Microeconometrics

Problem set 2 - Solutions

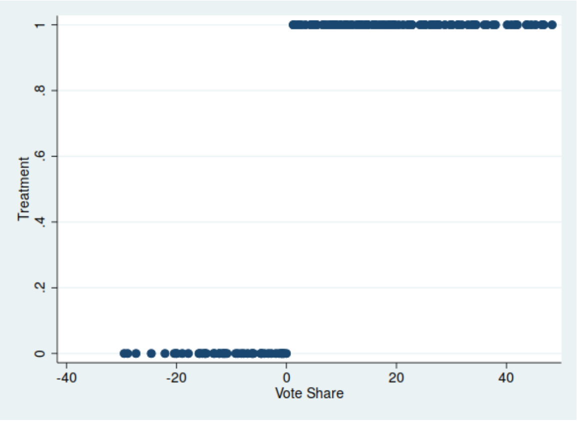
1. (50%) **Regression Discontinuity.** De Kadt and Rosenzweig investigate partisan ties and their effects on national resource allocation in Ghana. The specific research question of interest is whether constituencies in Ghana that elect MPs who are from the same party as the President (the "ruling party") receive more electrification over the next four years. Using electoral data from the 1996 parliamentary elections and nightlights data the authors use a sharp regression discontintuity (RD) design to investigate the effect of the treatment (an MP from the ruling party winning the 1996 election) on the outcome (change in nightlights over the next four years). The forcing variable the authors use is voteshare of the ruling party MP candidate. The unit of analysis is the constituency. In this problem you will similarly conduct a sharp RD analysis using the dataset **votes_Ghana**. The dataset contains 152 observations each corresponding to a constituency in Ghana, with the following variables:

   - *constit*: name of consistuency
   - *voteshare*: voteshare (% votes obtained) for the ruling party MP candidate
   - *treatment*: treatment status indicator (1 if the ruling party MP won the 1996 election in that constituency, 0 if the ruling party candidate lost)
   - *changeNL_1996_2000*: change in nightlights between 1996 and 2000
   - *mean_1995*: mean level of nightlights in 1995

   (a) Look at the data to see if sharp RD makes sense for our dataset. To do so, plot treament (y- axis) as a function of the forcing variable (x-axis), where the forcing variable is the margin of victory/loss for the ruling party MP candidate. Does it seem appropriate to use a sharp regression discontinuity design in this case? [*HINT:* to do the plot you can use the STATA command `twoway scatter`. For the rest of the exercise remember that some commands requires the definition of the cut-off value or for simplicity they require that the forcing variable has a discontinuity at zero.]

   SOLUTION: in STATA, you can generate Figure 1 using the command: `twoway scatter treatment voteshare`. Figure 1 reveals that our forcing variable does fully determine the treatment status. Ruling party MP candidates who receive greater than 50% of the voteshare do indeed win the election and those who do not lose. It seems that a sharp regression discontinuity design is appropriate.

Figure 1: Sharp discontinuity



(b) Estimate the local average treatment effect at the threshold. What are assumptions required for this estimation strategy? Interpret your resulting estimate.

SOLUTION: you need to assume **smoothness**, i.e. the potential outcomes $E[y_{i0}|S_i]$ and $E[y_{i1}|S_i]$ are continuous on S at $\underline{S}$. In STATA, You can estimate the local average treatment effect with the following commands:

```
gen adjvoteshare = voteshare - 50
reg changenl_1996_2000 treatment adjvoteshare
```

The point estimate is -0.582, which we can interpret as the Local Average Treatment Effect (LATE) of D on Y. Interpreting this estimate we see that moving from an opposition-won constituency to a ruling party-won constituency in the 1996 election results in a 0.58 reduction in nightlights. The estimate is statistically significant at the 10% level.

(c) Conduct the same analysis as in part (b) except that you should now use a *linear* model with different slopes for treated and control units [*HINT*: add the interaction between the forcing variable and the treatment variable]

SOLUTION: in STATA, you can estimate the local average treatment effect with the following commands:

```
gen treatvote = treatment * adjvoteshare
reg changenl_1996_2000 treatment adjvoteshare treatvote
```

The point estimate is essentially unchanged from the previous model. It is -0.63 and is statistically significantly at 10% significance level. In this case switching from an opposition MP to a ruling party MP candidate winning the 1996 election results in a 0.63 decrease in nightlights over the next four years.

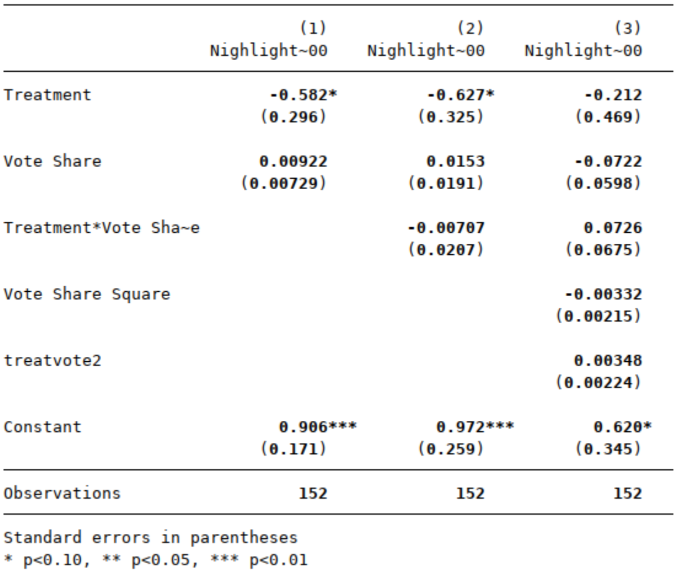
(d) Conduct the same analysis as in part (b) except that you should now use a *quadratic* model with different model coefficients for the treated and control groups.

SOLUTION: in STATA, you can estimate the local average treatment effect with the following commands:

```
gen vote2 = adjvoteshare^2
gen treatvote2 = treatment * vote2
reg changenl_1996_2000 treatment adjvoteshare treatvote vote2 treatvote2
```

The point estimate is now smaller (-0.21) with a very large p-value. In this case switching from an opposition MP to a ruling party MP candidate winning the 1996 election results in a 0.21 decrease in nightlights over the next four years.

Figure 2: Results section b), c), and d)



	(1) Nighlight~00	(2) Nighlight~00	(3) Nighlight~00
Treatment	-0.582* (0.296)	-0.627* (0.325)	-0.212 (0.469)
Vote Share	0.00922 (0.00729)	0.0153 (0.0191)	-0.0722 (0.0598)
Treatment*Vote Sha~e		-0.00707 (0.0207)	0.0726 (0.0675)
Vote Share Square			-0.00332 (0.00215)
treatvote2			0.00348 (0.00224)
Constant	0.906*** (0.171)	0.972*** (0.259)	0.620* (0.345)
Observations	152	152	152

Standard errors in parentheses
* p<0.10, ** p<0.05, *** p<0.01

(e) Use the `rd` command in STATA (or similar in other softwares) to estimate the Local Average Treatment Effect at the threshold using a local linear regression with a triangular kernel. Report your estimate with the s.e. and discuss your result in light of points (b), (c) and (d) [*HINT:* use the stata command `rd` with the default options, report the coefficient and the s.e. using a bandwidth of 100 percent]

SOLUTION: in STATA, you can use the following command:

```
rd changenl_1996_2000 adjvoteshare, gr mbw(100)
```

The LATE estimate is now positive, 0.302, and is not statistically significant, with a standard error of 0.260. This estimate suggests that switching from an opposition MP to a ruling party MP winning the 1996 election results in a 0.30 increase in nightlights over the next four years.

(f) How do the estimates of the LATE at the threshold differ based on your results from parts (b) to (e)? In other words, how robust are the results to different specifications of the regression? What other types of robustness checks might be appropriate?
SOLUTION: The estimates of the LATE seem to vary based on the functional form used.

The way we estimate the conditional expectation function of Y_1 and Y_0 is going to significantly impact the results we get. Recall that the LATE is only identified when X = c; this means that we really want to pay special attention to getting the conditional expectation function right in the local area around c. So we don't want to draw too heavily on the rest of the joint distribution. Local linear regression with a triangular kernel (the last estimator we use) is the current "best practice" for getting the conditional expectation right – this is the estimate we should trust since we did not limit the bandwidth around the threshold in our other estimation procedures. Other robustness checks we might want to conduct are examining continuity in covariates and the forcing variable around the threshold c. We could also conduct placebo regressions. We could also estimate the LATE for different sized bandwidths around the threshold.

(g) Finally conduct a placebo test using nightlights measured in 1995 ($mean_1995$) as the outcome in a sharp RD analysis for the 1996 election. Use local linear regression as you did in part (e). What does this placebo test say about the relationship between the 1996 election of ruling party MPs and nightlights measured in 1995?
SOLUTION: in STATA, you can use the following command:

```
rd mean_1995 adjvoteshare, gr mbw(100)
```

From this placebo regression we can see that there is little (statistically significant) relationship between the 1996 election close winners and losers from the ruling party and 1995 nightlights. The LATE estimate from the local linear regression is large but not statistically significant. Our large estimate may be due to outlying data points around the threshold.

2. (50%) **Panel data models.** Consider a subset of the data used by Vella and Verbeek (1998) "*Whose Wages do Unions Raise? A Dynamic Model of Unionism and Wage Rate Determination for Young Men*", **wagepan**, to estimate the effects of unions on workers' wages. The dataset is comprised of 545 men (identified by variable *nr*) who worked in every year from 1980 through 1987 (identified by variable *year*) in the United States. Consider the following model:

$$lwage_{it} = a_i + \theta_t D_t + \beta x_{it} + \pi z_i + u_{it}$$

where i denotes the worker and t the year. The vector x_{it} is comprised of *exper* (labor market experience) and its square *expersq*, *married* equals 1 if the individual is married, and *union* equals 1 if the worker is unionized. The vector z_i includes the variables *black* equals 1 for blacks, *hisp* equals 1 for hispanic workers, and *educ* denotes the number of years of education.

(a) Study the distribution of the variable *lwage*. Produce an histogram of the overall distribution and estimate the non-parametric density of the variable. Then compare the non-parametric density for black, hispanics and the remaining group ($black == 0$ and $hisp == 0$).

SOLUTION: Stata commands:
hist lwage
kdensity lwage
tw (kdensity lwage if black == 1 & hisp == 0)(kdensity lwage if black == 0 & hisp == 1)(kdensity lwage if black == 0 & hisp == 0)

(b) Explain which effects parameters θ_t and a_i are likely to capture.

SOLUTION: The coefficients of the time dummies (δ_t) capture the aggregate shocks in a specific year that affect wages of all individuals in that particular year.

The term α_i captures the time-invariant characteristics of a worker i that affect wages and might also be correlated with the explanatory variables. Omitting this term may lead to an omitted variable bias problem if the worker fixed effects are correlated with the explanatory variables.

(c) If unions are successful in their wage negotiations with employers, what should be the sign of β_{union}?

SOLUTION: The estimated coefficient $\widehat{\beta}_{union}$ is expected to be positive. Unionized workers earn more on average than non-unionized workers, ceteris paribus.

(d) Estimate the equation by pooled OLS. Do you find any evidence for a union effect? Are the assumptions required for these estimates to be consistent plausible? If not, what would be the asymptotic bias you would expect in the union estimate?

SOLUTION: According to the estimates, unionized workers earn on average approximately 18.25% more than non-unionized workers.

Stata command:
reg lwage educ black hisp exper expersq married union i.year, vce(robust)

The assumptions required for consistency are $E(u_{it}|x_{it}, a_i)$ and $Cov(x_{it}, a_i) = 0$, where $\mathbf{x}_{it}$ includes the explanatory variables in the estimated equation. In this case it is unlikely to hold. The classical example is worker's ability that affects wages and are likely to be correlated with the explanatory variables.

(e) Considering the time-varying variables, estimate the equation using the within (FE) estimators.[1] What are the necessary assumptions for consistency of the estimator? Can we estimate the returns to education? Why? What about race effects and experience?

SOLUTION: The variable educ is redundant because it assumes the same range of values for all individuals in the sample. This happens due to the deterministic nature of the variable. The same is true for the race variable. Because they are time-invariant, their coefficients are not identified in the fixed effects model. In contrast, exper has within variation and therefore we can estimate experience effects.

The assumptions for consistency of the FE estimator are FE.1 $E(\ddot{x}'_{it}u_{it}) = 0$, where $\ddot{x}'_{it} = x_{it} - \bar{x}_i$; and FE.2 $rank\left[\sum_{t=1}^{T} E(\ddot{x}'_{it}\ddot{x}_{it})\right] = K$

Stata command: xtreg lwage exper expersq married union i.year, fe vce(robust)

(f) Compare results from the pooled OLS and within (FE) estimator. What do you learn?

SOLUTION: According to the pooled OLS estimates, unionized workers earn on average approximately 18.25% more than non-unionized workers while the fixed effects estimates suggest that unionized workers earn about 8% more than non-unionized workers. This is consistent with the presence of unobserved heterogeneity correlated with the union variable.

[1] In most softwares you need to define what is the panel. In STATA, for instance, you would run the following code: *xtset nr year*. This tells the software that i = nr and t = year.