Wayne State University - Department of Economics

Master Field Exam: Econometrics

December 2019

1. Fourteen countries engaged in "expansionary austerity" policies in response to the 2008 financial crisis. I am interested to know if these austerity policies worked. For simplicity, treat austerity as a dummy variable equal to 1 for countries that engaged in it and 0 for others.

[1 point] (a) Write the regression model (math equation) for the difference-in-difference (DID) estimator. Be precise, use meaningful names for the variables, and write the subscripts of the variables and coefficients as well.

$$Y = \beta_o + \beta_1 after + \beta_2 austerity + \beta_3 after * austerity + u$$

[0.5 point] (b) Write in plain English the definition of the variables used in item "a".

β_o = intercept
after = dummy variable indicating 1 for time after crisis and 0 for time before crisis
austerity = dummy variable indicating 1 for countries using expansionary austerity measures after the 2008 financial crisis and 0 for countries which did not
δ_1 = Estimated effect of austerity measures
Y = some economic evaluator like economic growth or unemployment

[0.5 point] (c) According with item "a", what is the coefficient that gives the difference-in difference (DID) estimator?

β_3

2. Suppose we have panel data on voter opinions toward government spending in 2010, 2012, and 2014. Explain why we can or cannot estimate the effect of each of the following variables in a fixed effects regression.

[1 point] (a) Gender

Gender must have time-invariant values with time-invariant effects to include it in the fixed-effect model. Which means the value of the gender doesn't change across time, and the effect of gender on the outcome at time t is the same as the effect of gender at time t+1. Otherwise, we need to have explicit measurements of gender. We can include the interaction of gender with time in the case of time-varying effects. Also, if gender

interact with other variables in the model we need to have explicit measurements of time-invariant.

[1 point] (b) Income

Income could be included in a fixed effects equation and its impact could be estimated since its values change over time.
Fixed effects models can't control the variables that are vary over time like income. Thus, to include income in the model we need to include dummy variables for time or space unit.

3. Many school districts pay for new school buildings with bond issues that must be approved by voters. Supporters of these bond issues typically argue that new buildings improve schools and thereby boost housing values. Cellini, Ferreira, and Rothstein (2010) used Regression Discontinuity to test if passage of school bonds caused housing values to rise.

[1 point] (a) What is the assignment (running or score) variable?

The assignment variable is the election result, or the percentage of votes "yes" on the referendum.

[0.5 point] (b) Explain how to use Regression Discontinuity to estimate the effect of school bond passage on housing values.

Cellini, Ferreira, and Rothstein (2010) estimate the value of school facility investments to parents and homeowners using dynamic regression discontinuity. They distinguish the effects of capital investments on housing prices by comparing districts in which school bond passed or failed by narrow margins. The RD design effectively eliminates endogeneity by comparing districts in which the school bond referendum passed or failed by a narrow margin, and comparing the housing prices to find the treatment effect.

[0.5 point] (c) Provide a specific equation for Regression Discontinuity

$$\text{Housing values}_i = \beta_0 + \beta_1 T_i + \beta_2(\text{Election result}_i - 50) + u_i$$

Where $T_i = 1$ if the bond passes (election results > 50) and $T_i = 0$ if bond does not pass (election results < 50)

4. Earnings functions, whereby the log of earnings is regressed on years of education, years of on the job training, and individual characteristics, have been studied for a variety of reasons. Some studies have focused on the returns to education, others on discrimination, union non-union differentials, etc. For all these studies, a major concern has been the fact that ability

should enter as a determinant of earnings, but that it is close to impossible to measure and therefore represents an omitted variable.

Assume that the coefficient on years of education is the parameter of interest. Given that education is positively correlated to ability, since, for example, more able students attract scholarships and hence receive more years of education, the OLS estimator for the returns to education could be upward biased. To overcome this problem, various authors have used instrumental variable estimation techniques. For each of the instruments potential instruments listed below briefly discuss instrument validity (inclusion condition and exclusion condition).

[1 point] (a) Years of education for the individual's mother or father.

This is generally considered a valid instrument because it has been shown to be correlated with an individual's years of education (inclusion condition) and is probably uncorrelated or weakly correlated with anything that might be captured in the error term (exclusion restriction). Parent's ability might be correlated with parents' years of education, putting the instrument's exclusion restriction validity into question. However, this is uncertain and requires further research.

[1 point] (b) Number of siblings the individual has.

This instrument is not likely to have a very strong correlation with an individual's years of education, so it would not have a very strong first stage (inclusion condition). This variable is more likely tied to family culture or income than years of education, so it fails the exclusion restriction as well. Apparently, some research has shown a link correlation between number of siblings and years of education, so it may have a strong first stage but more than likely still fails the exclusion restriction.

5. Are members of Congress more likely to meet with donors than with mere constituents? To answer this question, Kalla and Broockman (2015) conducted a field experiment in which they had political activists attempt to schedule meetings with 191 congressional offices regarding efforts to ban a potentially harmful chemical. The messages the activists sent out were randomized. Some messages described the people requesting the meeting as "local constituents," and others described the people requesting the meeting as "local campaign donors."

Variables for Donor Experiment:

treat_donor: Dummy variable indicating that activists seeking meeting were identified as donors (1 = donors, 0 = otherwise)

staffrank: Highest-ranking person attending the meeting: 0 for no one attended meeting, 1 for non-policy staff, 2 for legislative assistant, 3 for legislative director, 4 for chief of staff, 5 for member of Congress

Python code:

```
import numpy as np

df['MOCMeeting'] = np.where(df['staffrank'] >=5, 1, 0)
```

[1 point] (a) Is there any endogeneity problem (omitted variable bias, self selection bias, reverse causality, etc) if I estimated the simple regression below via Linear Probability Model (LPM), Logit or Probit? Justify.

$$MOCMeeting = \beta_0 + \beta_1 treat\_donor + \epsilon$$

Since the researchers randomized the messages to Congress members indicating whether an individual donated or not, one-way causality can be inferred and there wouldn't be an endogeneity problem. If the messages sent to Congress members were not randomized, there would be an issue with reverse causality in this model. If it is common knowledge that donating to a Congress member's campaign increases an individual's chances of meeting with the Congress member, then people will be more likely to donate if they want to meet with a Congress member. Additionally, Congress members have more of an incentive to meet with and listen to donors because they are directly funding the Congress member's campaign advertisement ability and subsequently increasing their chances of becoming elected.

However, there could be sources of omitted variable bias. The Congress members might have access to information outside of these standardized "messages" to let them know whether an individual contributed to the campaign. Additionally, the size of the donation surely impacts how likely it is that the meeting will take place. This experiment does not control for such a factor.

[1 point] (b) Do you recommend me to add additional covariates (explanatory variables) to the regression above? Yes or No? Explain the logic?

Adding more variables that are explanatory would help decrease standard error. As this experiment was randomized already we would assume not many covariates would have to be added.

Wayne State University - Department of Economics

Final HW - Part 2

December 2019

[1point] 1. Let us assume we implement an AND function to a single neuron. Below is a tabular representation of an AND function:

X1	X2	X1 AND X2
0	0	0
0	1	0
1	0	0
1	1	1

The activation function of our neuron is denoted as:

$$f(x) = \begin{matrix} 0, & for & x < 0 \\ 1, & for & x \geq 0 \end{matrix}$$

What would be the weights and bias? Show the computation and select the right letter below:

$Y = w_1x_1 + w_2x_2 + b$

a) 1.5*0 + 1.5*1 + 1.5 > 0 → fails
b) 2*0 + 2*1 + 1.5 > 0 → fails
c) 1*1 + 1*1 -1.5 > 0 → pass
   1*0 + 1*0 - 1.5 < 0 → pass
   1*0 + 1*1 - 1.5 < 0 → pass
   1*1 + 1*0 - 1.5 < 0 → pass

a) Bias = 1, w1 = 1.5, w2 = 1.5
b) Bias = 1.5, w1 = 2, w2 = 2
c) Bias = -1.5, w1 = 1, w2 = 1
d) None of these

[1 point] 2. The steps below for using a gradient descent algorithm are in the wrong order. Write the right order.

4, 3, 1, 5, 2

1) Calculate error between the actual value and the predicted value.
2) Reiterate until you find the best weights of network.
3) Pass an input through the network and get values from output layer.

4) Initialize random weight and bias.
5) Go to each neurons which contributes to the error and change its respective values to reduce the error.

[1point] 3. In Deep Learning, what are the benefits of mini-batch gradient descent?

> Mini-batch gradient descent allows for a more robust convergence compared to batch gradient descent. It also makes the algorithm faster, and more efficient by split the training set into smaller sets, and implement gradient descent on each set.

[2points] 4. What are the differences between feedforward neural network and a recurrent neural network?

> In a feedforward network, information only moves in one direction - from the input nodes, to the hidden nodes, to the output nodes - and does not form a cycle. On the other hand, recurrent neural networks can use an internal state to produce sequences of inputs. Recurrent Neural Networks use backpropogation to adjust weights based on the gradient and improve the algorithm.

[2points] 5. For each of parts (a) through (d), indicate whether we would generally expect the performance of a flexible statistical learning method to be better or worse than an inflexible method. Justify your answer.

(a) The sample size n is extremely large, and the number of predictors p is small.

> A flexible method would perform better because it can take full advantage of the large sample size, and reduce bias.

(b) The number of predictors p is extremely large, and the number of observations n is small.

> A flexible method will cause overfitting due to the small sample size, which is more likely to have more noise in the model. Thus an inflexible method will perform better in general.

(c) The relationship between the predictors and response is highly non-linear.

> An inflexible method has a problem in finding the non-linear effect, so a flexible method will perform better in general.

(d) The variance of the error terms, i.e. $\sigma 2$ = V ar(), is extremely high.

A flexible model is more likely to overfit all the noise between the data when the variance of the error terms is extremely high. Therefore, an inflexible model will perform better it is less likely to overfit the noise.

[1point] 6. Explain what are the advantages and disadvantages of k-fold crossvalidation relative to:

(a) The validation set approach?

Validation set approach is computationally less demanding and simpler to use than LOOCV or k-folds cross validation. However, it may be biased because the validation set is usually arbitrarily chosen and isn't always representative of the actual data. The predictive accuracy of a model using such an approach could be skewed because it isn't "smoothed out" by multiple iterations as is the case in k-fold cross validation. Even with the simplest k=2 form of k-fold cross validation, predictions are comparably more valid without becoming overly computationally-demanding.

(b) Leave-one-out cross-validation (LOOCV)?

LOOCV is the same as k-folds cross validation when k=n. Thus, it is simply a special case of k-folds cross validation. The higher k is, the more computationally demanding so LOOCV requires more time and computational resources. One advantage of LOOCV over k-folds cross validation is that it uses almost the entire data set to train, reducing overall bias of the prediction.

[1point] 7. Describe the difference between Bagging, Random Forest, and Boosting.

Bagging is a simpler decision tree method that results in less variance and reduced overfitting. The prediction is based on random bootstrap samples of the dataset. Because it is a relatively simple method, it also performs slightly worse for prediction.

Random forest is based on applying bagging to decision trees with addition to sampling, also samples the variables. With random forest specifically, at each stage of running the algorithm, the choice of variable is limited to a random subset of variables. Random forest builds on adding bagging (mentioned earlier) and the bootstrap sampling of variables within each split in the algorithm.

Boosting is like bagging most commonly used in decision trees. However, boosting fits a series of models with each successive model fit to minimize the error of the previous models. Their are various forms of boosting, the most general and used is gradient boosting. In relation to random forest, stochastic gradient boosting adds randomness to an algorithm by sampling observations and predictor variables at each stage.

[1 point] 8. What are the differences between K-Means Clustering and Hierarchical Clustering? Give examples of clustering application.

Hierarchical clustering is a method were alike objects into clusters. These clusters are then grouped by similar factors. An example of this would be grouping states by similar amounts of income. Where states with state income per capita are grouped by how close they are to each other, and across the whole hierarchy, the objects across all clusters are similar as they measure the same thing.

K-Means clustering tries to keep objects in each cluster, grouped that are alike by things like value of mean(s) while keeping the clusters as far as possible from each other. Generally speaking K-means uses distance measurements to group objects together. One application might be for astronomy - clustering stars into galaxies based on distance from telescopic images.