**UNIVERSITY OF CALIFORNIA AT BERKELEY**

**BIG DATA & BETTER DECISIONS**

**MBA217: SPRING 2018**

**MIDTERM EXAM**

***WEDNESDAY, MARCH 21st, 2018***

STUDENT NAME: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

STUDENT ID: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

START TIME: **9:40am**

SUBMIT TO BCOURSES: **12:20pm (late exams not accepted)**

Your exam consists of four sections with a total of **21** questions on **14** pages (including the cover page) and is worth **235** points total. Please make sure to answer all questions.

The exam is open book and open note, and you are permitted to consult online references.  However, collaborating with other students or asking any other individuals to assist you with the exam questions is not allowed. As members of the UC-Berkeley community, our expectation is that you will adhere rigorously to the UC-Berkeley Honor Code. Anyone caught cheating on a quiz or exam in this course will receive a failing grade in the course and will also be reported to the University Center for Student Conduct.

**Part I: Increasing Deposits in Savings Accounts** **(65 Points)**

Savings is important. Despite high returns to savings only 22 percent of adults worldwide report having saved at a formal financial institution in the past year (Dupas and Robinson 2013).

USCEU, a credit union in California, decided to try to encourage the use of savings accounts. In 2010 they offered a lottery for cash prizes for 2 consecutive months-- September 2010 and October 2010. If an individual opened a new account with at least $50 or increased their account balance by $50 they were entered to win.

USCEU hopes that the temporary incentive of a 2-month lottery will increase number of account openings and encourage subsequent long-term use of these accounts. The number of accounts opened at each branch each month is on average 3 new accounts across all branches.

USCEU randomly selected 110 branches to partake in the study, and randomly assigned 70 to control and 40 to treatment. They checked for baseline balance on total number of accounts and number of new accounts for the first 6 months of 2010. They found balance at baseline (July 2010) and started putting up promotion flyers for the incentives in the treatment branches in August 2010.

**Questions:**

**1. What is the research question? (5 points)**

**How do short-term incentives to save affect long-term savings outcomes?**

**2. Describe a pathway (theory of change) between the intervention and the outcome of interest? (10 points)**

**The intervention is the lottery system which incentivizes opening a bank account. We expect this will increase in the number of accounts opened and in turn impact long term usage of the accounts and increased savings.**

**3. Please describe the evaluation methodology you will use to answer the research question. Clearly identify the treatment and the control group and how branches are assigned to the treatment and control. Please also explain why you believe that this method gives you an unbiased estimate of the counterfactual. (10 points)**

**Randomized assignment/randomized offering/randomized controlled trial (2 points)**

**Treatment is the branches offered the lottery, Control is branches not offered the lottery (4 points)**

**This method should provide an unbiased estimate of the treatment because randomization ensures that the only reason one group was treated versus another was chance/randomization means there should be no systematic reasons that one group received the treatment and another did not and so outcomes should not be correlated to treatment (i.e. there will be no selection bias). (4 points)**

Suppose USCEU tells you that they estimated the treatment effects using the following regression specification and estimated separate treatment effects for each month:

(1) ytj= α+BtTj+εjt  ,

where ytj = ATM accounts opened in month t in branch *j*, Tj = 1 if the branch is a treatment branch in period *j* and 0 otherwise. They run equation 1 for each month they have data. They have data for 2 months before the treatment periods, the 2 months of treatment, and 3 months after the treatment ends. The results of these analyses are presented in Table 1.

**4. Equation 1 is the same as performing what other statistical test? Please explain why. What conditions are required for the estimates to be causal? (10 points)**

**Equation 1 -running this month-by-month regression of the treatment effect- is the same as running a T-test for the differences in the means of the treatment and control groups for each month. The coefficient in the regression is the estimate of the difference in means and a ttest of whether is zero or not is a test of the difference in means. (5 pts)**

**Conditions for causal estimate:**

**1. We have a randomized trial**

**2. We have balance at baseline**

**Since we have balance at baseline, we can view this effect as the valid treatment effect and we have no need to control for other factors due to the randomization working. (5 pts)**

Table 1: Impact of Treatment on Total ATM Account Openings per Branch (Extensive Margin)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Month | Treatment Effect | SE | P-value | Control mean |
| 10-Jul | 0.06 | (0.74) | 0.93 | 3.61 |
| 10-Aug | 0.04 | (0.97) | 0.97 | 3.99 |
| 10-Sep | 0.42 | (0.98) | 0.67 | 4.26 |
| 10-Oct | 1.46\*\* | (0.74) | 0.05 | 3.21 |
| 10-Nov | 0.88\* | (0.53) | 0.1 | 2.27 |
| 10-Dec | 0.29 | (0.66) | 0.66 | 2.41 |
| 11-Jan | 0.66 | (0.45) | 0.15 | 1.71 |

*Notes:* Number of observations is 110 branches for all regressions. This table reports the estimated effect of lotteries treatment on the opening of ATM accounts. Each treatment effect comes from a separate linear regression corresponding to equation 1. The lotteries were held in Sep-10 and Oct-10 and ATM accounts were eligible for the lottery. \*\*\* p<0.1, \*\* p<0.05, \* p<0.10

**5. Based on the results in Table 1, would you say the conditions to interpret the results as casual are met? Please explain your answer. (5 points)**

**Yes – there is no difference in the number of accounts opened in treatment groups compared to control groups in the two months prior to the intervention. This is confirmed by the P-values for the test of differences in means. Hence the treatment and control groups are balanced.**

**6. Based on the results in Table 1, what would you say is the impact of the treatment is on ATM accounts opened? Please explain your answer. (10 points)**

**The impact is that in the second month of the lottery 1.46 more accounts were opened in treatment branches. Thus USCEU accomplished its goal of increasing account openings due to the lottery. The number is economically meaningful as well as statistically significant at the 95 percent confidence interval.**

**(6 pts positive effect, 2 pts for magnitude, 2 pts for statistical significance)**

Then USCEU tells you a little more about those accounts opened in October, the second month of the lottery. Particularly they are interested in how many of the accounts were kept open that is accounts that survived, the proportion of accounts that had at least one transaction per month, and the average number of transactions per month.

Table 2—Survival Rates of Account Openings During October of the Lottery

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Treatment | Control | Diff. | P-Value |
| Account Survival Rates (1 year) | 0.94 | 0.95 | -0.01 | 0.67 |
| Account Survival Rates (1 year > 50 pesos) | 0.81 | 0.76 | 0.05 | 0.26 |
| Number of Transactions Per Month (1 year) | 1.72 | 1.74 | 0.03 | 0.93 |
| Proportion of Accounts with Transactions (1 year) | 0.72 | 0.67 | -0.06 | 0.22 |

Notes: This table presents rates in October 2011 (1 year after the lottery).

**7.** **Based on the results in Table 2, did the short-term lottery incentive have a sustained long-term impact? Please explain your answer. (10 points)**

**No, there is no statistically significant difference in long-term outcomes. The p-values are all above .05.**

**OR**

**Yes, there is no statistically significant difference between survival rates meaning that the treatment didn’t effect closing of accounts. Since the treatment individuals are using their accounts at the same rates as the control group, there is a long term effect of usage.**

**8. Based on the results from the above study, would you recommend that the same policy be implemented in France? Why or why not? (5 points)**

**No, the study is not externally valid to a different population in Kenya. (2 points)**

**Kenya may have different cultural norms with banking, access to banks, and attitudes towards lotteries (3 points)**

**Part II: Private Pension Startup for the Gig Economy (70 Points)**

Pensions are one of the most important components of a social security system. The most common type of pension in the U.S and many other high-income countries is a contributory plan that is financed by taxing own labor income. However, the rise of the “gig economy” (e.g. Uber, Lyft) has led to concerns over employee contributions given their status as contract workers as well highly varying earnings. In addition, many older American’s are working in these jobs as a way to augment the existing pensions. Together, these factors point to a public policy challenge as well as a PR concern for “gig economy” companies as they see their employees retire and attempt to introduce driverless vehicles and other AI tools that require far less labor.

Enter Replaced by Robots Retirement (RRR), a startup that wants to provide a small non-contributory universal pension scheme for seniors financed by Uber, Lyft and other employers. The objectives of the pension program are to allow older adults to retire form the work force without lowering their family’s standard of living measured by consumption. By decoupling the payment from hours worked the scheme has the potential to address large losses for people who worked too little prior as well as to facilitate the PR and social challenges associated with the shift to self driving cars.

To garner support RRR is piloting the program in a number of small communities in California and trying to gather evidence on impact for their corporate customers and regulatory agencies. In the pilot adults were eligible for the pension if they were over 65 years old, had worked at least 1 year in one of the corporate sponsors and lived in communities with fewer than 2,500 inhabitants. Applicants must provide documentation to prove their age and place of residence. Beneficiaries receive a cash transfer of $1,000 every month. The beneficiaries of the program are also invited to take part in workshops and social development activities.

RRR carried out two surveys in the early stages of the program’s implementation. The first survey (or the baseline survey) was carried out between September and November 2016. The survey collected information on individuals and households before the disbursement of cash transfers took place. The second – or follow-up—survey was carried out between November and December 2017, once the program had been operating for a year. You are hired by the RRR and the State of California to conduct an impact evaluation of the program.

**Questions:**

**1)** **What is the research question? Suggest two indicators you would use to measure outcomes. (15 points)**

**Research question: What is the impact of the non-contributory universal pension scheme on senior adults’ labor force participation and economic and health well-being? (must refer specifically to pension program ) 5 pt**

**Indicators: e.g. Labor force participation, Consumption expenditure (must give two distinct, measurable, and relevant outcomes) 10 pt**

You are given access to both waves of the survey data on a sample of adults who were of age 65-70 during the time of the first survey from both the treatment localities (500-2,500 inhabits) and control localities (2,501-3,300 inhabitants).

**2)** **An intuitive way of calculating the treatment effect would be to simply estimate the difference between the average of the relevant indicator for adults in the treatment localities and the same average for those in the control localities after the program has been implemented. What identifying assumption(s) is required for this strategy to yield an unbiased estimate of the program’s impact?**

**The treatment and the control groups should be identical on average absent the pension.**

**(-5 points if assumption given is for difference-in-difference)**

In order to assess the validity of the assumption(s) above, you produce the following table using the data from the baseline survey:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Treatment (<2500 Residents) | Control (2500-3500 Residents) | P(value for test of Equality) |
| Age | 66.90 | 66.93 | 0.621 |
| Male | 0.50 | 0.35 | 0.000 |
| Years of Schooling | 10.86 | 10.39 | 0.005 |
| Married | 0.66 | 0.46 | 0.000 |
| Geriatric Depression Scale | 3.52 | 3.69 | 0.339 |
| Worked Last Week | 0.36 | 0.31 | 0.078 |
| Worked Last Week for Pay | 0.23 | 0.23 | 0.000 |
| Worked Last Week for no Pay | 0.13 | 0.09 | 0.029 |
| Hours Worked Last Week | 14.20 | 10.93 | 0.003 |
| Hours Worked Last week for Pay | 9.28 | 7.72 | 0.074 |
| Hours worked last week for no pay | 4.92 | 3.21 | 0.036 |
| Labor earnings | 176.81 | 200.38 | 0.509 |

**3)** **Based on the results in Table 1, do you think the identifying assumption(s) you propose in Question 3) is valid? Why or why not? (10 points)**

**No. There are significant differences along many characteristics between those in the treatment localities and those in the control localities in the baseline, such that it is very unlikely that these two groups will have the same outcomes even in the absence of the pension program.**

**4)** **Given the design of the program and the data you have,** **propose an alternative strategy to evaluate the impact of the program. What assumption is needed for this alternative strategy to be internally valid? Given the data that you have, can you test this assumption? (20 points)**

**The alternative strategy is difference in differences. The assumption required is that the evolution in the outcomes for the subsample in the treatment localities would have been the same as the evolution in the outcomes for those in the control localities in the absence of the pension program, i.e. no differential time trends between the treatment and control localities (after controlling for observable characteristics). No, you cannot test this assumption as you need two surveys before the intervention to do so.**

**Points:**

**-10 if RDD is given, with the running variable being age, as the data available do not contain those aged below 70**

**-10 if regression or matching is given and all questions were answered, as difference in differences is preferable**

**-15 if RCT is given, since it is not compatible with the design of the program and the data available**

**- 3 if RDD is given and all questions were answered, with the running variable being number of individuals in a community, as it is a group-level variable and there are unlikely to be sufficient number of observations around the threshold (see Table 1)**

You conduct the evaluation strategy proposed in Question (4) and produce the following results:

**Table 2: Pension Program Impact on Labor Force Participation and**

**Consumption Expenditure**

|  |  |  |
| --- | --- | --- |
| ***Dependent Variable*:** | **Worked Last Week for Pay** | **Consumption per Capita** |
| Treatment Effect | -0.047 | 63.342 |
| Standard Error | (0.016) | (16.401) |
| t-statistic | -2.938 | 3.862 |
| Control Group Mean | 0.23 | 422.91 |
| Number of Observations | 1950 | 1417 |

**5)** **Based on results in Table 2, would you say that the pension program has achieved its objectives? Why or why not? (15 points)**

**Yes the pension program decreases labor force participation and increases consumption, as both treatment effects are positive and statistically significant at the 5 percent level.**

**Points:**

**-5 if does not mention that the program achieved its objective by decreasing labor force participation**

**PART III: Evaluating a Sales Training Program**

You work for a company that sells office supplies to other businesses and relies on a sales team to market these products. You were dissatisfied with sales figures in 2016 and decided to send low-performing sales people to a week-long training conference at the beginning of 2017. All sales people who fail to hit a yearly sales goal of $800,000 in total revenue for 2016 were required to attend the training program.

After the first year of this program, you need to decide whether offering the training to lower performing employees succeeds in increasing their sales volume, and whether you should continue the program and require the training conference again this year. You hire a consultant to help you with data analysis. To carry out the evaluation, she proposes exploiting the rule you used to determine who was required to attend the training, i.e. that those with sales below $800,000 were required to attend. She suggests that you compare outcomes for those whose 2016 sales were very close to this cutoff, i.e. just above and just below.

**Questions:**

**1)** **What is the main research question? Suggest a measurable outcome of interest. (10 points)**

***Research Question: “What is the causal effect of the training program on 2017 sales revenue for sales associates who attended the program?” (5 points)***

***Outcome of Interest: Total Sales Value in 2017 (could also say total sales margin, increase in sales from 2016-2017, etc) (5 points)***

**2)** **Which evaluation methodology is being proposed? Identify the treatment and control groups. (10 points)**

***Methodology: Regression Discontinuity (6 points)***

***Treatment: Associates with total sales value just below $800k (2 points)***

***Control: Associates with total sales value at $800k or just above (2 points)***

Assume for Questions (3) and (4) that the required assumption to apply the methodology that you proposed in Question (2) are proven valid. Your consultant shows you the following figure:



**3)** **What can you say about the treatment effect based on the figure above? Be specific about what the treatment is. What other statistics would you request (on the same outcome) that help you evaluate the treatment effect in this case? (10 Points)**

***From the graph above, the treatment of attending the training appears to have increased sales in 2017, based on the vertical shift in the trend of sales to the left of the cutoff value of 800k in 2016 sales (appears to be around $1000-$2000). (5 points)***

***However, we cannot conclude based on this graph whether this difference is statistically significant. Knowing the exact value of the shift (vertical difference at the cutoff) and standard error will tell us whether the training increased sales significantly (pvalue or t statistic would also tell us this). (5 points)***

Additionally, your consultant runs the regression below using data for associates between $750000 and $850000 of 2016 sales, where “Treat” takes a value of 1 if the sales associate was required to attend the training session (2016 sales below $800,000), Sales2016 is the dollar amount of sales in 2016 and Sales2017 is the dollar amount of sales in 2017.

She presents you with the following table:

**Table 1: Regression Results on Treatment Effect in Test using RDD**

|  |  |
| --- | --- |
| **REGRESSION RESULTS** | **Total Value of Sales in 2017 ($)** |
| Coefficient on “Treat” (β1) | *1301* |
| *SE* | *209.7* |
| *T-statistic* | *6.20* |
|  |  |
| Coefficient on “Sales2016” (β2) | *1.049* |
| *SE* | *0.044* |
| *T-statistic* | *23.86* |
|  |  |
| *Constant (α)* | *12748* |

**4)** **Please interpret the results in Table 1. Is the training program effective in improving sales and if so by how much? (10 points)**

***Based on the Coefficient β1 , the program increased sales by $1301 in 2017. (5 points)***

***This difference is statistically significant at the 1% level based on the T statistic (also significant at 5% level). (5 points)***

***There is also a strong linear association between 2016 and 2017 sales ($1 increase in 2016 sales is associated with a $1.05 increase in 2017 sales).***

**5) Suppose your department head decides that in addition to continuing the program this year for low-performing sales associates, he wants to require all sales associates to attend the training, regardless of their previous year’s sales performance. Would you have any concerns with this decision, given your results above? Discuss why or why not you think this is a good idea. (10 points)**

***You should be skeptical of this decision because our results are only valid for sales associates near the cutoff in 2016. We cannot necessarily extrapolate to the rest of the associates, so we essentially have no evidence on how sales would be affected for high performing associates. It may be that the training is only effective for low-performing associates in which case the company would be wasting lots of money & time by making everyone complete the training. An alternative would be to test the training again this year by using a randomly selected group of higher performing employees for the treatment group, and make the decision about whether to scale the training program up in the following year. (10 points)***

Suppose that you decided to continue the program for low-performing associates only for another year. However, this year, sales associates knew in advance that if they did not hit their yearly sales goal of $800,000 in 2017, they would be required to attend the training program. At the end of the year, you decide to use the same approach as above to estimate the treatment effect of the training program in its second year. Before conducting the analysis, your consultant shows you the following histogram of the distribution of yearly sales amounts for different associates.

****

**6) Do you think the same approach can be applied to compute the treatment effect of the training program in its second year? Explain why (or why not) by discussing the identification assumptions that are fulfilled (or violated). (10 points)**

***The assumption that individuals don’t select into or out of treatment is likely violated in this case. Individuals knew ahead of time that they would have to attend the training if they failed to meet the cutoff, so some of those who were close may have worked extra hard to push themselves above the cutoff. Those who remained just below the cutoff then might be on average less informed or motivated, due to these selection issues rather than the treatment itself. The bunching in the distribution of 2017 sales hints that this could be a possibility. Additionally, it is the second year of the program, so many who used to be below the cutoff were pushed above it in 2017 as a result of improvement after the training. Those remaining below the cutoff may be selected to be the “non-responders”. (10 points)***

**­PART IV: Evaluating Overnight Delivery**

You work for a retail company that sells products to consumers online. Up to 2014, you charged a flat rate for shipping of $5 per order for standard delivery (5-7 days) and $10 per order for express delivery (2-3 days). In 2015, you started offering overnight shipping in a small number of zipcodes for $20 per order. These zipcodes were selected based on ease of offering the overnight delivery option. You are considering expanding this shipping option to other areas, but want to better understand how offering overnight shipping affected your sales.

You decide to compare the change in sales before and after offering the new shipping option with the change in sales of nearby zipcodes that were not offered overnight shipping. The new shipping option was offered to selected communities starting in the 7th month of the year. The data scientist on your team produces the following graph for you.

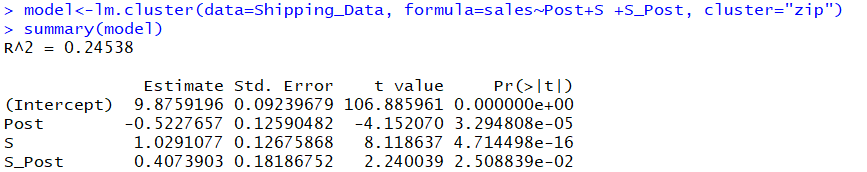


**1)** **What evaluation methodology is being proposed? Do you think the approach is appropriate given the graph above? Explain why or why not by discussing the identifying assumptions required for this method. (10 points)**

***Methodology: Difference in Differences (5 points)***

***DiD is likely an appropriate method in this case because the treatment and comparison group have parallel trends in the pre-period (5 points)***

Now, you ask your data scientist to estimate the treatment effect of offering overnight shipping using average sales in each zipcode from months 5-8. So, each zipcode has 4 observations. She gives you the following R output.



The variable **S** is equal to 1 if the observation is from one of the zipcodes that eventually received the new shipping option (0 otherwise), while **Post** is equal to 1 if the observation is from a time period after the new shipping option was offered (0 otherwise). **S\_Post** is the interaction of these two variables and is equal to 1 only if the observation is in a treated zipcode after the shipping option was offered (0 otherwise). The outcome you are analyzing is total monthly sales (**sales**) in millions, with the variable **zip** identifying the zipcode of each observation.

**2)** **Please interpret the regression output above. What is the effect of the new shipping option on sales? (10 points)**

**The treatment effect is a $407,390 increase in sales. (5 points)**

**This is statistically significant at the 5% level (pvalue=0.025, t=2.24). (5 points)**

**BONUS: Notice that we’ve used the lm.cluster() function instead of lm(). Why might we want to use lm.cluster() instead of lm() in this case. How do you think the results from lm() would compare to those above? (10 points)**

***We use lm.cluster in order to cluster the standard errors by zipcode. One of the assumptions required for linear regression is random sampling or independence of error: This assumption states that the errors of the outcome variables are uncorrelated with each other. Intuitively, it means that each observation of Y (sales) must be independently drawn from the population from other outcomes. In practice, the 4 observations from the same zipcode at different points in time will be correlated (we call this autocorrelation). Therefore, the assumption of independent Y is not valid-- we must use clustered standard errors to mitigate the problem.  If we had used lm instead of lm.cluster, our standard errors would likely have been lower. With lm.cluster we lower our effective sample size and statistical power. However, in this case, results are still highly significant even using the clustered standard errors.***