HEVAL5140 - Exam - Part I - Guide to answers

Max 6 pages. Deadline: 10. December, 23.59.59

Context: In this paper will will examine questions related to the effect of eliminating copayments for youths when going to the physician. In order to do this, you will use data from Norway which tells you how often females in different age groups went to the physician (in different years). The copayment was eliminated in 2010. Before 2010 you had to pay about 15 EURO for every visits when if you were 12 or older. In 2010 the age threshold was increased to 16 years. This means that, for instance, 14 year old females had to pay in 2009, but not in 2010. The general question is whether this led to an increase in the number of visits to the physician.

Please see the pdf notebook for examples of code and figures. Note that because students select different age groups as control group, they may get different number answers for some of the questions. Note also that the importance of the interpretation of the results and the discussion (since much of the calculation was aided by what we did in the class).

1. **Make a figure which shows that average number of visits to the physician among 14 year old females for the different years. Give some brief comments about the shape of the figure.**

- Remember to label both axis
- Use average not total number of visits (as some did)
- Main comment: Increasing trend over time, but mainly an increase between 2009 and 2010
- Bonus if comment on the steepness of the curve and make a quick calculation of the percent increase between 2010 and 2011

2. **Calculate the mean number of visits before and after the interventions for 10, 12, 14 and 16 year old females.**

- OK (see appendix)

3. **Make a figure showing the average number of visits in different age group in one year.**

- OK (see appendix)
- Decrease from 0 to about 11, then increase, peak around 30 (pregnancy is a contributing factor), increase in old age and peak around 78, then decrease (possible explanation: nursing home not recorded).

4. **Run a standard regression with the average number of visits of 14 year old females as the dependent variable and a time trend as the independent variable. Is there a time trend?**

- Yes, there seems to be a (positive) time trend (significant), but the increase could also be the result of the intervention, so we need to include both variables (time and intervention dummy).
- Should comment on how large the trend is (bonus if calculate percentage)

5. **Run the same regression (as in 4), but with a dummy variable after the reduction of the copayment. Does it look like the introduction of copayment increased the number of visits to the physician for 14 year olds?**

- Both copayment dummy and trend dummy (year) are significant
- Bonus if comment on the reduction of the trend coefficient compared to regression in q4 (copayment will take some of the effect of trend)
- List some standard assumptions: (No omitted variables, no autocorrelation, no heteroskedasticity, linearity)
- Comment on key causal assumptions (constant causal effects? Are all confounding variables included?)
- May also comment on the relatively few observations (but also that behind each annual aggregate observation there are many individuals)
- Also bonus for comparing and discussing size of coefficients (time trend vs. copayment coefficient size)

6. **Discuss some factors that affect whether you think the coefficient for the intervention dummy in the regression (in 5) can be interpreted as the causal effects of the reduced copayment.**

- In order to assume that the effect is causal:
    - No omitted variables (all confounders are included)
    - Constant causal effect
- May also comment on technical assumptions: linearity, no autocorrelation and so on (unless it is discussed in q5)

7. **Run a difference-in-difference regression to test whether the intervention had an effect for 14-year old females. Interpret the results and discuss the assumptions behind the method.**

- See notebook appendix for one possible set of results.
- Key assumption: Parallel trend (control group and intervention group)

- Useful (bonus) if also present the equations and explain the method (not just run the regression)

8. **Discuss whether it would be possible to use a propensity score method to analyze whether copayments affect visits to the physicians for young females. Assume you could get all the data you want: What kind of data or information would you like to have in order to do a propensity score method to estimate the effect of the intervention.**

- First point: We have aggregate data, but to run propensity score methods we would like to have have individual data with information about many covariates.
- However, even with this information (and bonus to recognize this): Propensity score method is not ideal (regardless of home many more data points you get) since in this case it does not make sense to speak about “the probability of receiving the treatment” (=being a 14 year old). It is not like some people have a low probability of being 14, who are not!

9. **Discuss whether it would be possible to use the instrumental variables method to analyze whether copayments affect visits to the physicians for young females.**

- Explain iv method and assumptions (correlation with intervention, not directly correlated with outcome)
- Relate the method and the assumptions to the current case: it seems difficult to use the IV method in this case.

10. **How would you test whether the difference you observed in the regression from (7) is significant and whether the method works well?**

    - Test the statistical significance of the interaction term
    - Conduct placebo tests to test method (try different years and age groups and see if you still get an effect when there should be no effect)
    - Examine pattern: If the method works, the effect should be bigger the closer the used interventions year is to the true intervention year

HEVAL5140, Home Exam, part II

Max 6 pages

Deadline (together with part I which has been handed out previously): 10. December at 23.59.59

Background
Register data can be used to study the effect of interventions. In this exam, you will examine the effect of Right Heart Catheterization on critically ill patients.

The original dataset is available at: http://biostat.mc.vanderbilt.edu/wiki/Main/DataSets . Search for rhc, download the .csv file for data and see the html file for variable description). For a presentation and key results on the topic, see: http://www.mc.vanderbilt.edu/crc/workshop_files/2008-04-11.pdf

1. **What percentage of patients died in the group that received RHC vs. No RHC? Is the average difference a good indication of the causal effect of RHC? Explain why or why not.**

More deaths among patients with RHC, but this is may be because those patients are sicker, not because the intervention is bad.

2. **Repeat the analysis from (1) in the following sub-groups: male vs. female, young vs. old age and one or more sub-groups (or sub-groups of sub-groups) you think is interesting. Do you believe that this is a better estimate of the causal effect than the average difference in (1)?**

Comparison between more homogeneous groups could get us closer to the true causal effect because we are comparing more "like-with-like" not, for instance: old vs. young. In short, it is a simple form of (one variable) matching.

3. **Simulate how many times you get a statistically significant coefficient (for the coefficient b) when you run an ordinary least squares regression of x on y (y = a + b x), where x and y are 100 randomly drawn integers between 0 and 100. Do the simulation 1000 times. Draw a histogram that shows the distribution of the p-values you get (for the coefficient b) and interpret the results.**

In general you will get a "significant" p-value between two series (using the 5% threshold) 50 times when you do make completely random completely random numbers 1000 times.

4. **Assume an a researcher examines the difference between many sub-groups, and finds a sub-group with a significant difference at the 5% level. She claims that RHC has an effect in this sub-group. Do you agree?**

- Unless the sub-group distinction is rooted in strong theory or that the hypothesis was proposed before the test was carried out, you are bound to get some random significant results if you compare many groups (bonus if refer back to q3).
- May discuss the problem of “fishing” in the data
- May also discuss Bonferroni correction

5. **Estimate the probability of receiving RHC (using a logistic regression) given some of the variables you believe might be important (gender, age etc. You do not need to include all kinds of variables, but select some you think are most relevant).**

See notebook.

6. **Create a figure that shows the distribution of the probability of receiving RHC for:**
    a. **All patients**
    b. **Patients who received RHC**
    c. **Patients who did not receive RHC**
    d. **Why is this figure important?**

See notebook.
The figure is important because a key assumption for propensity score matching is that there is overlap i.e. for each probability there be should be some people with and without the intervention. The graph visualises whether this is true.

7. **Create some sub-groups of individuals with similar probability of receiving RHC and compare the proportion of deaths for RHC vs. No RHC within these sub-groups.**

    See notebook.

8. **Discuss under which assumptions the estimate in (7) is a good measure of the causal effect.**

The two key assumptions are:
- No omitted variables
- Overlap (see q6)

9. **Make a figure which shows survival within 30 days after discharge for RHC vs. No RHC patients. Is the difference statistically significant?**

- Graph and use a statistical test (example log-rank)
- The conclusion depends on the level of significance demanded.

10. **Calculate the mean age and gender compositions for RHC and No-RHC patients. Next, calculate the mean age and gender values (for RHC and No-RHC patients) for all sub-groups from question (7). Comment on whether you believe you are comparing similar types of patients.**

The mean group difference in terms of age, gender etc should be smaller for individual in the same probability subgroup than the overall difference. (This is sometimes called the ‘balance’ of the data).