POLSCI 787 | | **Due Date:**
Fall 2017 **Problem Set 5** **Friday December 8**

- You may work in (small) groups while solving this assignment.
- Submit individual solutions via Canvas in one single report collecting everything (e.g., derivations, figures, tables, computer code). Use `RMarkdown` to create this report in PDF format.
- Please provide written explanations whenever those are requested; reporting only code and output automatically leads to $\checkmark^{-}$.

Question 1

(a) Let X and Y be two any random variables with realizations x and y, respectively. We use $\mathbb{E}[X]$, $\mathbb{V}[X]$, and $\mathbb{C}\text{ov}[X, Y]$ to denote expectation, variance and covariance, respectively.

Let $\theta \equiv (\theta_x, \theta_y)$ be a particular point. For any real, bivariate, and twice continuously differentiable function $f(x, y)$, the second-order Taylor expansion around θ is:

$$\begin{aligned} f(x,y) \;&=\; f(\theta) + f'_x(\theta)(x-\theta_x) + f'_y(\theta)(y-\theta_y) + \\ &\frac{1}{2}\left\{f''_{xx}(\theta)(x-\theta_x)^2 + 2f''_{xy}(\theta)(x-\theta_x)(y-\theta_y) + f''_{yy}(\theta)(y-\theta_y)^2\right\} + \texttt{remainder} \end{aligned} \tag{1}$$

Use this expansion to show that we can approximate $\mathbb{E}[f(X, Y)]$ around $\theta = (\mathbb{E}[X], \mathbb{E}[Y])$ as follows

$$\mathbb{E}[f(X,Y)] \approx f(\theta) + \frac{1}{2}\left\{f''_{xx}(\theta)\mathbb{V}[X] + 2f''_{xy}(\theta)\mathbb{C}\text{ov}[X,Y] + f''_{yy}(\theta)\mathbb{V}[Y]\right\}$$

(b) Consider two random variables, R and S, and define their ratio $G = g(R, S) = R/S$. (We assume that dividing by zero is not a problem, so we assume that S has no mass at zero (if it is discrete) or its support is $(0, \infty)$ if it's continuous.)

Use the result in 1(a) to obtain the following result, where the expansion is made around $\theta = (\mathbb{E}[R], \mathbb{E}[S])$:

$$\mathbb{E}\left[\frac{R}{S}\right] \equiv \mathbb{E}[g(R,S)] \approx \frac{\mathbb{E}[R]}{\mathbb{E}[S]} - \frac{\mathbb{C}\text{ov}[R,S]}{(\mathbb{E}[S])^2} + \frac{\mathbb{V}[S]\mathbb{E}[R]}{(\mathbb{E}[S])^3}$$

Question 2

Let Y be an outcome of interest, Z be the binary instrument and D be the binary endogenous variable. We have a sample $\{Y_i, D_i, Z_i\}_{i=1}^{n}$. We assume conditions A1-A5 in Angrist, Imbens, and Rubin (1996) hold. As we saw in class, under those assumptions, the LATE can be expressed as a function of observable variables only:

$$\tau_{\texttt{LATE}} \equiv \mathbb{E}\left[Y_i(1) - Y_i(0) \ \mid \ (D_i(1) - D_i(0)) = 1\right] = \frac{\mathbb{E}\left[Y_i(1, D_i(1)) - Y_i(0, D_i(0))\right]}{\mathbb{E}\left[D_i(1) - D_i(0)\right]}$$

Consider the estimated intention-to-treat (ITT) effects of Z_i on Y_i and Z_i on D_i, and their ratio:

$$\begin{aligned}
\widehat{\tau}^{ZD} &= \frac{\sum_i D_i Z_i}{\sum_i Z_i} - \frac{\sum_i D_i(1 - Z_i)}{\sum_i (1 - Z_i)} \\
\widehat{\tau}^{ZY} &= \frac{\sum_i Y_i Z_i}{\sum_i Z_i} - \frac{\sum_i Y_i(1 - Z_i)}{\sum_i (1 - Z_i)} \\
\widehat{\tau}^{IV} &= \frac{\widehat{\tau}^{ZY}}{\widehat{\tau}^{ZD}}
\end{aligned}$$

(a) Is $\widehat{\tau}^{IV}$ an unbiased estimator of $\tau_{\texttt{LATE}}$? Formally show your answer. The result you derived in question 1 above should be helpful to answer this question (but it's OK to use other results in your answer if you want).

Question 3

Hahn, Todd, and van der Klaauw (2001) proved non-parametric identification of the RD treatment effect in Theorems 1 and 2. In their notation, the binary treatment is x, the score or running variable is z, the cutoff is z_0, the potential outcomes are y_0,y_1 and the observed outcome is y. In Theorem 2, the "individual treatment effect", $y_{1i} - y_{0i}$, is allowed to be different across i.[1] Theorem 2 relies on the following main assumptions:

- Assumption RD (page 202): (i) The limits $x^+ \equiv \lim_{z\downarrow z_0} \mathbb{E}[x_i|z_i = z]$ and $x^- \equiv \lim_{z\uparrow z_0} \mathbb{E}[x_i|z_i = z]$ exist; (ii) $x^+ \neq x^-$.
- Assumption A1 (page 202): $\mathbb{E}[y_{0i}|z_i = z]$ is continuous in z at z_0.
- Assumption A2 (page 203): $\mathbb{E}[y_{1i} - y_{0i}|z_i = z]$, regarded as a function of z, is continuous in z at z_0.
- Assumption of local conditional independence (page 203, inside Theorem 2): x_i is independent of $y_{1i} - y_{0i}$ conditional on z_i near the cutoff z_0.

Under these assumptions, the authors show that:

$$\mathbb{E}[y_{1i} - y_{0i}|z_i = z_0] = \frac{\lim_{z\downarrow z_0} \mathbb{E}[y_i|z_i = z] - \lim_{z\uparrow z_0} \mathbb{E}[y_i|z_i = z]}{\lim_{z\downarrow z_0} \mathbb{E}[x_i|z_i = z] - \lim_{z\uparrow z_0} \mathbb{E}[x_i|z_i = z]} \tag{7}$$

[1] This is in contrast to Theorem 1, where the authors make the strong assumption that $\beta_i \equiv y_{1i} - y_{0i} = \beta$.

In this question, you will prove the conclusion of Theorem 2 using different assumptions.

(a) First, show the particular form that Assumption RD takes when the RD design is sharp.

(b) Prove the result in Equation (7) using the following assumptions: (i) the RD design is sharp, (ii) Assumption A1, and (iii) Assumption A2.

(c) Explain (i) why you were able to prove Equation 7 without the local independence assumption; and (ii) why you need the local independence assumption to prove Equation 7 when the RD design is fuzzy instead of sharp.

Question 4

This exercise asks you to re-analyze the data from Caughey and Sekhon (2011). In this RD design, the unit of analysis is U.S. House congressional districts, and we are interested in the effect of the Democratic party barely winning election t on the probability of the Democratic winning the following election, which occurs at $t+1$. The period is 1942-2010. The running variable is the Democratic party margin of victory. There are also eight predetermined covariates that you are asked to analyze to investigate the plausibility of the RD assumptions.

The variables of interest are the following:

```
DifDPct: Running variable/score ==> Democratic Party's Margin of Victory at Election t
DWinNxt: Outcome of interest ==> Democratic Party's Victory at Election t+1
DWinPrv : Covariate 1 ==> Democratic Party's Victory at Election t-1
DifDPPrv: Covariate 2 ==> Democratic Party's Margin of Victory at Election t-1
DemInc  : Covariate 3 ==> Whether there is a Democratic incumbent in t race
NonDInc : Covariate 4 ==> Whether there is a non-Democratic incumbent in t race
PrvTrmsD: Covariate 5 ==> No. of previous terms served by Democrat in t race
PrvTrmsO: Covariate 6 ==> No. of previous terms served by Republican in the t race
DSpndPct: Covariate 7 ==> Democratic Percentage of Money Spent in the t race
DDonaPct: Covariate 8 ==> Democratic Percentage of Donations in the t race
```

(a) Analyze this RD design empirically using a randomization-based approach, assuming that the window in the support of the running variable `DifDPct` in which the treatment was randomly assigned is $[-0.50, 0.50]$—that is, races decided by a margin of victory equal to or less than 0.50 percentages points. You are free to chose a Neyman or Fisherian approach in this window.

The analysis should consist of at least the following steps:

- Graphical analysis of outcome and covariates with RD plots
- Formal Falsification: density tests and covariate tests
- Estimation and inference of RD effect on main outcome of interest.

(b) Analyze this RD design empirically using a continuity-based approach. Use local polynomial methods with mean-squared-error optimal bandwidth choices. The analysis should consist of at least the following steps:

- Graphical analysis of outcome and covariates with RD plots
- Formal Falsification: density tests and covariate tests
- Estimation and inference of RD effect on main outcome of interest.

(c) Based on your randomization-based analysis, do you see evidence that this RD design is invalid? What does invalid mean exactly in this approach, what assumption seems to be failing?

(d) Based on your continuity-based analysis, do you see evidence that this RD design is invalid? What does invalid mean exactly in this approach, what assumption seems to be failing?

(e) Do you reach different conclusions regarding the validity of the design from both approaches or the same? If you reach different conclusions, do you think one conclusion is more plausible than the other? Why? Do you think the Democratic Party incumbency advantage estimated with this RD design is credible? Why/Why not?

References

Angrist, Joshua D, Guido W Imbens, and Donald B Rubin. 1996. "Identification of causal effects using instrumental variables." *Journal of the American statistical Association* 91 (434): 444–455.

Cattaneo, Matias D., Michael Jansson, and Xinwei Ma. 2016. "Simple Local Regression Distribution Estimators with an Application to Manipulation Testing." working paper, University of Michigan.

Caughey, Devin, and Jasjeet S Sekhon. 2011. "Elections and the regression discontinuity design: Lessons from close US house races, 1942–2008." *Political Analysis* 19 (4): 385–408.

Hahn, Jinyong, Petra Todd, and Wilbert van der Klaauw. 2001. "Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design." *Econometrica* 69 (1): 201–209.