" growth and -1.5 is “low" growth.
Consider the following four scenarios:
1) High growth, incumbent Democratic candidate running.
2) High growth, incumbent candidate not running.
3) Low growth, incumbent Democratic candidate running.
4) Low growth, incumbent candidate not running.

Across the range of values of ADAACA in the dataset, plot the predicted value of Prob(DWIN = 1) for each of the four scenarios. (You can use the predict() function in R to generate their respective predicted values). Generate two figures: one with the plots for scenarios 1 and 3, and the other with those for scenarios 2 and 4. Please provide a substantive discussion of your results in light of the above hypotheses.