GSERM 2020 – Regression for Publishing

Final Examination

Instructions

1. This examination comprises three questions. **You are to choose one question and answer it.** *Do not attempt to answer all three questions.*

2. Your answer should take the form of an empirical analysis of the data that answers the question(s) asked. That should include plots, tables, text, and any other techniques you think will be useful in answering the question. There are no minimum or maximum length requirements for any answer.

3. All data necessary for completing each and any of the exam questions are available on the course CANVAS page, and at the course github repository.

4. You have two timeline options for completing the exam:

   (a) You may choose to complete the exam "in class" and submit it (following the instructions below) no later than 19:00 CET on Friday, June 19, 2020. All exams submitted at or before this time will be graded together.

   (b) Alternatively, you may submit the exam as a "take home" exam, no later than 17:00 CET on Monday, June 22, 2020. All exams submitted at or before this time (but after 19:00 on 19 June) will be graded together.

5. When you have completed the exam, you will submit your answer to the instructor electronically as a PDF file by e-mailing it to `zorn@psu.edu`.

Exam Questions

Question One. The topic of this question is leverage, discrepancy, influence, and outlier detection. The substantive question is crime, and the data in question (`GSERM-2020-Final-Exam-Crime-Data.csv`) are 1993 statistics on a number of state-level variables ($N = 51$, since we have the District of Columbia in there too) in the United States. Specifically, they are:

- `state` is a two-letter USPS state identifier.
- `vcrimerate` is the violent crime rate for that state in 1993 – that is, the number of violent crimes reported per 100,000 population.
- `blackpct` is the percentage of the state's population that is African-American.
- `urbanpct` is the percentage of the population that lives in urban/metropolitan areas.
- `unemployed` is the unemployment rate (as a percent of the state population).
- `police` is the number of police officers per 100,000 in population.
- `prisonpop` is the number of people imprisoned per 100,000 in population.

The dependent variable of interest is the violent crime rate; for purposes of this exam, we'll assume that all five of the other variables belong on the right-hand side of the regression model.

Directions

1. Estimate the model discussed above, and (very) briefly discuss your "findings."
2. Address the question of whether – and, if so which of, and to what extent – the findings are being driven by a small number of particularly influential observations. It is probably wise to start with / rely upon the discussion from the June 17 class for this, though you should also use your own judgement as to what kinds of things can and should be considered.
3. Finally, estimate and provide a brief discussion/justification of a "final" model – that is, one that deals with outliers, if any. Please note: *Your "final" model need not be any different from your initial one.* What I **do** ask, however, is that you justify your decisions about your final model in light of whatever you find (or do not find) in your analysis of influence and outliers.

Question Two. In December 1998, the members of the U.S. House of Representatives voted on four "articles of impeachment" of President William Clinton. Articles which passed would then be sent to the U.S. Senate for trial; if convicted, President Clinton would have been removed from office. The articles, therefore, were widely (and correctly) viewed as a referendum on President Clinton's legitimacy and leadership.

We will examine data on the House's impeachment votes on President Clinton, in an effort to assess the important predictors of those votes. The data (available on github and CANVAS as `GSERM-2020-Final-Exam-Impeachment-Data.csv`) comprise the 433 voting members of the U.S. House of Representatives. The variables in the data are as follows:

- `votesum` is the number (from zero to four) of articles in favor of which a given member voted. That is, a `votescore` of 1 means that the member in question voted in favor of one of the four articles of impeachment; a `votescore` of two means s/he voted for two of those articles, and so forth. Since a vote in favor of an article was a vote against President Clinton, higher numbers denote stronger opposition to Clinton's leadership. All 433 members voted ("yea" or "nay") on all four articles.
- `pctbl96` equals the percentage of that member's House district that was African-American, as of 1996.
- `unionpct` is the proportion of that member's House district that were members of organized labor. Both African-Americans and union members are generally more supportive of the Democratic party, and of Clinton.
- `clint96` equals the percentage of the two-party presidential vote that President Clinton received in that district in the 1996 election.
- `GOPmember` is a dummy variable, coded 1 if the representative was a member of the Republican party and 0 if they were a Democrat.
- `ADA98` is the member's 1998 Americans for Democratic Action score, a measure of how liberal that member's voting behavior is; it ranges from a low of zero (most right/conservative) to a high of 100 (most left/liberal).

Directions

1. Choose a statistical model from those we discussed in class to assess the relationship between the five covariates and `votesum`. Discuss briefly the reason(s), statistical and/or substantive, for your choice.
2. Fit the model, and discuss in substantive terms the marginal impact of each of the covariates on the response variable `votesum`, using whatever methods you deem appropriate.
3. Finally, compare your findings using your chosen method to an alternative (and presumably less plausible) statistical approach. What, if any are the differences between the results, and which do you feel are more accurate/correct?

Question Three. The substantive focus of this question is the influence of political ideology on U.S. Supreme Court voting over time. In particular, the conventional wisdom states that political actors often become more politically moderate over time. To test this hypothesis, you'll examine data on the voting patterns of justices sitting on the Supreme Court during the Vinson, Warren, Burger, and Rehnquist courts (1946-1994) ($N = 32$, $T = 49$, unbalanced). The data are available on the course CANVAS page and github repository as `GSERM-2020-Final-Exam-Court-Data.csv`. The variables are:

- `justice` (the justice identifier variable),
- `year` (the year identifier),
- `civlib` (the percentage of left/liberal votes cast by that justice in civil rights and liberties decisions in that year),
- `econs` (the percentage of left/liberal votes cast by that justice in economics decisions in that year),
- `score` (the normed "Segal/Cover" (1989) ideology score of the justice, ranging from -1 (most right/conservative) to 1 (most left/liberal) and
- `tenure` (the number of years the justice has served on the Court, as of that year).

If the conventional wisdom is correct, one possible manifestation is that the effect of `score` on voting liberalism should be positive, but the interaction of `score` and `tenure` should be negative (as justices moderate their extremism later in their careers).

Directions

1. First, examine voting liberalism in civil rights and liberties cases (`civlib`).
    - Specify and estimate one or more panel-data models for this outcome, and discuss your results, both substantively and statistically.
    - Briefly discuss the basis for your choice of model(s) in the first step.
2. Repeat the above steps for the variable on economics cases (`econs`).
3. Talk in general terms about which model(s) you prefer for these analyses, and why.