Empirical Project for ECON 318 Xinhui Yu

1. The Question

From the data set, I can analyze the effect of working from home both in terms of its impact on the firm and the impact on the employees. In my report, I'd like to focus on the home-working impact on the performance of employees. I will use Performance_Panel.dta, EmployeeCharacteristics.dta, Performance.dta and EmployeeStatus.dta to test whether there is obvious difference of the performance scores between the treatment and control groups.

2. Data Cleaning

a) Describing Data Sets

EmployeeStatus.dta shows employees' id numbers and their status (whether in the treatment group or not). EmployeeCharacteristics.dta shows detailed background about the employees, containing personal information on each employee assigned to the pilot at the start of the pilot (e.g., marital status, age, if completed high school). Performance_Panel.dta is monthly performance data on workers for the months before and after the experiment. Data is at the monthly level, where the variable post indicates if the month is in the pre-experiment period (January to November 2010) or in the period during the experiment (December 2010 to August 2011). I mainly use the performance evaluation variable, which is rating of their performance in that month (on a scale from 0 to 100). Performance.dta stores average performance of workers over the period before and after the experiment. Each worker has one observation in the pre-experiment period and one in the period during the experiment (denoted by the variable *post*).

b) Cleaning Variables

From the four data sets, I can find some abnormal values which are out of range or obviously impossible. I change the values into missing to avoid the influence of outliers on the regression result. Detailed changes made in the four data sets are shown in the Table 1.

c) Merging, Rearranging and Reshaping Data

I merged EmployeeStatus.dta into the other three data sets (Performance.dta, Performance_Panel.dta and EmployeeCharacteristics.dta) to bring the variable *treatment* into the data sets. And I also created 5 variables in total. First, to make further regression more convenient, I created the variable *treatmentXpost*, which equals to the product of variable *treatment* and variable *post*. The variable *treatmentXpost* captures the differences in differences effect of home-working. That is, the differential effect of home-working and after the experiment has started. What's more, I created variable *time* to combine *year* and *month* into one variable to work as x-axis. In addition, I created variable *m* as the mean of variable *performance_score* grouped by the same *time* and *treatment*. I created *performance_treatment* and *performance_control* variables to categorize *m* by treatment and control groups. Finally, I created variable *lperformance* to record the logarithm of variable *performance_score*, to make outliers less severe.

d) Testing Random Grouping

EmployeeCharacteristics.dta shows detailed background about the employees. I'd like to check that

the backgrounds of employees in treatment and control groups look the same on average, so I can attribute differences to the launch of the home-working policy. Table 2 is the balance table using data from EmployeeCharacteristics.dta. From the table, we can see that only variable *married* has significant difference between treatment and control groups. In conclusion, we can consider the two groups are selected randomly and have similar backgrounds.

3. Empirical Model and Strategy

a) Empirical Strategy

In this model, I used ***Differences-in-Differences*** to get the equation (1):

$$lperformance_{i,t} = \beta_0 + \beta_1 treatment_{i,t} \times post_{i,t} + \beta_2 post_{i,t} + \beta_3 treatment_{i,t} + \varepsilon_{i,t} \quad (1)$$

where $lperformance_{i,t}$ is the logarithm of average performance evaluation on a scale from 0 to log(100); $treatment_{i,t}$ is a dummy variable that equals 1 if an individual belongs to the treatment group; and $post_{i,t}$ is a dummy variable that equals 1 during the experimental period. Finally, $\varepsilon_{i,t}$ is the error term.

Because I want to know the effect of home-working policy, but in the experiment, I cannot know what changes would happen to the treatment group if they hadn't been selected to the treatment group. For example, without the home-working policy, people in treatment group may increase in performance scores due to proficiency, which will affect our estimation about the home-working impact. I can use ***Differences-in-Differences*** to deal with the panel data. So, I need the variable to represent the before/after condition (variable *post*) and the variable to represent the participant/non-participant condition (variable *treatment*), and I also create a variable *treatmentXpost* to get difference-in-differences estimates. These variables are based on Performance.dta.

The key assumption of this model is that the change in the control groups is a good approximation of what the change in the treatment groups would have been if they had not worked from home. To test whether the assumption is true, I need to use Performance_Panel.dta. I plot the performance scores for the treatment and control groups from Jan 2010 until the end of the experiment in August 2011. As Figure 1, trends of both groups are similar before the date of experiment began (Dec 2010, which is drawn in the red vertical line). But once the experiment began, the treatment group started to outperform the control group. Then I can conclude that the two groups follow parallel trends and the difference between the two is steady, which suggests the differences-in-differences estimator will be valid.

b) Regression Output

There are 496 observations in total in this model, and the standard errors in the output are adjusted for 249 clusters in personid. F(3, 248) is the F-stat which the null hypothesis states that the explanatory variables are not useful at all in explaining the dependent variable, i.e., $H0: \beta_1 = \beta_2 = \beta_3 = 0$. But according to $F_{3,248}$, the critical value is 3.00 at 5% level. 10.98 > 3.00, so we can reject the null hypothesis at a significance level of 5%, which is also proved by "Prob > F = 0.0000".

The variable *post* captures the difference in performance scores from before to after the home-working policy launch in the control group (i.e., for employees with *treatment*=0). From the output, we can conclude that performance scores of employees in the control group decreases by 4.14% after the

home-working policy launch, significant at the 0.1% level. The variable *treatment* captures the difference in performance scores between employees working from home and working in office before the policy (i.e., for the period with *post*=0). However, we cannot reject the null hypothesis that variable *treatment* has no effect on the outcome performance scores at the 5% level. The effect of the home-working policy on performance scores is measured by variable *treatmentXpost*, which is statistically significant. Overall performance of the treatment group is found to be 3.73% higher than the control group after the experiment started, significant at the 0.1% level.

4. Conclusion

The frequency of working from home has been rising rapidly around the world due to Covid-19 recently. I can conclude the results of the experiment on working from home from the data, which employees who volunteered to work from home were selected randomly. I can find a 3.73% increase in performance scores from home-working, which is very significant. Although the result seems to be supportive for working from home, we should note that some factors may contribute to the possibly unrepeatable result. For instance, the employees are working in call center, which means the job has less limitation on the working environment and fewer requirements of communication between coworkers. However, if the company want to expand working from home for their full workforce, many positions may not fulfill similar requirements, such as product managers.

There are also many companies against work from home. Tesla CEO Elon Musk enacted a strict return-to-office policy this spring, informing employees suddenly by email on May 31 that they would need to “spend a minimum of forty hours in the office per week.” In conclusion, working from home increases employees’ performance scores statistically, but more research should be taken before the company expands working from home for their full workforce.

Table 1: Errors in EmployeeStatus.dta and Performance.dta

Data set	Observation	Variable	Value	Change and reason for change
EmployeeStatus.dta	No problem values			
EmployeeCharacteristics.dta	personid = 5018	prior_experience age tenure	-99	This is an outlier and does not seem possible. I changed it to a missing value
EmployeeCharacteristics.dta	personid = 26618	prior_experience age tenure	-99	This is an outlier and does not seem possible. I changed it to a missing value
EmployeeCharacteristics.dta	personid = 21648	prior_experience	24.888889	It's more reasonable to take an integer month. I changed it to 25.
Performance_Panel.dta	personid = 29996 year = 2011 month = 7	performance_score	1000	This is an outlier and does not seem possible. I changed it to a missing value
Performance_Panel.dta	personid = 29996 year = 2011 month = 3	performance_score	1000	This is an outlier and does not seem possible. I changed it to a missing value
Performance_Panel.dta	personid = 38046 year = 2011 month = 4	total_monthly_calls	-999999	Negative calls in a month aren't possible. I changed it to a missing value
Performance_Panel.dta	personid = 31292 year = 2011 month = 4	calls_per_hour	200	This is an outlier and does not seem possible. I changed it to a missing value
Performance.dta	personid = 29216 post = 1	performance_score	176.3382	This is an outlier and does not seem possible. I changed it to a missing value
Performance.dta	personid = 29996 post = 1	performance_score	278.1695	This is an outlier and does not seem possible. I changed it to a missing value
Performance.dta	personid = 38046 post = 1	total_monthly_calls	-108951.9	Negative calls in a month aren't possible. I changed it to a missing value
Performance.dta	personid = 31292 post = 1	calls_per_hour	42.67397	This is an outlier and does not seem possible. I changed it to a missing value
Performance.dta	personid = 21710 post = 0	total_monthly_calls	0	Total monthly calls are less than calls per hour, which is not possible. I changed it to a missing value
Performance.dta	personid = 39530 post = 0	total_monthly_calls	0	Total monthly calls are less than calls per hour, which is not possible. I changed it to a missing value

Table 2: Comparison between treatment and control groups

	Treatment	Control	Difference
prior_experience	16.754	19.256	2.502
	(23.818)	(27.771)	(3.307)
age	24.347	24.403	0.056
	(3.536)	(3.572)	(0.453)
tenure	28.254	25.674	-2.580
	(21.938)	(21.547)	(2.769)
basewage	1,562.799	1,539.864	-22.935
	(185.397)	(136.157)	(20.480)
bonus	1,092.587	1,030.901	-61.685
	(655.717)	(597.664)	(79.430)
grosswage	3,003.362	2,949.730	-53.632
	(825.599)	(758.050)	(100.362)
costofcommute	8.338	7.892	-0.446
	(5.554)	(8.031)	(0.884)
rental	0.203	0.244	0.041
	(0.404)	(0.431)	(0.053)
male	0.466	0.466	-0.000
	(0.501)	(0.501)	(0.064)
married	0.322	0.221	-0.101*
	(0.469)	(0.417)	(0.056)
high_school	0.864	0.824	-0.040
	(0.344)	(0.382)	(0.046)
Observations	118	131	249

Table 3: The performance impact of home-working was positive

	(1)
	lperformance
post	-0.0414***
	(-5.63)
treatment	0.00259
	(0.28)
treatmentXpost	0.0373***
	(3.94)
_cons	4.369***
	(683.83)
N	496

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Figure 1: Performance changes by time in treatment and control groups