HW 2

2023-10-22

HW 2 Development Economics

Group: Chiara Luna Pilato, Casey Shu, Rowen Kliethermes, Callie Cesewski

```
tidy=TRUE

#First I make sure I load the required libraries
library(pacman)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v forcats   1.0.0     v readr     2.1.4
## v ggplot2   3.4.3     v stringr   1.5.0
## v lubridate 1.9.2     v tibble    3.2.1
## v purrr     1.0.2     v tidyr     1.3.0

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become error
```

```
pacman::p_load(tidyverse)

#Then I make sure I am in the right working directory
setwd("/Users/chiaraluna/Desktop/Development Econ/HW2")


#Now I load the data

rural_roads <- read.csv('Rural_Roads.csv')
rural_roads_extended <-read.csv('Rural_Roads_long_format.csv')
```

Diff in Diff Impact Evaluation

Question a) Descriptive statistics: What is the mean population size across all villages in the sample in 1980?

```
#Get the mean population size
mean_pop_all_1980<-mean(rural_roads$Pop.1980)
print(mean_pop_all_1980)
```

```
## [1] 44429.64
```

The mean population size across all villages for the sample in 1980 is 44,429.63. We are also able to say then that we round up so that we have whole numbers, since we cant have 0.63 of a person. The mean population size across all villages for the sample in 1980 is 44,430.

Question b) Baseline differences in means: What is the difference in population size means in 1980 between villages that would be assigned to the program in the future and those that would not?

```
mean_pop_treatment_1980<-mean(rural_roads$Pop.1980[rural_roads$
  Road.Investment.Ever..Treatment.group.==1])
mean_pop_notreatment_1980<-mean(rural_roads$Pop.1980[rural_roads$
  Road.Investment.Ever..Treatment.group.==0])

difference_inmean_1980<-mean_pop_treatment_1980-mean_pop_notreatment_1980

print(mean_pop_treatment_1980)
```

```
## [1] 49877.2
```

```
print(mean_pop_notreatment_1980)
```

```
## [1] 40803.98
```

```
print(difference_inmean_1980)
```

```
## [1] 9073.219
```

The mean difference between the mean population size of the villages assigned to the program and those not assigned for the program would be 9,073.219. Here, once again we round so that we have whole numbers. We say the mean difference between the mean population size of the villages is 9,073.

Question c) If the program had been randomly assigned, what would you have expected as an answer in b)?

If the program was randomly assigned, we would have expected to have a same mean population; hence our value for b (the difference in these means) would have been or very close to zero. This is because when we have randomized trial, we have a control group and a treatment group. In our RTC programs, we need everything in these two groups to be the same, so that we are sure that the measured effect of the treatment can be attributed to the treatment only; not inherent differences between the groups.

In this case then, its important to have the same population means, so that population size differences can be discarded when analyzing differences in the effect of the treatment. In a diff-in-diff, it doesn't matter what the baseline differences are because as long as the trends are parallel, then that means the variable we are testing is causal in any changes in the outcome of the different villages.

Question d) Parallel trends: Population growth between 1990 and 1980 (the two pre- treatment data rounds) is described by the variable Delta Pop 90-80 defined as: (Pop90-Pop80)/Pop80. Use a regression to determine whether population growth is different between villages that would be assigned to treatment in the future and those that would not be (the "road investment ever" variable). Does the parallel trends assumption hold?

```
regression_d <-lm(Delta.pop.90.80 ~ Road.Investment.Ever..Treatment.group., data=rural_roads)
summary(regression_d)
```

```
##
## Call:
## lm(formula = Delta.pop.90.80 ~ Road.Investment.Ever..Treatment.group.,
##     data = rural_roads)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -9.223e-04 -9.020e-06 -3.010e-06  3.860e-06  8.821e-04
##
## Coefficients:
##                                          Estimate Std. Error   t value Pr(>|t|)
## (Intercept)                             1.000e-01  1.364e-06 73333.787   <2e-16
## Road.Investment.Ever..Treatment.group. -1.383e-06  2.157e-06    -0.641    0.522
##
## (Intercept)                            ***
## Road.Investment.Ever..Treatment.group.
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.283e-05 on 2498 degrees of freedom
## Multiple R-squared:  0.0001644,  Adjusted R-squared:  -0.0002358
## F-statistic: 0.4108 on 1 and 2498 DF,  p-value: 0.5216
```

As seen in the summary of this regression table and the value of the "Road.Investment.Ever..Treatment.group" coefficient being both very small and not statistically significant on the regressand Delta.pop.90.80; we can see that there is no effect on being on the treatment or not pre-treatment when it comes to population growth. Which means yes, the parallel trend holds and that the population growth rate is parallel between the two.

Question e) Impact:Estimate the effect of road construction on village population in the period 1990-2000 by regressing the percent change in population between 1990 and 2000 against the dummy of road construction between 1990 and 2000. How much does a small city grow if it gets a road in 1990-2000 compared to all other villages in India? Show your regression output and explain.

```
#Make a regression model against the dummy
regression_f<-lm(Delta.pop.00.90~Road.Investment.between.1990.2000, data=rural_roads)
summary(regression_f)
```

```
##
## Call:
## lm(formula = Delta.pop.00.90 ~ Road.Investment.between.1990.2000,
##     data = rural_roads)
##
```

```
## Residuals:
##        Min         1Q     Median         3Q        Max
## -6.552e-04 -6.590e-06 -5.800e-07  5.590e-06  4.989e-04
##
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       5.000e-02  8.559e-07   58417   <2e-16 ***
## Road.Investment.between.1990.2000 3.000e-01  1.752e-06  171280   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.734e-05 on 2498 degrees of freedom
## Multiple R-squared:      1,  Adjusted R-squared:      1
## F-statistic: 2.934e+10 on 1 and 2498 DF,  p-value: < 2.2e-16
```

As seen in the regression model summary above, the dummy variable of road construction between 1990 and 2000 has an effect of 3x10ˆ-1, which is equivalent to 0.3, which means that being in the treatment group of having a road built increases population by 30 percentage points in village population in the period 1990-2000.

Question f) Estimate the effect of road construction on village population growth in the period 2000-2010 by regressing the percent change in population between 2000 and 2010 against the dummy of road construction between 2000 and 2010. Show your regression output and explain.

```r
rural_roads<-rural_roads%>%
  mutate(difference_pop_10_00=(Pop.2010-Pop.2000)/Pop.2000)

regression_f<-lm(difference_pop_10_00~Road.Investment.between.2000.2010, data=rural_roads)
summary(regression_f)
```

```
##
## Call:
## lm(formula = difference_pop_10_00 ~ Road.Investment.between.2000.2010,
##     data = rural_roads)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -6.436e-04 -6.630e-06 -9.800e-07  5.020e-06  9.797e-04
##
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       3.000e-02  9.879e-07   30370   <2e-16 ***
## Road.Investment.between.2000.2010 3.000e-01  2.463e-06  121776   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.525e-05 on 2498 degrees of freedom
## Multiple R-squared:      1,  Adjusted R-squared:      1
## F-statistic: 1.483e+10 on 1 and 2498 DF,  p-value: < 2.2e-16
```

As seen in the regression model summary above, the dummy variable of road construction between 2000-2010 has an effect of 3x10ˆ-1, which is equivalent to 0.3, which means that being in the treatment group of having a road built increases population by 30 percentage points in village population in the period 2000-2010.

Question g) One time changes vs permanent shifts in the growth rate:Now regress village population growth in the period 2000-2010 against road construction in the 1990-2000 period as well as road construction in the 2000-2010 period (you now have 2 explanatory variables in the regression).

Is the growth rate between 2010-2000 higher in villages with roads built in 1990-2000 compared to those that never got roads? Ie, do roads have one-time or permanent effects on city growth rates? (Show your regression output and explain.)

```r
regression_g<-lm(Delta.pop.10.00 ~ Road.Investment.between.1990.2000+
                     Road.Investment.between.2000.2010, data=rural_roads)
summary(regression_g)
```

```
##
## Call:
## lm(formula = Delta.pop.10.00 ~ Road.Investment.between.1990.2000 +
##     Road.Investment.between.2000.2010, data = rural_roads)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -6.438e-04 -6.640e-06 -8.500e-07  4.900e-06  9.794e-04
##
## Coefficients:
##                                     Estimate Std. Error    t value Pr(>|t|)
## (Intercept)                        3.000e-02  1.168e-06  25683.893   <2e-16 ***
## Road.Investment.between.1990.2000 -9.074e-07  2.190e-06     -0.414    0.679
## Road.Investment.between.2000.2010  3.000e-01  2.542e-06 118039.599   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.526e-05 on 2497 degrees of freedom
## Multiple R-squared:      1,  Adjusted R-squared:      1
## F-statistic: 7.412e+09 on 2 and 2497 DF,  p-value: < 2.2e-16
```

As seen in the regression model above, in the years 1990-2000, the estimate coefficient is both very small and also statistically insignificant (As seen by the t values) However, we do see a growth in population in the years 2000-2010; by about 30 percentage points as well. Because of this, it is clear that roads have an effect on the immediate 10 years after being built, but they don't continue to have an effect 20 years after. Road construction has a one time effect.

Question h) Using the diff-in-diff specification: Up to this point you have performed an analysis using percent changes in population and checked that these growth rates are correlated to road construction. When you show your results to the government, they say they are promising but request that you implement a diff-in-diff specification:

#In order to do this, you must use the data in long format (use the Rural_Roads_Long_Format sheet). It has the same information as Rural Roads, except the data are stacked on top of each other for the different years and you use log population instead of changes in population as your dependent variable. Regress log pop in city i at time t against the year dummies (1990, 2000, 2010), as well as Treatment Group Dummy and TreatXPost dummy. What is the regression coefficient you obtain on Treatment GroupXPost (i.e. the variable: ProgramCityXPostTreatment)? Your regression output should report a 95% confidence interval. Is your estimated coefficient in e) and f) contained within the confidence interval in h) ? Show your regression output and explain.#

```
#First we build the regression
regression_h <-lm(log.pop~X1990.dummy+X2000.dummy+X2010.dummy+
Road.Investment.Ever..Treatment.group.+Treat.groupXPost.treatment, data=rural_roads_extended)
summary(regression_h)
```

```
##
## Call:
## lm(formula = log.pop ~ X1990.dummy + X2000.dummy + X2010.dummy +
##     Road.Investment.Ever..Treatment.group. + Treat.groupXPost.treatment,
##     data = rural_roads_extended)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5831 -0.3888  0.2940  0.6546  0.9581
##
## Coefficients:
##                                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)                             10.34014    0.01985 520.843  < 2e-16 ***
## X1990.dummy                              0.09531    0.02500   3.813 0.000138 ***
## X2000.dummy                              0.14431    0.02629   5.488 4.16e-08 ***
## X2010.dummy                              0.17470    0.02847   6.135 8.82e-10 ***
## Road.Investment.Ever..Treatment.group.   0.21740    0.02261   9.615  < 2e-16 ***
## Treat.groupXPost.treatment               0.25047    0.03411   7.343 2.25e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8839 on 9994 degrees of freedom
## Multiple R-squared:  0.04797,    Adjusted R-squared:  0.0475
## F-statistic: 100.7 on 5 and 9994 DF,  p-value: < 2.2e-16
```

```
#Building the confidence interval
cfi_upperbound <- 0.25047 +1.96* 0.03411
cfi_lowerbound <- 0.25047 -1.96* 0.03411

print(cfi_upperbound)
```

```
## [1] 0.3173256
```

```
print(cfi_lowerbound)
```

```
## [1] 0.1836144
```

The regression coefficient we obtained on Treatment Group X Post was .25047.

We have a 95% confidence interval of (0.1836144,0.3173256). This interval includes the values of 0.30 we got previously in questions e) and f), further supporting our claims that having the road built (having the treatment) does have an effect on population growth by 30 percentage points (increasing it); at least in the 10 years right after.

Question i) What do you think explains the difference in estimated coefficients between question h and questions e and f?

The e and f program has same effect in the ten years and the building of the roads always has the immediate 10 year effect on increasing the population growth of the village that received that treatment. In g, we see that there is no effect on population growth in the 20 years after receiving the treatment. e and f have higher coefficients because it only measures that immediate effect, but h takes into account those immediate years as well as all the post years which had no effect at all. Thus, it makes sense that the coefficient in h is slightly lower than .3.