```
*****************************************************************************
* Author: Aadya Mishra (adya82001.m@gmail.com)
* Task: Perform power calculations, compute compliance, ITT, TOT & LATE, conduct Diff-in-Diff, RDD & Propensity Score Matching for a CCT-based program
* Latest Update: 28 March 2024
*****************************************************************************

cap clear all
cap program drop _all
cap log close
set more off
pause on

* Set global data directory when file path is local
global DataDir "/Users/aadyamishra/Desktop/Evaluation"
use $DataDir/DataFinal_ps3.dta, clear

*******************************
*PART1:POWER CALCULATIONS
*******************************

*Estimating sample size for a small impact (0.2 standard deviations)*
//m2=m1 + MDE
//m2 = 2933.233 + 0.2*5526.565

display (0.2*5526.565)+2933.233
power twomeans 2933.233 4038.546, sd(5526.565) power(.90) alpha(0.05)

*Change in sample size for power of 50%, 60%, 70%, 80% and 90%.*
power twomeans 2933.233 3883.233, sd(5526.565) alpha(0.05) power(0.5(0.1)0.9) graph(y(power))

*Estimating sample size for cluster RCTs with 20 samples per cluster; mean difference between treatment and control groups of 1500*
power twomeans 2933.233, diff(1500) sd (5526) m1(20) m2(20) rho (0.20) alpha (0.05) power (0.80) cluster


*****************************************************************************************************************
*PART2_A:Assesing extent of Imperfect Compliance in the Analysis of Labor Participation among Female Children Aged 8-12 Years in PROGRESA Program Villages, where villages are randomly assigned
*****************************************************************************************************************
* Open Dataset
```

```
use $DataDir/PanelPROGRESA_97_99year_Ind.dta, clear


*restrict the sample to female children between>=8 and <=12 years and year 1999*

capture drop fem1
        gen fem1 = .
        replace fem1 = 1 if (sex==0 & age<=12 & age>=8 & year== 1999)
        tab fem1
        lab var fem1 "Females aged 8-12 in 1999"

*cross tabulate frequency of observation by village level randomized assignment (D) and household level enrollment (D_HH)
tabulate D D_HH if fem1==1, cell

* Relationship between the PROGRESA program (D), household participation (D_HH), and mean labor participation among female children aged 8-12 years. *
tabulate D D_HH if fem1==1, sum(labor) mean

******************************************************************************************************
* PART2_B. Estimate effect of random assignment to treatment group (ITT) and actual participation (TOT)
******************************************************************************************************

*regressing on D and D_HH gives ITT estimate*
regress labor D_HH if fem1==1, cluster(D)
regress labor D if fem1==1, cluster(D)
save $DataDir/PanelPROGRESA_97_99year_Ind.dta, replace

*********************************************************************************
*PART2_C. Estimating Local Average Treatment Effect (LATE) using Instrumental Variables
*********************************************************************************

*Random Assignment of villages treated as instrument for one-sided compliance (D_HH) using 2sls *
ivregress 2sls labor (D_HH=D) if fem1==1, cluster(D)

*Estimating LATE manually*
//ITT = 0.28 and 0.78
regress labor D_HH if fem1==1, cluster(D)
    scalar ITT_hh= e(b)[1,1]
regress labor D if fem1==1, cluster(D)
    scalar ITT_vill=e(b)[1,1]

```

```
* Estimating compliance rate
count if D==1 & D_HH==1 & fem1==1 & !mi(labor) & !mi(D) & !mi(D_HH)
    scalar treated_hh= r(N)
count if D==1 & fem1==1 & !mi(labor) & !mi(D) & !mi(D_HH)
    scalar treatment_group= r(N)

scalar compliance = (treated_hh/treatment_group)
    display treated_hh/treatment_group

* TOT for households and villages
dis ITT_hh/compliance
dis ITT_vill/compliance


*******************************************************************************
*Part C. Difference in Difference - treatment effects of being a PROGRESA beneficiary in 2000 on labour conditions of children between 12 and 16 years age in 2003
*******************************************************************************

*************************************
* Processing the Data for DID Analysis
*************************************
* Set global data directory when file path is local
global DataDir "/Users/aadyamishra/Desktop/Evaluation"
use $DataDir/Panel_OPORTUNIDADES_00_07_year.dta, clear

*not interested in endline data, hence we drop it
drop if year==2007
* indicator variable for year 2003 and 2000
capture drop year_2003
gen year_2003 = 1
replace year_2003 = 0 if(year==2000)

* create a flag variable to identify unique household level observation each year
capture drop unique_hhid
egen unique_hhid = tag(hogid2 year)
lab var unique_hhid "Unique household ID, different for periods"

save $DataDir/Panel_OPORTUNIDADES_00_07_year.dta, replace

******
*Testing hosuehold-level variables using unique households for baseline balance when baseline period is 2000, treatment is assigned by village
ttable2 pov_HH Income_HH famsize agehead sexhead age sex edu_child enroll_child labor if (year==2000 & unique_hhid==1), by (D)
```

```
/*Since the year 2000 isn't a true baseline, and many treated villages were already enrolled in the program even at the baseline, enrolled households are expected to have higher household income from receiving the CCT. Hence, treated and control groups may not be truly comparable, which is essential for valid casual inference. Another quasi-experimental method would be required to adjust for any pre-existing differences in control and treatment villages.*/

*******************************************
* B) Estimate average treatment effect (ATE)
*******************************************

*i) restrict to: children 12-16 years (inclusive)
capture drop adolescent
    gen adolescent=.
    replace adolescent=1 if age<=16 & age>=12
    label var adolescent "Child is age 12-16 (both ends inclusive)"

*ii) Let us estimate a regression model clustered at village level (D) with an interaction term between village level treatment and period
    *Creating interaction term
capture drop treat_village_period
    gen treat_village_period = D * year
    label var treat_village_period "(treat_village*period) interaction term"
reg labor D year treat_village_period if adolescent==1, cluster(D)
    display _b[treat_village_period]

*iii) estimating model with additional covariates
reg labor D year treat_village_period age sex Income_HH famsize if adolescent==1, cluster(D)
display _b[treat_village_period]

***************************************************************************************
* PART D: Propensity Score Matching - to find 15 treatment and control villages as reasonably balanced at the baseline as possible
***************************************************************************************

use $DataDir/Homework4_PSM.dta, clear

* generating propensity scores using logit
psmatch2 treat bpl_hhs num_hp IsPW IsReach, logit common
* creating weights for PSM
pstest bpl_hhs num_hp IsPW IsReach, t(treat) graph both
```

```
**Assessing common support using graphing function
*i) Histogram
psgraph, bin(100) treated (treat)

*ii) Kdensity plot; Labelling lines at the bottom of the table
(control and treatment groups )

twoway kdensity _pscore if treat==1 || kdensity _pscore if treat
==0, legend(label(1 "Control Group")) legend(label(2 "Treatment
Group"))

** using output of pstest to identify matched pairs
* dropping observations which haven't been matched to any
control/treatment observations
codebook _weight
drop if missing(_weight)

* _id stores unique identification for a treated matched
observation with a corresponding value in _n1, that represents
the _id of its control counterfactual.
sort _id
sort _pdif
browse

*********************************
* PART E: Regression Discontinuity Design
*********************************
use $DataDir/PanelPROGRESA_Enrollment_97_99.dta, clear

* identify & assign treatment/control status by villages for all
years of observations, based on assignment status in 1998
capture drop temp1
gen temp1 = D if year==1998

capture drop D_assig
egen D_assig=mean(temp1),by(villid)

* identifying a household as poor/non-poor for all years of
observations based on poverty status in 1998
capture drop temp2
gen temp2=pov_HH if year==1998

capture drop temp3
egen temp3=mean(temp2),by(hogid)

*replace pov_HH = cond(year == 1997, temp3, .)
* flag unqiue household each year as multiple records are
collected from same household
capture drop uniqhh
egen uniqhh = tag(year hogid)
label var uniqhh "Unique HH level flag for each year"

```

```
*** Determining cutoff points of forcing variable in each region ***

* Distribution of the sample: eligibility and treatment status
tab D_assig pov_HH if year==1997 & uniqhh == 1, row
* inspecting assignment variable distribution
twoway (kdensity yycali if temp3==1 & D_assig==1 & year==1997) ///
       (kdensity yycali if temp3==0 & D_assig==1 & year==1997), ///
       legend(lab(1 Poor) lab(2 Non-Poor)) graphregion(fcolor(white)) title("Density of households in the distribution of yycali")

* Finding cutoff values for each region
capture drop maxcut
gen maxcut = 0
levelsof entidad, local(entidades) // we are creating a "vector" of values of regional IDs
foreach j of local entidades {  // for each region ID we will do the following
    summ yycali if year==1997 & D_assig == 1 & pov_HH==1 & entidad==`j'
    replace maxcut=r(max) if entidad==`j'

}
tab maxcut

*center the cutoff value around 0
capture drop Z
gen z = yycali - maxcut //assignment variable

twoway (kdensity z if pov_HH==1 & D_assig==1 & year==1997)  ///
       (kdensity z if pov_HH==0 & D_assig==1 & year==1997), ///
       legend(lab(1 Poor) lab(2 Non-Poor)) graphregion(fcolor(white)) title("Density of households in the distribution of assignment variable")

* identifying eligible households for receiving treatment
capture drop E
gen E = z<=0
label var E "Eligible for program"

**Identifying relevant RDD sample of eligibles in treatment villages (assigned treatment and poor) and ineligibles from control villages (not assigned treatment and non-poor)**
capture drop sampleRD
gen sampleRD = ( (D_assig==1 & pov_HH == 1 & z>=-200 & z<=0) | ( D_assig==0 & pov_HH == 0 & z>0 & z<=200) ) & year == 1999
* maintain relevant sample only to perform regression later
preserve
keep if sampleRD == 1
```

```
* using xtile to create new variables based on the values of an existing variable z
xtile hl = z if D_assig==1 & sampleRD == 1, n(30) // partition the relevant z values into 30 equally sized groups
xtile hu = z if D_assig==0 & sampleRD == 1, n(30)

* creating a new variable hd for storing values around the cutoff point of zero
capture drop hd
gen hd = -hl if D_assig==1
replace hd = hu if D_assig==0

* creating variables for mean z scores, child enrolment rates and assignment status for each bin of hd
egen meanZ = mean(z), by(hd)
egen meanEnroll = mean(enroll_child), by(hd)
egen meanD_assig = mean(D_assig), by(hd)

* creating mean Z score squared for polynomial fitting
capture drop meanZ2
gen meanZ2 = meanZ^2

* predicting mean enrolment rates for eligibles and ineligibles by fitting a regression model
reg meanEnroll meanZ meanZ2 if meanD_assig==1
predict yhat1 if e(sample)

reg meanEnroll meanZ meanZ2 if meanD_assig==0
predict yhat0 if e(sample)

sort meanZ

*********************************************************************************
*plotting a graphical analysis of RDD, comparing enrolment effects for eligibles in treatment villages and ineligibles in control villages
*********************************************************************************
twoway  (scatter meanEnroll meanZ if meanD_assig ==1) (line yhat1 meanZ if meanD_assig ==1) || ///
    (scatter meanEnroll meanZ if meanD_assig ==0) (line yhat0 meanZ if meanD_assig ==0), ///
    ylabel(0 1) xline(0) legend(off) graphregion(fcolor(white))

**********************************
* Regression analysis for SHARP RD
**********************************

restore

```

```
gen z2 = z^2

* With perfect compliance, restricting sample to eligibles from treatment villages and ineligibles from control villages
replace sampleRD = ( (D_assig==1 & pov_HH == 1 & z>=-200 & z<=0) | (D_assig==0 & pov_HH == 0 & z>0 & z<=200) )

* 1997
reg enroll D_assig z if sampleRD==1 & year==1997, vce(cluster villid)
estimates store r1_97

reg enroll D_assig z z2 if sampleRD==1 & year==1997, vce(cluster villid)
estimates store r2_97

* 1999
reg enroll D_assig z if sampleRD==1 & year==1999, vce(cluster villid)
estimates store r3_99

reg enroll D_assig z z2 if sampleRD==1 & year==1999, vce(cluster villid)
estimates store r4_99

* export meta-analysis of regression table to excel
findit xml_tab
ssc install xml_tab, replace

xml_tab r1_97 r2_97 r3_99 r4_99, replace save("DataDir$RD_TableI.xml") ///
    title("Table I: Sharp RD for Enrollment") below stats(N r2)
```