**Govt 702**

Final Exam

Spring 2021

Name: \_\_\_\_Timothy Liptrot\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

“I commit myself to be honest in any academic endeavor and to respect and uphold the Georgetown University Honor System.”

Signature \_\_\_\_\_\_\_Timothy Liptrot\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**NOTES: NOTES:**

1. 150 total points are possible on this exam. Each answer is worth 5 points unless otherwise noted.

2. Show your work for partial credit. Explanations should be brief, and calculations may be approximate.

3. The boxes show the approximate space needed to write an answer. If you need additional space, please indicate “continued on back” and write on the back of the page.

4. You may use your notes, book and (not recommended) the internet. You may not communicate with anyone or get any other help during the exam.

**Good luck!**

1. For this question, we will analyze the relationship between daily Facebook stock prices (“FB”) and the probability the Donald Trump would win the 2020 presidential election (“Trump”). The data on Trump win probability is from the PredictIt betting market. We also control for the overall stock market performance using the Dow Jones Industrial Average (“DJI”). As appropriate, refer to these models to answer the following questions.

**## Model A: Facebook prices as a function of Trump and DJI**

lm(formula = FB ~ Trump + DJI, data = stkPres\_pre)

Estimate Std. Error t value Pr(>|t|)

(Intercept) 8.037e+01 2.064e+01 3.894 0.000134 \*\*\*

Trump -2.592e+02 2.263e+01 -11.457 < 2e-16 \*\*\*

DJI 1.006e-02 6.143e-04 16.382 < 2e-16 \*\*\*

---

Residual standard error: 20.94 on 204 degrees of freedom

Multiple R-squared: 0.6985, Adjusted R-squared: 0.6956

F-statistic: 236.3 on 2 and 204 DF, p-value: < 2.2e-16

**## Model B: Residuals from the above model regressed on lagged residuals**

lm(formula = Residuals ~ Lag.Residuals, data = stkPres\_pre)

Estimate Std. Error t value Pr(>|t|)

(Intercept) -0.06614 0.60831 -0.109 0.914

Lag.Residuals 0.91606 0.02967 30.879 <2e-16 \*\*\*

---

Residual standard error: 8.73 on 204 degrees of freedom

Multiple R-squared: 0.8238, Adjusted R-squared: 0.8229

F-statistic: 953.5 on 1 and 204 DF, p-value: < 2.2e-16

**## Model C: Use the Cochrane.Orcutt function on the model produced in Model A**

summary(cochrane.orcutt(OLS.1))

Estimate Std. Error t value Pr(>|t|)

(Intercept) 7.6874e+01 2.3568e+01 3.262 0.001298 \*\*

Trump -4.4314e+01 1.0791e+01 -4.107 5.818e-05 \*\*\*

DJI 6.9302e-03 5.9943e-04 11.561 < 2.2e-16 \*\*\*

---

Residual standard error: 5.1102 on 203 degrees of freedom

Multiple R-squared: 0.411 , Adjusted R-squared: 0.4052

F-statistic: 70.8 on 2 and 203 DF, p-value: < 4.648e-24

Durbin-Watson statistic

(original): 0.18059 , p-value: 1.361e-40

(transformed): 1.70952 , p-value: 1.933e-02

* 1. Why should we be concerned about possible autocorrelation for Model A?

We should be concerned for two reasons. Firstly, the facebook stock price is non-stationary throughout the period and Trump’s probability was non-stationary in the beginning of the campaign. This can give spurious correlation. Secondly, if the variables are autoregressive we will overestimate our number of observations (effectively).

fsdfsdf

* 1. Is there autocorrelation? Be specific about the evidence and indicate whether there is strong, weak or no evidence of autocorrelation.

**Yes there is. The simple autoregression test found a rho of .9, which is high. No autocorrelation means a rho of .9.**

**In the CO model I’m not sure which value is rho, usually that output is printed as well.**

* 1. Does Model C account for autocorrelation? Why or why not? If so, how?

**Model C is a start, but I am unsure if we are done. Facebook stock prices should be non-stationary, but model c assumes stationary autoregression. It’s possible that stocks are stationary over certain time-frames, like a year.  
Also non-stationarity should bias the beta\_1 upward, be we find a negative beta\_1 in model C (to the extent that stocks usually gain value at 4%/year).**

* 1. Below are the results from a differenced model. Do you think this model with have more or less autocorrelation? Why or why not?

**## Model D: Variables with “\_ret” in name are differenced (measured as the percent change each day)**

lm(formula = FB\_ret ~ Trump\_ret + DJI\_ret, data = stkPres\_pre)

Estimate Std. Error t value Pr(>|t|)

(Intercept) 0.001461 0.001539 0.949 0.344

Trump\_ret -0.064473 0.015657 -4.118 5.62e-05 \*\*\*

DJI\_ret 0.877515 0.061053 14.373 < 2e-16 \*\*\*

---

Residual standard error: 0.02177 on 197 degrees of freedom

Multiple R-squared: 0.5183, Adjusted R-squared: 0.5134

F-statistic: 106 on 2 and 197 DF, p-value: < 2.2e-16

**This model deals with the non-stationarity problem I described above! Great!**

**The model should have less autocorrelation therefore.**

* 1. Below are results for stationarity tests for the variables in Models A and D. Given these results would you recommend using a differenced model? Why or why not?

> library(tseries)

> adf.test( na.omit(stkPres\_pre$Trump), k = 1)

Dickey-Fuller = -3.4124, Lag order = 1, p-value = 0.05369

alternative hypothesis: stationary

> adf.test( na.omit(stkPres\_pre$FB), k = 1)

Dickey-Fuller = -2.1079, Lag order = 1, p-value = 0.5307

alternative hypothesis: stationary

> adf.test( na.omit(stkPres\_pre$Trump\_ret), k = 1)

Dickey-Fuller = -10.889, Lag order = 1, p-value = 0.01

alternative hypothesis: stationary

> adf.test( na.omit(stkPres\_pre$FB\_ret), k = 1)

Dickey-Fuller = -10.57, Lag order = 1, p-value = 0.01

alternative hypothesis: stationary

I would recommend using a differenced model. In the two first regressions we almost reject the stationary hypothesis, indicating the results are non-stationary, and in the 3rd and 4th we soundly reject the stationary hypothesis. Nonstationary results should be analyzed with differenced models. That said, I would include both were I to publish such a paper.

1. For this problem we look at the influence of a tax on beer on traffic fatalities. We have data on the 48 continental U.S. states from 1982 to 1988 on per capita traffic fatality rate (the dependent variable), the beer tax (standardized), income (standardized) and whether there were mandatory jail sentences for drunk driving.



* 1. Describe a state characteristic that is controlled for in model (3) but not in model (1).

In some states people drive more than in others. This driving is in the error term. But high driving is correlated with libertarian culture, which may restrain vice taxes. Thus high-driving is in the error term and correlated with the dependent variable.

* 1. Tell a story that could explain why the coefficient on beer tax flips in sign and significance from model (1) to model (3).

State legislatures pass high beer taxes in response to high drinking rates. So the states that pass beer taxes also have culturally-driven higher drinking rates, creating endogeneity that biases the relationship in the pooled model.

However, beer taxes do reduce the amount of drinking. By controlling for SFE we remove these cultural attributes and find a negative effect of taxes on drunk driving.

* 1. What is the assumption that needs to be true for the OLS pooled model to produce unbiased

estimates? Is this assumption true for this data? Explain **using evidence from results table**.

The independent variable of interest must be uncorrelated with the error term, otherwise we have endogeneity. From model 2 to model 4 the coefficients on two IV’s switch direction, which suggests that some factor correlated with the IV’s was in the error term but removed when SFE’s were included.

* 1. The coefficient on income changes a lot from specification (2) to specification (4). Explain.

States with higher income had higher traffic fatalities, possibly because they had more cars or more driving. However, with fixed effects we are seeing the effect of variation in state income over the period. So states that grew a lot had more traffic accidents (I am assuming the data is on the state level, not individual or county).

* 1. Below we provide results from a hypothesis test. Explain.



This test checks if the effect of both beertax and income could be 0. If finds with high confidence that they do jointly have an effect.

Both were significant in the original specification, so this result is not surprising.

.

1. Suppose we have cross sectional data on corona virus deaths by county in April 2020 in the U.S. We are interested in knowing whether social distancing reduced mortality from the virus. Our measure of social distancing equals 1 if a county was subject to stay at home orders by April 10, 2020. For each of the following variables, discuss whether you should control for, should not control for it or it doesn’t matter. Explain using statistical concepts.
   1. Health care spending per capita in 2019

Health care spending is correlated with SHO’s, and does affect mortality. Also, the spending in 2019 is not downstream of SHO’s. Therefore recommend to include.

* 1. Change in mobility as measured by cell phone trackers in April 2020

The change in mobility is downstream of the cell phone tracks, and is affected by possible U variables like economic changes. This creates risk of collider bias. Recommend against inclusion.

* 1. Suppose someone proposed using whether a state had a Republican governor as an instrument for social distancing policy. Would that be a good instrument?

That would not be a good instrument. Republican governors would likely pass the inclusion condition. However, they would not pass the exclusion condition.

* Republican governors are correlated with republican voters, who assign lower salience to the pandemic.
* Republican governors can change state policy in other ways than SHO’s

1. Suppose you have a health care data set with 3 variables: whether someone tested positive for coronavirus (the dependent variable), whether they were hospitalized and whether they were pregnant. Assume for the sake of this question that pregnant women are no more or less likely to get corona virus than others. Describe a model specification in which you would (given sufficient data) get a statistically significant coefficient on the pregnant variable. Explain.

**Positive = B\_0 + B\_1 \* Hospitalized + B\_2 \* Pregnant + epsilon**

**A simple way to see this is to suppose your friend says “I just got out of the hospital” you would assign a high probability to them having covid. If they then revealed that they were pregnant you would think “oh, they were probably just in the hospital for their baby” and you subjective probability that they have covid would decrease.**

**If I have time I will return and formalize this.**

1. Suppose that a group of countries provide reports on human rights in Egypt each year and that these reports include a rating on a scale of 1 to 5, with 1 being human rights are being respect and 5 indicating that human rights are being grossly violated. As a first cut on our analysis we wish to estimate average global opinion human rights in Egypt using this variable. A challenge is that only half of U.N. member states provided this information.
   1. Write down a model for selection and outcome.

Pr(Selection) = B\_0 + B\_1 \* Needs\_egypt\_favors + B\_2 \* cares\_about\_HR + B\_3 \* sovereignty\_importance + mu

Report = B\_0 + B\_1 \* Needs\_egypt\_favors + B\_2 \* cares\_about\_HR + B\_3 \* sovereignty\_importance + epsilon

B\_1 will be negative in the first equation and negative in the second. So the sample that reports is unusually low in favor\_need, so they would report more violations than an unbiased sample. That is a belief falsification model though, so we need to distinguish strategic bias from statistical bias.

* 1. Discuss how one could weight the data. Use continent as the weighting variable; that is we know how many countries responded by continent and we know how many countries there are on each continent.

**We could increase the importance of each observation by the inverse of the response rate of the continent.**

* 1. What is the necessary condition for weighting to be valid?

**There must be little remaining correlation between mu and epsilon (between the error term of the selection equation and the error term of the model).**

**Personally, I would add Voeten’s UN voting factions as a regressor.**

1. The models below are based on an analysis of 20 Democratic members of Congress. In particular, we would like to understand which candidates use the term “Medicare for all” on their website.
   1. Here is the distribution of the dependent variable. Draw a likelihood function for a model in which we seek to use MLE to estimate p, the probability a candidate (in the sample) will refer to Medicare for All.

> table(dta$medicare4all)

0 1

14 6

Pr(MFA) = phi^-1 ( B\_0 + B\_1 \* age + B\_2 \* committee rank + epsilon).

* 1. Here are LPM results. The variable nominate\_dim1 is the “nominate” ideology score, with more negative values indicating more liberal members. The variable ranges from -0.7 to 0 for this sample. What is the fitted probability of referring to Medicare for All for the most liberal member?

lpm.1 = lm(medicare4all ~ nominate\_dim1, data = dta)

Estimate Std. Error t value Pr(>|t|)

(Intercept) 0.1100 0.1986 0.572 0.575

nominate\_dim1 -1.0000 0.5000 -2.00 0.024

Residual standard error: 0.4789 on 18 degrees of freedom

Multiple R-squared: 0.1401, Adjusted R-squared: 0.09234

F-statistic: 2.933 on 1 and 18 DF, p-value: 0.104

Lol I forgot about ideology, I am too cynical.

The most liberal member has pr(MFA) = phi^-1 (.2 – 1 \* -.7) = phi (.9) = ~.81= 81%

* 1. For the probit results below, what is the fitted probability for someone with nominate\_dim1 = -0.5? (that’s negative 0.5, to be clear)

prob.1 = glm(medicare4all ~ nominate\_dim1,

family = binomial(link = probit))

Estimate Std. Error z value Pr(>|z|)

(Intercept) -1.0000 0.5654 -1.876 0.0606 .

nominate\_dim1 -2.0000 1.4027 -1.678 0.0934 .

Phi\_inv(-1 – 2\*-.5) = phiinv(0) = .5

* 1. Explain a case in which probit will provide similar results as a LPM model.

The LPM is linear, so the second derivative is always 0. The second derivative of the probit function is only 0 at a probability of .5.

If the average probability is .5 and the IVs explain relatively little of the variation. The second derivative of the probit function is 0 at probability=.5

1. There is considerable evidence that exposure to lead can lead to development damage, particularly related to impulse control. Feigenbaum and Muller (2015) analyze the relationship between crime and lead water pipes in mid-twentieth U.S. cities from 1921 to 1936. Here, we’ll explore a simplified version of their data and analysis. The dependent variable will be either Homicides\_per\_million or LnHomicide\_per\_million.

**Independent variables**

lead: dummy variable indicating presence of lead pipes

black\_share: black population share

acidity: a pH based measure of water acidity, scaled so that high values

indicate acidic water (ranges from 0.7 to 7.8)

Acidity\_x\_Lead: an interaction of the lead and Acidity variables. Lead leaches

(absorbs) into water more readily in acidic water.

foreign\_share: foreign-born population share

single\_men: share of single men between 18 and 40

region dummies: northeast (NE), midwest (MW), south and west

year: a variable indicating the year of the observed data

**MODEL A: Base model**

lm(homicide\_per\_million ~ lead + black\_share + foreign\_share + single\_men + NE + MW + South)

Estimate Std. Error t value Pr(>|t|)

(Intercept) 39.5000 5.9235 6.685 2.48e-11 \*\*\*

lead 9.8000 2.1552 4.580 4.72e-06 \*\*\*

black\_share 4.7000 0.1440 32.734 < 2e-16 \*\*\*

foreign\_share -0.5000 0.1104 -4.931 8.38e-07 \*\*\*

single\_men 2.8000 0.3292 8.631 < 2e-16 \*\*\*

NE -33.0000 4.1927 -8.058 8.95e-16 \*\*\*

MW -8.0000 4.3322 -1.890 0.0588 .

South 53.0000 7.1079 7.496 7.36e-14 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 89.84 on 7426 degrees of freedom

Multiple R-squared: 0.5418, Adjusted R-squared: 0.5414

F-statistic: 1255 on 7 and 7426 DF, p-value: < 2.2e-16

1. How do homicides in the South compare to homicides in the Northeast in Model A (controlling for the other independent variables)? Provide a specific number.

Cities in the NE had 86 fewer homicides per million (hpm) than cities in the South, controlling for other variables.

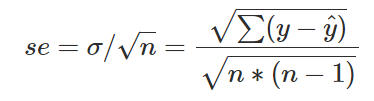
I would point out that other parasites and diseases were more common in the South at the time, and could produce a lead-like effect.

1. The coefficient on Midwest is not statistically is not statistically significant at conventional levels. Would MW be statistically significant if South were the reference category? Why or why not? You will not be able to definitively answer this based on the table, but can give a very good guess.

If South were the reference category, then MW would be stat sig. The value on MW would then be 60 and the standard error is currently 4, so the t-stat would be about 15.

1. What is the standard error of the regression () in Model A? Explain what it means.

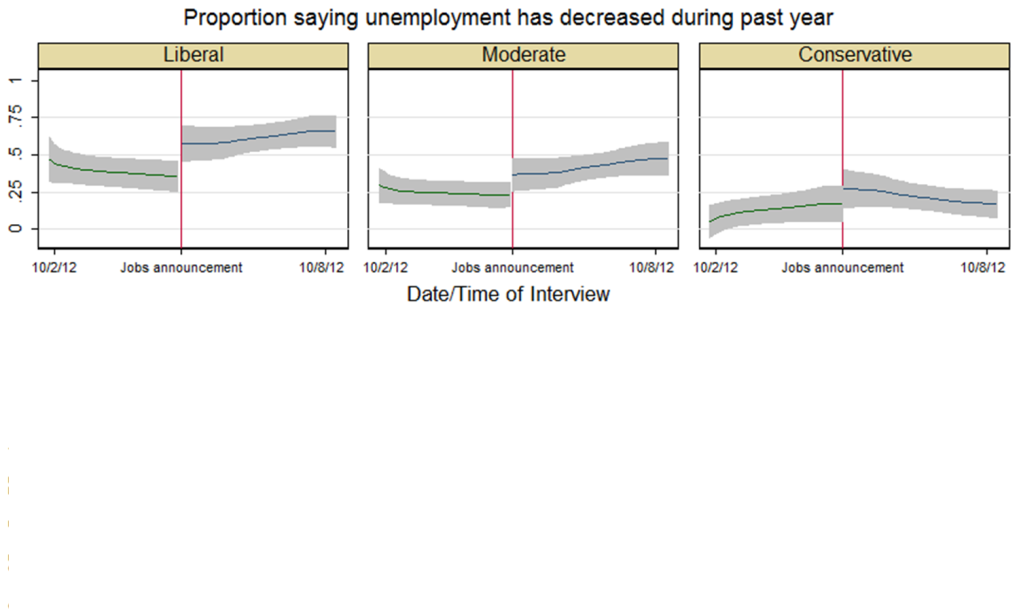
**The standard error is 90. The formula for standard error is (below). More intuitively, it is the square-averaged differences between the observed and predicted value. This is very similar to the average absolute value of the residuals.**



1. We want to know if information from the government affects public views about the economy. We’re particularly interested in the possibility that people who identify with different political ideologies (as opposed to political parties)—liberals, moderates and conservatives—respond differently.

The September 2012 jobs report from the federal government indicated that unemployment had fallen. The figures below are from time series data on the percent of a national sample who said that unemployment has decreased during the last year. The panels show these series for liberals (panel on the left), moderates (panel in the middle) and conservatives (panel on the right). We want to use this data to assess whether the jobs report caused liberals, moderates and conservatives to change their assessments of unemployment in the US.

In each plot, the Y-axis indicates the percent of liberals (or moderates or conservatives) who said at time t that unemployment has decreased during the past year. The horizontal axis indicates the date of the survey.



http://themonkeycage.org/blog/2012/11/04/ideological-bias-and-the-jobs-report/

1. Describe how to do a “regression discontinuity” analysis of whether the treatment affected the percent of people who said unemployment has increased (where the treatment is the announcement of jobs numbers on 10/5/2012, labeled “jobs announcement” in the figure). In your explanation, write a ***single equation*** that we could use separately for each of the three groups (liberals, moderate and conservatives); in other words, you will NOT need to have a variable indicating ideology in the equation. Be sure to allow for slopes to vary on different sides of the job announcement

**y = B\_0 + B\_1 \* date + B\_2 \* liberal + B\_3 \* conservative + B\_4 \* announcement + B\_5\* announcement\*liberal + B\_6\*announcement\*conservative + B\_7\*announcement\*date + epsilon**

**You could allow the slopes to vary between ideologies on either sides with four more terms.**

1. For each of the three ideological groups, describe the sign of the treatment effect as estimated by regression discontinuity.

**In liberal and moderate the treatment will be positive. In the conservative group the announcement variable will be 0 or maybe slightly positive. The conservative slope will turn downward however.**

1. Discuss how to diagnose potential problems in a RD analysis.

**You should carefully check that the populations before and after the event are the same. This can be checked with a simple regression of the variable of interest on the correlate variables.**

**You should also check the histogram of observations graphically to ensure there is no bump of observations just after the discontinuity, and graph you Y variables and Y-hats.**

1. SHORT ANSWER (3 points each)
   1. Which of the following is R code that implements a two way fixed effect model?
      1. lm(Y~X1 + X2, twoways)

**Answer**

**ii**

* + 1. plm(Y~X1 + X2, index = c(state, year), effect = “twoways”)
    2. \cexY 2-way !(\*state[0])//4[\end].
    3. xtreg Y X1, fe
  1. Suppose we have a survey and our sample has 20% white working class citizens, which happens to be the exact proportion in the overall population as reported by the Census. True or false: the non-response is ignorable.

**Explanation**

**This is not enough to show that the non-response is non-ignorable. We must also show that no other factors in the selection error term are in the survey response error terms.**

**True or false**

**False**

* 1. True or false: If an instrument does a poor job predicting X, 2SLS will produce a coefficient with more bias than OLS.

**Explanation**

**The poor prediction in the first stage does not create bias. The formula for bias in 2SLS is**

**E[B\_1 hat] = B\_1 + corr(z,e) \* sigma\_e / corr(z,x) \* sigma\_x**

**However, if there is any corr(z,e) it will cause larger bias with a weak instrument.**

**True or false**

* 1. True or false: Adding a lagged dependent variable as an independent variable will typically cause autocorrelation to decrease.

**Explanation**

**Autocorrelation/autoregression are different from non-stationarity. It can actually increase bias.**

**True or false**

**F**

* 1. True or false: Non-linear models are always better for regression discontinuity.

**Explanation**

**A slope change are some other point in time (not the focal point) could create a misleading impression of a regression discontinuity where non exists. You should plot your y-values by time and y-hats to check for this.**

**True or false**

**F**