CSSS/POLS 512:
Time Series and Panel Data for the Social Sciences

Problem Set 3

Professor: Chris Adolph, Political Science and CSSS

Spring Quarter 2020

Due in class on Tuesday 26 May 2020

General instructions for homeworks: Homework can be handwritten or typed. For any exercises done with R or other statistical packages, you should attach all code you have written and all (interesting) output. Materials should be stapled together in order by problem. The most readable and elegant format for homework answers incorporates student comments, code, output, and graphics into a seamless narrative, as one would see in a textbook. Working in groups on R code is allowed, but (1) each member of the group must provide his or her own writeup and (2) you must list all members of your group.

Problem 1: Analyzing partisan seat shares in US state legislatures

[100 points total.] In the last homework, we examined the dynamic relationship between the number of seats held by Democrats in the US House of Representatives and a set of political and economic variables testing the effect of election cycles, presidential coattails, and judgment of the president's economic performance on the size of the Democratic majority. If we turn to American state legislatures, we can test similar questions but with much more data.

We will restrict our scope in several ways to make the problem more manageable (a more complete analysis might avoid these restrictions, but would require a more flexible panel data model). First, we look only at legislatures elected in 1978 or later, up through the 2016 elections. Second, we exclude South Carolina (for which some data are missing), Nebraska (which has non-partisan legislative elections) as well as Alaska and Hawaii. Third, we restrict our attention to the lower house of each state's legislature (typically known as the "House"). Fourth, we include only states that elect house members to two-year terms, which excludes Alabama, Louisiana, Maryland, Mississippi, and North Dakota. Finally, states vary in the year they hold gubernatorial elections, but most hold them only in the sequence of years that runs 2010, 2014, 2018. . .. We include only these states (note this means we also exclude New Hampshire and Vermont, which hold gubernatorial elections in every even year). As a result, our scope includes 28 states each observed for 20 lower house election cycles.

Because each state has a varying number of total house seats, in this analysis, we will focus on the *share* of seats held by the Democrats. Following on our analysis in the last homework, we consider three kinds of explanations for shifts in Democratic control.

Economic performance. When a state's unemployment rate is higher than the long-term national average on election day, it may hurt the electoral prospects legislators belonging to the party in government – but for state legislative elections, which party is that? The party of the governor or the party of the president?

Spillovers from votes for president, congress, or governor. Few voters pay close attention to state legislative elections, so partisan swings in the state house may reflect spillovers from votes on the presidential, congressional, or gubernatorial races. This could involve presidential coattails – where a new party sweeps into the presidency and lower offices at the same time. It could also take the form of "gubernatorial coattails" within a state, whereby a new governor brings in a larger seat share for his party. Finally, just as voters use congressional races to vote against the president during midterms, they could use state house races in the same fashion.

Trends. American politics passed through a major re-alignment in the late twentieth century as white voters in the South switched from solid support of the Democratic Party to solid support of Republicans. This proceeded at different rates in different states. Possibly in reaction to this and other socioeconomic trends, other regions may be shifting in partisanship as well.

Variable	Description
`State`	Two letter state abbreviation
`Statename`	Full state name
`FIPS`	Unique numeric code for each state (non-consecutive)
`Year`	Start year of the legislative session; duplicated as `trend`
`GovCycle`	1 indicates states that last held gubernatorial elections in 2014, 2 indicates 2015, 3 indicates 2016, 4 indicates 2017
`HouseTerm`	The length in years of elected house terms
`DemHouseShare`	Proportion of lower house seats held by Democrats $[0, 1]$
`PartisanMidterm`	1 for legislatures elected in midterms in which the Democrats held the presidency, -1 for legislatures elected in midterms in which there was a Republican president, and 0 for legislatures elected in other years
`PresUnem`	equal to the difference between this state's election-year unemployment and the national 1978–2016 average (5.97%) for legislatures elected when a Democrat was president, and to $-1\times$ this quantity for legislatures elected under Republican presidents
`GovUnem`	equal to the difference between this state's election-year unemployment and the national 1978–2016 average (5.97%) for legislatures elected when a Democrat was governor, and to $-1\times$ this quantity for legislatures elected under Republican governors
`PresCoattails`	1 if the presidency shifted to the Democrats when this legislature was elected, -1 if the presidency shifted to the Republicans, and 0 if the party of the president was unchanged
`GovCoattails`	1 if the governorship shifted to the Democrats when this legislature was elected, -1 if the governorship shifted to the Republicans, and 0 if the party of the governor was unchanged
`Midwest`	Region dummy for the Midwest, broadly defined
`West`	Region dummy for the West, broadly defined
`Northeast`	Region dummy for the Northeast, broadly defined
`South`	Region dummy for the South, broadly defined

Table 1. Codebook for State House Seats data. Data are in `statehouse.csv`, and are taken from the Bureau of Labor Statistics (unemployment), the Book of States (legislative shares) and Wikipedia (governors and cycles), or constructed from these data by your instructor & TA.

Throughout the homework, we will consider two basic specifications. In the first specification, called M1, we control for midterm effects, presidential and gubernatorial unemployment effects, and presidential and gubernatorial coattails.[1] In the second specification, called M2, we control for all these variables and region specific trends.[2]

In the file `statehouse.csv`, you will find the variables described in Table 1. Examine the data file and note well the behavior of these variables over time. Then work through the following exercises. *This is potentially a long assignment with several tricky steps, so start now, work in groups, and email the class and instructors often for advice. It's okay to share as much code as you like in seeking help – we're all working together on this one!*

a. [5 points.] Preprocess the data to remove unwanted states. Specifically, you should create a new dataframe that removes all states that have a `GovCycle` other than 1 or a `HouseTerm` other than 2. Confirm that you now have 28 states and 20 election cycles left in your dataset. Then, create the first four lags of the outcome variable using `lagpanel()` in the `simcf` library.[3]

b. [15 points.] For each state left in the analysis, plot the time series `DemHouseShare` and plot its ACF and PACF (there is no need to show these plots; please offer general impressions). Perform augmented Dickey-Fuller and Phillips-Peron tests for unit roots for each time series and examine their distribution using histograms. Finally, conduct two Im-Pesaran-Shin panel unit root tests of `DemHouseShare`, assuming fixed intercepts and trends, respectively. Sample code for the first test is given by:

```
ts <- with(statehouse,
           data.frame(split(DemHouseShare, as.character(State))))
purtest(ts, pmax = 4, exo = "intercept", test = "ips")
```

Describe your findings, being sure to describe what kind(s) of time series process may be at work.

1 The specific variables to be control are thus `PartisanMidterm`, `PresUnem`, `GovUnem`, `PresCoattails`, and `GovCoattails`.

2 To add region specific trends, you could include the following terms in a model specification: `trend:South`, `trend:Midwest`, `trend:Northeast`, and `trend:West`. Note that in a model without state fixed effects, you should also control for the regions themselves (why can't you do this in a fixed effect model?).

3 I recommend naming these lags `DemHouseShareL1`, `DemHouseShareL2`, etc.

c. [10 points.] Fit M1 using a model that treats the intercept as a state random effect. To deal with the dynamic nature of the outcome, consider and estimate a variety of ARMA(p,q) specifications.[4] Using the insights gleaned from part **b.** and goodness of fit tests, select the best model of the time series and call this the "best RE model for M1."

d. [10 points.] Fit M2 using a model that treats the intercept as a state random effect. To deal with the dynamic nature of the outcome, consider and estimate a variety of ARMA(p,q) specifications. Using the insights gleaned from part **b.** and goodness of fit tests, select the best model of the time series and call this the "best RE model for M2."

e. [10 points.] Fit M1 using a model that treats the intercept as a state fixed effect. To deal with the dynamic nature of the outcome, consider controlling for one or more lags of the outcome variable.[5] Using the insights gleaned from part **b.**, goodness of fit tests, , and tests of serial correlation, select the best model of the time series and call this the "best FE model for M1."

f. [10 points.] Fit M2 using a model that treats the intercept as a state fixed effect. To deal with the dynamic nature of the outcome, consider controlling for one or more lags of the outcome variable. Using the insights gleaned from part **b.**, goodness of fit tests, and tests of serial correlation, select the best model of the time series and call this the "best FE model for M2."

g. [30 points.] Using each of four "best" models, forecast what will happen to the size of the Democratic majority in the average state in the 2019 and 2021 sessions for the following single scenario. Assume the Democrats resume this state's governorship in 2019 and the presidency in 2021, and compute appropriate counterfactual values of `PartisanMidterm`, `PresCoattails`, `GovCoattails`. Assume unemployment falls to 3.6% for both elections and construct `PresUnem` and `GovUnem` accordingly. Set all trend variables at the average value they will take across regions in 2019 and 2021, respectively. Make appropriate assumptions for the prior value(s) of the outcome variable (e.g., the average Democratic House share in 2017).

4 For this problem and the next, I recommend you use `lme` in the `nlme` package for estimation.

5 For this problem and the next, I recommend you use `plm` in the `plm` package for estimation and the lags premade using `lagpanel` as controls.

For each model, report or graph the predicted Democratic majority and its 95% confidence (or predictive) interval for the 2019 and 2021 sessions. Describe the substantive impact of your forecast results in as much detail as you feel comfortable, as well as how much confidence we should have in the forecasts. Be sure to consider the scale of the outcome variable in assessing what counts as a substantively large or small change.

NB: As a check on your work, report the table of counterfactual covariate values you used to make your forecasts. Be very careful when constructing these values to capture to logic of the covariates; each one is tricky in its own way. To carry out the forecasts, use the `simcf` library's `ldvsimev()`, pay close attention to the example code, and think through all modifications you need to make.

h. [10 points.] Using everything you have learned in this assignment and in the course, which of the four best models should we use to write-up our results, and why? (You may argue for multiple models if you think that's appropriate.) What are your final substantive conclusions? Substantively, does it make much difference which model we choose? How does this affect the way you would write this analysis up in a paper?

There's a lot we've left out to make this analysis managable: panel-corrected standard errors, cross-validation tests of fit, efficient presentation of the full model results through simulation, and so on. Projects using panel data are often complex, with many modeling choices and opportunities to exploit the model to answer substantive and statistical questions. This assignment is just a start...