Daniel Masters

Professor Bassi

ECON 318

May 1, 2022

Empirical Project

Introduction

As a consultant of the Chinese travel agency, we have decided to run a pilot project that introduces some of our call center employees to working from home rather than working in-person at our central office. This idea came about after we had issues of our workers quitting because of their long commutes. However, we are skeptical about working from home because it could lead to less productivity from our workers. To measure the effectiveness of the pilot project we decided to split our 249 call center employees into two groups. Our treatment group is composed of 131 employees that will work from home while the control group, 118 employees continued to come in to work in-person. The data I analyzed comes from before and after the pilot. After analyzing the data, I will assess the results to recommend working from home or not.

Data: Merging and Cleaning

Data collected before and after the pilot include: “Employee Status”, “Employee Characteristics”, “Quits”, “Quit Date”, “Attitudes”, “Performance Panel”, and “Performance.” “Employee Status”, “Employee Characteristics” and “Quits” are examples of cross-sectional data which means that these datasets feature many subjects during a point in time. More specifically, these three datasets observe all 249 employees and their various characteristics without specification of different time periods. Using the statistical software, Stata, these cross-sectional datasets were merged one-to-one on each employee using their unique id “personid.” A one-to-one merge on “personid” allowed me to create a combined dataset of all 249 employees with information provided in “Employee Status”, “Employee Characteristics”, and “Quits.” This was important because it allowed me to maximize the amount of data of each employee without duplicating them. This combination of datasets is called the master-data. I merged the master-dataset with “Performance Panel” using a one-to-many merging technique. In this case, “Performance Panel” is called the using-dataset and is considered panel data due to its inclusion of time periods: 2010 and 2011 and multiple observations of each employee over time. The one-to-many merging technique matches the unique observations in the master-dataset (“personid”) with many observations in the using-dataset. I repeated this process and merged the master-dataset with using-dataset, “Attitudes” which is also considered a panel dataset for the same reason as “Performance Panel”. I utilized the one-to-many merging technique for the master-dataset and “Attitudes” as well. Both these datasets are almost ready for analysis.

When merging data, it is always important to check if the data makes sense. To increase the quality of our data we must remove errors or nonsensical values present in the dataset before we can continue analyzing it. After carefully looking at the merged data, I found several categories with values that did not seem right. When looking over the data after merging the master-dataset with the “Performance Panel” using-dataset, I found that the column “performance_score” had values that were not between 0 and 100. I replaced these values in

Stata using the replace command and substituted the erroneous values with a period which designates it as a missing value. After merging the master-dataset with the "Attitudes" using-dataset, I found that the "age", "tenure", and "pior_experience" columns also contained non-sensical values that were less than 0. I replaced these values with same Stata replace command in all three cases.

Data Analysis and Results

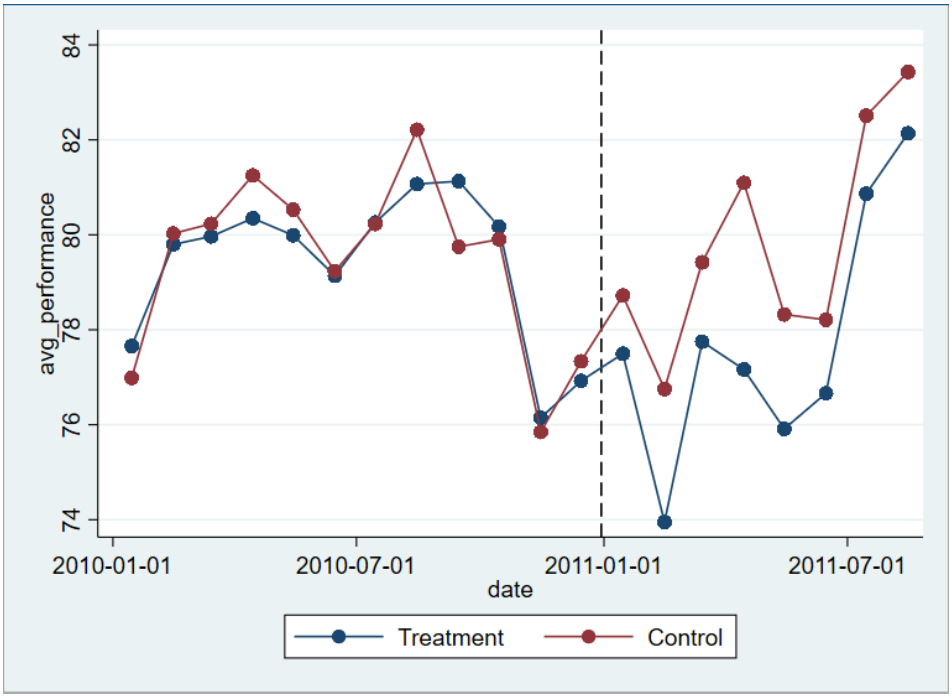
With the merged and cleaned datasets, I can now analyze the data. The question becomes what is the best way to analyze this data? Since we want to measure the effectiveness of working from home, we can use difference-in-difference estimation (DID). DID can help us determine the effect of the treatment (working from home) by comparing the change in outcomes of employees working from home versus those that go to work at the central office. The DID approach is useful because it removes biases in post-experiment period comparisons of the treatment and control group. It can also remove biases that form overtime with comparisons in the treatment group. To set up the DID regression with the merged master-data and the "Performance Panel" using-data, we need to declare and define a variable that combines the treatment and the survey collected after the pilot began. This variable will be called "post_treat" which equals the product of the given variables "treatment" and "post." This can be referred to as the interaction term. The "post" variable's inclusion in the regression equation will get rid of any time trend issues. The "Treatment" variable will take care of the differences in our treatment and control group, and the interaction term gives us the estimated treatment effect. Now, I regress on the variable within the dataset that gives the best measurement of working from home. The average performance score variable ("performance_score") best fits this description and I will use it as the dependent variable. Following the dependent variable, I include the independent variables: "treatment", "post", and "post_treat". The output of this regression is depicted in the output below.

VARIABLES	(1) performance_score
treatment	0.0265
	(0.337)
post	-2.181***
	(0.378)
post_treat	1.967***
	(0.512)
Constant	79.63***
	(0.243)
Observations	4,401
R-squared	0.010

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

The coefficient of the output is listed on the right, the standard error is listed below it in parenthesis and the observations are listed at the bottom. The coefficient of "treatment" is positive which means that working from home has a positive impact on performance scores. The coefficient of the "post" variable is negative which means that the performance scores are trending downwards over time. Lastly, the coefficient of "post_treat" suggests that working from home had an overall positive impact on performance score. The importance of this regression comes from measuring the statistical significance of the coefficient for "post_treat." Using the *** next to coefficient of "post_treat" one can see that the P-Value is less .01 (0.00) which means it is less than a 1% confidence level. A P-value that is less than the confidence level means that is a statistically significant and changes in this independent variable are related to the changes in the dependent variable. Since "post_treat" is statistically significant, DID estimation is strong in this case. DID achieved this coefficient value by taking the difference between the treatment group before and after the working from home experiment and subtracting this by the difference between the control group before and after the experiment.

With stronger assumptions, DID can be used to estimate the causal effect in the population. One of the most critical assumptions is the parallel trend assumption. It states that when the treatment is absent, the difference between the treatment and control group are constant over time. Violation of the parallel trends assumption leads to biased estimation when determining the causal effect. There is no statistical test for the parallel trends assumption, but visual inspection will suffice. The graph below shows my regression's fulfillment of the parallel trends assumption.



The dashed line in the graph shows the introduction of the treatment. It can be observed that before the dotted line the difference between the control and treatment is constant with only slight variation. After the treatment is introduced both the control and the treatment group are

parallel. DID estimation was successfully used to get rid of biases and after meeting the parallel trends assumption I can safely conclude that the coefficient for "post_treat" identifies the causal effect of working from home on performance scores of the employees.

Following the same logic above, I built a regression with a one-to-many merge of the master-dataset and the "Attitudes" using-dataset. The dependent variable that I selected to be regressed on was "satisfaction" which provides sufficient insight for an employee's satisfaction with their job and how working from home impacts this. I used "treatment", "post", and the interaction term "treat_post." The resulting table is below.

VARIABLES	(1) satisfaction
treatment	0.0450
	(0.210)
post	-0.227
	(0.185)
treat_post	0.570**
	(0.234)
Constant	4.562***
	(0.166)
Observations	855
R-squared	0.040

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Very similar to the table on the first page, we see that the "treatment" coefficient is positive which means working from home has a positive effect on employee satisfaction. Also, the "post" coefficient is negative which means that working from home has a negative trend on employee satisfaction. The "treat_post" coefficient is positive and statistically significant at a 5% confidence level because the P-Value is less than .05 (.015). Since "treat_post" is statistically significant, using DID estimation will provide strong results like the other regression. Following the same intuition that was developed in the performance score case, we can state that the coefficient for "treat_post" identifies a causal effect of working from home on employee satisfaction.

Conclusion

Both tables and regression have produced statistically significant independent variables in the DID estimation. In both cases, working from home had causal effects on employee performance scores and satisfaction. Both coefficients of these interaction terms were positive which means that the overall effect of the treatment was positive. With clear analysis and reasoning of the provided data, it seems that working from home will benefit the company by providing employees increased performance scores and satisfaction/retention.

Appendix A: Stata Code

```
cd "C:\Users\Daniel\Documents\ECON 318"
ssc install outreg2

use "EmployeeStatus.dta", clear
sort personid
merge 1:1 personid using "EmployeeCharacteristics.dta"
drop _merge
merge 1:1 personid using "Quits.dta"
drop _merge
merge 1:m personid using "Attitudes.dta"
drop _merge
replace age = . if age < 0
replace tenure = . if tenure < 0
replace prior_experience = . if prior_experience < 0
gen treat_post = treatment*post
reg satisfaction treatment post treat_post
outreg2 using attitudes_reg, word

use "EmployeeStatus.dta", clear
sort personid
merge 1:1 personid using "EmployeeCharacteristics.dta"
drop _merge
merge 1:1 personid using "Quits.dta"
drop _merge
merge 1:m personid using "Performance_Panel.dta"
drop _merge
replace performance_score = . if performance_score < 0
replace performance_score = . if performance_score > 100
gen post_treat = treatment*post
```

```
reg performance_score treatment post post_treat

outreg2 using performance_panel_reg, word

gen day = 15

gen date = mdy(month,day,year)

format date %tdccyy-NN-DD

bys date treatment: egen avg_performance = mean(performance_score)

twoway (connected avg_performance date if treatment == 0, xline(18626,
lpattern(dash)lcolor(black)))(connected avg_performance date if treatment == 1), legend(label(1
"Treatment") label(2 "Control"))
```