Spring 2023 Gregory Bruich, Ph.D.
Dept. of Economics, Harvard University

Lab 9: The Long-Run Causal Effects of HOLC "Redlining"

Methods/concepts: differences in differences; parallel trends assumption; diff in diff versus RDD

LAB DESCRIPTION

In this lab, you will use **differences in differences** and **regression discontinuity design** to study the long-run *causal effect* of the Home Owners' Loan Corporation (HOLC) "Redlining" maps on homeownership rates using data from 1910-2010.

Our empirical strategy will be based on the following historical details. In the 1930s, the HOLC did not draw maps for every city in the United States. In particular, cities whose populations in 1930 were less than 40,000 residents were not mapped. Most (but not all) cities with 1930s populations above 40,000 were mapped. We will focus on 53 cities with 1930s populations between 30,000 and 50,000. The 1930 Census took place before the HOLC maps were drawn, so we have 1910, 1920, and 1930 Census data before the HOLC maps were drawn, and 1940 to 2010 Census data after the HOLC maps were drawn. The data are described in Table 1. A list and description of each of the Stata and R commands needed for this lab are contained in Table 2 and Table 3, respectively.

QUESTIONS

1. Start with a graphical **regression discontinuity design (RDD)** analysis:

   a. Draw a binned scatter plot to show that the likelihood of having a HOLC map drawn changes discontinuously if a city's 1930 population exceeds 40,000 residents. Restrict the data to 1930. Include your graph in your solution write up.

   b. Draw binned scatter plots to test for smoothness of 2-3 city characteristics measured in a pre-treatment year (i.e., 1910, 1920, or 1930) across the 40,000 resident threshold. Include your graphs in your solution write up.

   c. What do you conclude about the validity of the regression discontinuity research design using the 40,000 threshold based on your graphs? Explain clearly what you see in your graphs that leads you to your conclusion.

   d. Draw a binned scatter plot to evaluate whether homeownership rates pooling all the data from 1940 to 2010 changes discontinuously if a city's 1930 population exceeds 40,000 residents. Include your graph in your solution write up.

2. Next run the following **regression discontinuity design (RDD)** regression pooling data for 1940 to 2010. Note that this is the same as the regression you ran in Lab 5.

$$own\_home_i = \beta_0 + \beta_{RD} above_i + \beta_2 dist\_from\_cut_i + \beta_3\, interaction_i + v_i \qquad (1)$$

   Here, the dependent variable is the homeownership rate own_home_i in city i. The indicator variable $above_i$ is 1 if the city i's 1930 population was more than 40,000 and 0 otherwise. The variable $dist\_from\_cut_i = pop\_1930_i - 40000$ is the difference between city i's 1930 population and the threshold. The variable $interaction_i = above_i \times dist\_from\_cut_i$ equals the product between the indicator $above_i$ and the distance from the threshold $dist\_from\_cut_i$.

   Generate the necessary variables and run this regression. Report and interpret the regression discontinuity estimate $\hat{\beta}_{RD}$ of the causal effect of the HOLC maps on homeownership rates.

3. Now we will turn to a graphical **differences in differences** analysis:

   a. Plot average homeownership rates in 1910 through 2010 for cities with 1930s populations below the 40,000 population threshold and cities above the 40,000 population threshold. Include your graph in your solution write up.

   b. Is the *parallel trends identification assumption* plausibly satisfied in the data? Explain clearly what you see in the graph that leads you to your conclusion.

4. Now calculate and report the following conditional means:
   - $\bar{Y}_{T,Pre}$ = mean home ownership rate in the years 1910 to 1930 in the *treatment group* cities with 1930 population above 40,000 residents
   - $\bar{Y}_{T,Post}$ = mean home ownership rate in the years 1940 to 1960 in the *treatment group* cities with 1930 population above 40,000 residents
   - $\bar{Y}_{C,Pre}$ = mean home ownership rate in the years 1910 to 1930 in the *control group* cities with 1930 population below 40,000 residents
   - $\bar{Y}_{C,Post}$ = mean home ownership rate in the years 1940 to 1960 in the *control group* cities with 1930 population below 40,000 residents

   Use these averages to calculate the impact of the HOLC "Redlining" maps on home ownership rates using **differences in differences**. Include your calculation in your solution write up.

5. Next run the following simple **differences in differences** regression using data in 1910-1960:

$$own\_home_{it} = \beta_0 + \beta_1 treat_i + \beta_2 post_t + \beta_{DD} post_i \times treat_t + u_{it} \qquad (2)$$

   where own_home_{it} is the homeownership rate in city i in year t, the indicator $treat_i$ is 1 if city i's 1930 population was greater than 40,000 and 0 otherwise; $post_t$ is 1 if the year is 1940 or later and 0 if the year is 1930 or earlier. Confirm that the coefficient $\hat{\beta}_{DD}$ equals what you calculated in the previous question. In order for this to work, the regression has to be run over exactly the same sample (same years and same cities) as in the previous question.

6. In practice, differences in differences would usually be implemented by replacing $post_t$ with separate indicator variables for each year and replacing $treat_i$ with indicators for each city. Run the following "fixed effects" regression using data in 1910-1960:

$$own\_home_{it} = \beta_0 + \beta_1 city2_i + \cdots + \beta_{52} city53_i + \beta_{53} year1920_i + \cdots + \beta_{57} year1960_i + \beta_{DD} post_i \times treat_t + v_{it} \quad (3)$$

   Report and interpret the difference-in-differences estimate $\hat{\beta}_{DD}$.

7. Putting together the results from the regression discontinuity design and differences in differences analyses you did above, what do you conclude about the causal effect of the HOLC "Redlining" maps on home ownership rates? Explain your conclusions and reasoning clearly.

8. The files to submit for this lab are:
   a. Your well annotated .do/.R/.rmd file replicating all your analyses above (with enough comments that a principal investigator on a research project would be able to follow and understand what each step of the code is doing). Please submit these files to Gradescope.
   b. For the Stata labs, please submit a log-file with the log showing the output generated by your final do-file. Please submit this file to the same gradescope assignment as the do-file. (Please do not submit a .smcl file: we can only read .log files in gradescope).
   c. A PDF version of the solutions to the above questions. For graphs, save them as .png files and insert them into the document. Please submit this file to the same gradescope assignment as the code. (Please do not submit a word document: we can only read PDFs in gradescope. Using R Markdown is never required; but if you have chosen to use it, you can *knit* the file to generate the PDF).

DATA DESCRIPTION, FILE: holc.dta

The data consist of 53 cities with 1930s population between 30,000 and 50,000 for a total of 581 observations. We observe these 53 cities in 11 Censuses (1910-2010). These data were generously provided by Professors Daniel Aaronson and Daniel Hartley at the Federal Reserve Bank of Chicago. For more details on the construction of these data and background on the HOLC Redlining, see Aaronson, Hartley, and Mazumder (2021).

TABLE 1
Variable Definitions

Variable (1)	*Description* (2)	*Obs.* (3)	*mean* (4)	*sd* (5)	*min* (6)	*max* (7)
city_m	Name of city (string)	581	n/a	n/a	n/a	n/a
city_id	Numeric city identifier (1-53)	581	n/a	n/a	1	53
year	Year	581	1960	31.70	1910	2010
pop_1930	1930 population	581	39,400	5,952	30,729	48,764
ownhome	Home ownership rate	572	0.513	0.114	0.179	0.740
holc_map	1 if HOLC drew “redlining map” for city, and 0 otherwise	581	0.511	0.500	0	1
shraa	Share African-American	518	0.0971	0.118	0	0.616
median_gross_rent	Median gross rent	415	503.9	213.5	76.07	1,688
median_house_value	Median house value	417	104,468	99,178	7,783	969,200
median_contract_rent	Median contract rent	466	472.5	191.3	76.07	1,627
pop	City's population in current year	258	78,135	78,020	20,226	520,116
foreign_born	Share foreign-born	469	0.0987	0.101	0.00400	0.557
employment	Employment rate	159	0.734	0.0960	0.385	0.902
nonwhite	Share non-white	571	0.133	0.143	0	0.728
median_rent	Median rent	257	400.9	138.9	76.07	910.9
labforce	Labor Force Participation	212	0.571	0.0395	0.469	0.685
read_write	Fraction Literate	159	0.958	0.0401	0.728	0.998
mortgage	Fraction with a Mortgage	106	0.399	0.150	0.0212	0.778
radio	Share with Radio	53	0.416	0.151	0.0797	0.796
rent	Mean Rent	106	498.6	653.5	172.6	6,501

TABLE 2
Stata Commands

STATA command	Description
*clear the workspace clear all version 17 *change working directory and open data cd "C:\Users\gbruich\Ec 50\Lab 9\" use holc.dta, clear *Display all variables in the data describe *Report detailed information on all variables codebook	This code shows how to clear the workspace, change the working directory, and open a Stata data file. To change directories on either a mac or windows PC, you can use the drop down menu in Stata. Go to file -> change working directory -> navigate to the folder where your data is located. The command to change directories will appear; it can then be copied and pasted into your .do file. The describe and codebook commands will report information on what is included in the data set loaded into memory.
*Create running variable, centered at threshold gen dist_from_cut = pop_1930 - 40000	This code shows how to create a new variable *dist_from_cut* that equals the 1930 population minus the threshold 40000.
*Install binscatter ssc install binscatter, replace *Draw graph (command all goes on one line) binscatter yvar dist_from_cut if year == 1930, rd(0) line(lfit) *Save graph graph export figure1_linear.png, replace *Draw graph (command all goes on one line) binscatter yvar dist_from_cut if inrange(year, 1940, 2010), rd(0) line(lfit) *Save graph graph export figure2_linear.png, replace *Draw with quadratic best fit line binscatter yvar dist_from_cut if inrange(year, 1940, 2010), rd(0) line(qlfit) *Save graph graph export figure3_quadratic.png, replace	The first command installs binscatter, which only has to be done once. The second command produces a binned scatter plot of *yvar* against *dist_from_cut* with a linear best fit line, restricting the graph to observations year equal to 1930. The third line saves the graph. The next block of code shows how to pool all the data with year between 1940 and 2010. The last bock of code shows how to change the best fit line to be quadratic by changing *line(lfit)* to *line(qfit)*. To find all the options for binscatter, type "help binscatter" For example, the option `nq(100)` would divide data into 100 equal size groups for purposes of binning (the default is 20 bins)
*Create running variable, centered at pop_1930 40000 gen dist_from_cut = pop_1930 - 40000 * Create indicator for being above threshold gen above = 0 replace above = 1 if dist_from_cut >= 0 * Interact dist_from_cut with above threshold gen interaction = dist_from_cut * above *Estimate regression (all goes on one line) regress yvar above dist_from_cut interaction if inrange(year, 1940, 2010), vce(cluster city_id)	These commands show how to run a regression to quantify the discontinuity in *yvar* at the 40000 threshold. We first create a new variable *dist_from_cut* the equals 1930 population minus the threshold 40000. We then generate an indicator variable *above* for *dist_from_cut* being positive. We next generate *interaction* that is the product between *dist_from_cut* and the indicator. Finally, we run a regression of *yvar* on these three variables, restricting the regression to observations with *year* between 1940 and 2010. The coefficient of interest is coefficient on *above*, the indicator for being above the threshold. The `, vce(cluster city_id)` option computes standard errors that take into account that there are repeated observations on each city.
*Create indicator for treated city gen treat = 0 replace treat = 1 if pop_1930 > 40000	This code shows how to create a new variable *treat* that equals 1 for the treated cities and 0 for all other cities. Note that this is exactly the same as the variable *above* defined earlier.

*Install binscatter ssc install binscatter, replace *Draw graph binscatter yvar year, by(treat) linetype(connect) discrete *Save graph graph export figure1_dd.png, replace	The first command installs binscatter, which only has to be done once. The second command produces a binned scatter plot of *yvar* against *year* with separate averages and lines for the cities with treat = 0 and treat = 1. The options shown are: 1. `by(treat)` will show separate binned averages and lines for each value of the variable treat 2. `discrete` divides data into groups based on the discrete values of the x-axis variable (year) for purposes of binning 3. `linetype(connect)` connects the binned averages with a line (instead of showing a best fit line). The third line saves the graph.
*Summary statistics (command all goes on one line) sum yvar if treat==1 & inrange(year, 1910, 1930)	These commands display summary statistic for the variable yvar for observations meeting two criteria: the variable treat equals 1 and the variable year takes on a value between 1910 and 1930. It uses the `inrange()` function in the conditional statement.
*Create indicator for treated state gen treat = 0 replace treat = 1 if pop_1930 > 40000 * Create indicator for after HOLC maps drawn gen post = 0 replace post = 1 if year >= 1940 * Interact treat and post gen dd = treat*post *Estimate regression (all goes on one line) regress yvar post treat dd if inrange(year, 1910, 1960), vce(cluster city_id)	These commands show how to run a simple differences in differences regression. We first create a new variable *treat* that equals 1 for the treated state (in this example, Arizona) and 0 for all other states. We then generate an indicator variable *post* for *year* being greater than or equal to 1940. We next generate *dd* that is the product between *post* and *treat*. Finally, we run a regression of *yvar* on these three variables and restrict it to years 1910 through 1960. The coefficient of interest is coefficient on *dd*. The `, vce(cluster city_id)` option computes standard errors that takes into account that there are repeated observations on each city.
*Estimate regression (all goes on one line) regress yvar dd i.year i.city_id, inrange(year, 1910, 1960), vce(cluster city_id)	These commands show how to run a differences in differences style regression with separate indicators for each year and each city. The `i.year` term in the regression generates the indicators for each year automatically. The `i.city_id` term generates the indicators for each city automatically. The `i.` prefix is an example of Stata's factor variable syntax.
*close any possibly open log-files cap log close *start a log file log using lab9.log, replace *commands go here *close and save log file log close	These commands show how to start and close a log file, which will save a text file of all the commands and output that appears on the command window in stata. The first line is short for "capture log close" which will close any open log files, and otherwise just proceed to the next step. Then the "log using lab9.log, replace" starts the log file and changes the default in two ways. First, it changes the file type to have a .log file extension, which creates a plain text log file (which is readable in Gradesope so is important!). Second, it also adds the ", replace" option which will save over any other log file that has the same name. This is usually what you want. The rest of your lab code can go below the "log using lab5.log, replace" line. At the end of your do-file you can include the last line which is "log close" which will close and save the log-file.

TABLE 3
R Commands

R command	Description

#clear the workspace
rm(list=ls()) # removes all objects from the environment
cat('\014') # clears the console
 	
#Install and load haven package
if (!require(haven)) install.packages("haven"); library(haven)
 	
#Change working directory and load stata data set
setwd("C:/Users/gbruich/Ec 50/Lab 9")
dat <- read_dta("holc.dta")
 	
#Report detailed information on all variables
summary(dat)
	This sequence of commands shows how to open Stata datasets in R. The first block of code clears the work space. The second block of code installs and loads the "haven" package. The third block of code changes the working directory to the location of the data and loads in holc.dta. To change the working directory in R Studio, you can also use the drop down menu. Go to session -> set working directory -> choose working directory.
	
	The easiest way to open a Stata data set in R Studio is to use the drop down menu. Go to file, then import data set, and finally browse to locate the file you want to open. This option will be available after you install the haven package.
	
	The summary command will report information on what is included in the data set loaded into memory, including information on the number of missing observations NAs for each variable.

#Create running variable, centered at 1930 population = 40000 dat$dist_from_cut <- dat$pop_1930 - 40000	This code shows how to create a new variable *dist_from_cut* the equals 1930 population minus the threshold 40000.

#Load packages
if (!require(tidyverse)) install.packages("tidyverse"); library(tidyverse)
if (!require(rdrobust)) install.packages("rdrobust"); library(rdrobust)
 	
#Subset data to observations in years 1940 to 2010
narrow <- subset(dat,year<=2010 & dist_from_cut>=1940)
 	
#draw binned scatter plot with linear fit
rdplot(dat_narrow$yvar, #outcome variable
    dat_narrow$dist_from_cut, #running variable
    p = 1,
    nbins = c(20, 20),
    binselect = "es",
    y.lim = c(0, 1.1),
    x.label = "City Population in 1930 minus 40,000",
    y.label = "Outcome variable (yvar)"
)
 	
#Save graph
ggsave("figure1_linear.png")
	The first command installs rdrobust, which only has to be done once.
	
	The second command subsets the data to only observations with *dist_from_cut* between -1.2 and 1.2.
	
	The third block of code produces a binned scatter plot of *yvar* against *dist_from_cut* with a linear best fit line. The options shown are:
	p = 1, #p = 1 is linear best fit line. p = 2 is quadratic
	nbins = c(20, 20), #number of bins on each side of threshold
	binselect = "es", #option to use "equal spaced" binning
	y.lim = c(0, 1.1), #Set y-axis scale
	x.label = "City Population in 1930 minus 40000", #x axis label
	y.label = "Outcome variable (yvar)" #y axis label
	
	The fourth block of code saves the graph.

#Load packages
if (!require(sandwich)) install.packages("sandwich"); library(sandwich)
if (!require(lmtest)) install.packages("lmtest"); library(lmtest)
 	
#Create running variable, centered at 1930 population = 40000
dat$dist_from_cut <- dat$pop_1930 - 40000
 	
#Create indicator for being above probation threshold
dat$above <- 0
dat$above[which(dat$dist_from_cut >= 0)] <- 1
 	
#Interact dist_from_cut with non-probation
dat$interaction <- dat$dist_from_cut*dat$above
	These commands show how to run a regression to quantify the discontinuity in *yvar* at the 1.60 GPA threshold. We first create a new variable *dist_from_cut* the equals 1930 population minus the threshold 40000.
	
	We then generate an indicator variable *above* for *dist_from_cut* being positive. We next generate the a variable *interaction* that is the product between *dist_from_cut* and the indicator.

<pre>##Subset data to [1940,2010] with new variables added
dat_narrow <- subset(dat, year<=2010 & year>=1940)
 	
#Estimate regression
linear <- lm(yvar ~ above + dist_from_cut + interaction , data =
dat_narrow)
 	
#Report coefficients and standard errors
coeftest(linear, vcovCL(linear, cluster = dat_narrow$city_id))</pre>
	Then we subset the data to a new data frame with *year* between 1940 and 2010.
	
	Finally, we run a regression of *yvar* on these three variables, restricting the regression to observations with *year* between 1940 and 2010. The coefficient of interest is coefficient on *T*, the indicator for being above probation threshold.
	
	The `vcovCL()` function computes standard errors that take into account that there are repeated observations on each city.

<pre>#Create indicator for treated city
dat$treat <- ifelse(dat$pop_1930>40000, 1, 0)
mean(dat$treat)</pre>
	This code shows how to create a new variable treat that equals 1 for the treated cities and 0 for all other cities. Note that this is exactly the same as the variable above defined earlier.

```
#install ggplot and statar packages
if (!require(tidyverse)) install.packages("tidyverse"); library(tidyverse)
if (!require(ggplot2)) install.packages("ggplot2"); library(ggplot2)

#Bin scatter plot – connected dots
ggplot(dat,
       aes(x=year,y=ownhome,
           colour = factor(treat, labels = c("1930 pop < 40K", "1930 pop > 40K")),
           shape = factor(treat, labels = c("1930 pop < 40K", "1930 pop > 40K")))) +
  geom_vline(xintercept=1935) +
  stat_summary(fun = "mean",geom="point") +
  stat_summary(fun = "mean",geom="line")  +
  labs(x = "Year", y = "y-axis title", shape = "", colour = "") +
  theme(legend.position="bottom")

#Save graph
ggsave("binscatter_connected.png")
```

The first command loads the tidyverse and ggplot libraries. The second block of code a binned scatter plot of yvar against year with separate dots and lines for the cities with treat = 0 and treat = 1. The options shown are:

1. shape = factor() will show separate binned averages and lines for each value of the variable treat using different shapes to connect the binned averages
2. colour = factor() will make the lines and connectors different colors based on the variable treat
3. geom_vline() as a vertical line
4. stat_summary() divides the data into groups based on the discrete values of the x-axis variable (year) for purposes of binning and reports means with geom="point" and lines with geom="line"
5. labs() adds axis labels
6. theme(legend.position="bottom") puts the legend at the bottom instead

The last line saves the graph.

<pre>#Summary stats for one variable
mean(dat$yvar, na.rm=TRUE)
 	
#Summary stats for observations with dvar==1 and year in 1910-1930
#Subset data
new_df <- subset(dat, dvar == 1 & year >= 1910 & year <= 1930)
 	
#Report mean
mean(new_df$yvar, na.rm=TRUE)
 	
#Alternatively, do it all at once using the with() function
with(subset(dat, dvar == 1 & year >= 1910 & year <= 1930), mean(yvar,
na.rm=TRUE))</pre>
	We used these commands in Lab 1. These commands report mean of *yvar*. The first line calculates these statistics across the full sample.
	
	The other lines illustrate how to calculate these statistics for observations meeting certain criteria: when another variable in the data is equal to 1 AND the variable year is between 1910 and 1930.
	
	The subset() function will pick out only the observations in a data frame that meet certain criteria. One way to proceed is to create a new data frame and then apply the mean() function to yvar in this new data frame. The second way to proceed is to do it all at once using the with() function. The with() function in R takes two arguments: a data frame and an expression. The data frame argument is dat and the expression applies the mean() function to the variable yvar: mean(yvar).

<pre>#Load packages
if (!require(sandwich)) install.packages("sandwich"); library(sandwich)
if (!require(lmtest)) install.packages("lmtest"); library(lmtest)
 	
#Create indicator for treated city
dat$treat <- ifelse(dat$pop_1930>40000, 1, 0)
 	
#Create indicator for after HOLC maps drawn
dat$post <- ifelse(dat$year>=1940, 1, 0)
 	
#Interact treat and post
dd <- dat$treat*dat$post
 	
#Data frame with subset of years and new variables generated
dat_narrow <- subset(dat, year>=1910 & year <= 1960)
 	
#Estimate regression (all goes on one line)
reg1 <- lm(yvar ~ dd + post + treat, data=dat_narrow)
 	
#Report coefficients and standard errors
coeftest(reg1, vcovCL(reg1, cluster = dat_narrow$city_id))</pre>
	These commands show how to run a simple differences in differences regression. We first create a new variable treat that equals 1 for the treated state (in this example, Arizona) and 0 for all other states. We then generate an indicator variable post for year being greater than or equal to 1940. We next generate dd that is the product between post and treat. Finally, we run a regression of yvar on these three variables and restrict it to years 1910 through 1960. The coefficient of interest is coefficient on dd. The , vcovCL() option computes standard errors that takes into account that there are repeated observations on each city.

<pre>#Estimate regression (all goes on one line)
reg2 <- lm(ownhome ~ dd + factor(year) + factor(city_id),
        data=dat_narrow)
 	
#Report coefficients and standard errors
coeftest(reg2, vcovCL(reg2, cluster = dat_narrow$city_id))</pre>
	These commands show how to run a differences in differences style regression with separate indicators for each year and each city. The factor(year) term in the regression generates the indicators for each year automatically. The fator(city_id) term generates the indicators for each city automatically. The , vcovCL() option computes standard errors that takes into account that there are repeated observations on each city.