Essex Summer School in Social Science Data Analysis
Panel Data Analysis for Comparative Research

Practice Problems

Christopher Adolph
Associate Professor, Department of Political Science
and Center for Statistics and the Social Sciences
University of Washington, Seattle, USA

General instructions for homeworks: Homework can be handwritten or typed. For any exercises done with `R` or other statistical packages, you should attach (a.) all code you have written, and (b.) all 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.

For many of these problems, students will find it useful and interesting to analyze their own time series cross-section dataset, rather than the default datasets provided.

Problem 1: R practice and linear regression review

For this problem, you will need the `democracy.csv` file in your working directory. This file contains data from Przeworksi, Alvarez, Cheibub & Limongi. *Democracy and Development: Political Insitutions and Well-being in the World, 1950–1990*. The data have been slightly recoded, to make higher values indicate higher levels of political liberty and democracy. Missing values are coded as "." For this problem, we will not make use of the time series cross-section aspect of these data.

The answer to each of the following questions can be found using a single line of `R` code. The code you use does not need to be this concise, but if it's taking a lot of code, there is a simpler way. Regardless, it is essential you show your work and all code.

a. Load the Democracy dataset into memory as a dataframe.

b. Attach it so that each variable in the dataset is accessible by name.

c. Report a correlation matrix of all the variables in the dataset. Watch out for missing values.

d. Create a histogram for political liberties

e. Create a histogram for GDP per capita.

f. Create a scatterplot of political liberties against GDP per captia.

g. Create a boxplot of GDP per capita for oil producing and non-oil producing nations.

Variable	Description
GDPW	GDP per capita in real international prices
EDT	average years of education
ELF60	ethnolinguistic fractionalization
MOSLEM	percentage of Muslims in country
OIL	whether oil accounts for 50+% of exports
STRA	count of recent regime transitions
NEWC	whether county was created after 1945
BRITCOL	whether country was a British colony
POLLIB	degree of political liberty (1–7 scale, rising in political liberty)
REG	presence of democracy (0=non-democracy, 1=democracy)

h. On average, how many times smaller or greater is GDP per capita in countries with at least 40 percent Catholics, compared to those with fewer than 40 percent Catholics?

i. On average, how many times smaller or greater is GDP per capita in countries with more than 60 percent ethnolinguistic fractionalization, compared to those with less than 60 percent ethnolinguistic fractionalization?

j. What was the median average years of education in 1985 for all countries?

k. What was the median average years of education in 1985 for democracies?

l. What were the 25th and 75th percentiles of ethnolinguistic fractionalization for new countries?

m. Specify a linear regression of your choosing and run it using `lm()`. For our purposes, this regression need not be (and is in fact very unlikely to be) a theoretically and statistically sound model. For pedagogical purposes, it will help to have a non-binary variable as the response, and at least one non-binary variable among the covariate. Print and explain the summary table given by `lm`.

n. What are some possible sources of bias, inefficiency, or incorrect standard errors in the regression you ran? Note each briefly, but list as many as you can.

Problem 2: Some elementary R programming

As you saw in the last problem, R can do the same basic statistical operations as any other statistics package. Where R differs from most packages is that allows you to "easily" program your own statistical functions. Let's get acquainted with some programing basics.

a. Generating random data. *i.* Using `sample()`, randomly sample 5 elements from the set $\{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$, with replacement (i.e., if you draw a 5 on the first try, you could still draw a 5 on the next draw with probability 0.1). *ii.* Using `rnorm`, randomly sample ten draws from the normal distribution with mean 10 and variance 2.

b. Sourcing in scripts. Write a small script that performs the two exercises in a, and prints the results to the screen. Use the `source()` command to run the script.

c. Writing a simple function. Write an `R` function, called `squared()`, that takes an input, a, and returns a^2.

d. Writing a useful function. One example of a useful function is one which returns the percentiles of a vector of data. (If you are unfamiliar with percentiles, note that the median is the 50th percentile, because 50 percent of the data lies below it; the 23rd percentile is the value below which 23 percent of the data lie, the 99th percentile is the value below which 99 percent of the data lie, etc.)

We are going to write a function, `ptile()`, which takes as inputs a vector **x** of data, and a single number p, which is a percentile (for convenience, we will express the p as a decimal between 0 and 1), and return a single output, the value of x below which $p \times 100$ percent of the data lie.

This may seem challenging at first, but can be accomplished with a few lines of code. To help, here is an algorithm (or list of instructions) that we want to turn into code.

- First, we need to sort **x** from least to greatest. (Note `R` has a command called `sort()` that will do this for us.)
- Second, we need to find out how long the vector **x** is. We can use the `R` function `length()` to do so.
- Third, we need to figure out which row of the sorted **x** corresponds to the $p \times 100$th percentile. All we need to do is multiply the length of **x** by p to get an idea of approximately how many rows down in **x** the desired percentile is. Call this row `target`.
- Fourth, `target` may be a non-integer, but rows are always integers. This is a common problem, with a simple solution: chop off, or "truncate" the unwanted decimal. We can use the function `trunc()`
- Fifth, and finally, have our function return the `trunc(target)`th row of the sorted **x**. That is our percentile.

Test your procedure on the vector $\{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$.

e. Putting it together. Let's check our programming and our statistics together. Draw a vector of 100,000 random numbers from the normal distribution with mean zero and variance one. Now, use `ptile` to find the 67th, 95th, and 99th percentiles of the vector. Do these numbers look familiar? (Think t-statistics...)

Problem 3: R practice and linear regression review, continued

For this problem, you will need the `democracy.csv` file in your working directory.

a. Specify a linear regression of your choosing and run it using `lm()`. For our purposes, this regression need not be (and is in fact very unlikely to be) a theoretically and statistically sound model. For pedagogical purposes, it will help to have a non-binary variable as the response, and at least one non-binary variable among the covariate. Print and explain the summary table given by `lm`.

b. **Counterfactual first differences.** For each covariate, calculate the conditional effect of shifting the value of that covariate from its mean to some (reasonable) higher value, while holding all other covariates at their means. For continuous or ordinal covariates, it may make sense to increase the covariate by one unit or by one standard deviation. For binary covariates, you calculate instead the effect of raising the covariate from 0 to 1. Use the `predict()` function, and provide 95% confidence intervals, printing your results in a nice table.

c. **Counterfactual expectations.** Consider two complete counterfactual scenarios, or hypothetical values of all your covariates, and calculate the conditional expected value of your response variable, including 95% confidence intervals. Compare these to the conditional expectation of your response with all covariates at their means. Explain the substantive import of your findings in clear language.

d. What are some possible sources of bias, inefficiency, or incorrect standard errors in the regression you ran? Note each briefly, but list as many as you can, with particular reference to the problems posed by panel data.

Problem 4: Assessing dynamics of mystery time series, part I

> *NB: A more intricate version of this problem – where the time series potentially also contain seasonality and/or deterministic trends – is available as* Bonus Problem A *below. If you choose to work the bonus problem instead of this one, please indicate this in your answer.*

In the file `mysterytsEssex.csv`, you will find ten columns of time series data. Each column is an independent time series generated by your instructor to have a particular AR(p) or MA(q) structure. For each series, you task is to make your best guess of what process and order the time series was generated from, and whether the series is covariance-stationary. For each series **a.** to **j.**, you should conduct the following tasks:

- Graph the time series against time using `plot()`. There is no need to show this graph in your write up, but you may describe whether it appears mean-reverting or not.
- Produce the correlograms of autocorrelations and partial autocorrelations using `ACF()` and `PACF()`. You do not need to show the graphs, but do write a sentence describing what sort of patterns would be consistent with the sample ACF and PACF you found.

- Make you best guess as to the underlying process: whether AR or MA, what order, your approximate guesses as the values of the ϕ's or ρ's, and whether the process is stationary.

Problem 5: Assessing dynamics of mystery time series, part 2

In this part, we will focus more closely on the problem of stationarity. In the file `mysteryts2Essex.csv`, you will find five additional time series, **k.** to **o.** Once again, each column is an independent time series generated to have a particular structure. All of these series follow some AR(p) process. For each time series, conduct the following tasks:

- Subset the first 20 observations. Plot them against time, and plot the ACF and PACF. (There is no need to show these graphs on your write-up.) Based on these plots, guess the order of the AR(p), and the approximate values of the ϕ's.
- Subset the first 100 observations, and repeat the analysis.
- Conduct your analysis on all 1000 observations. Did any of your conclusions change? If so, what implications does this have for assessing stationarity in comparative politics time series?

Problem 6: Assessing unknown dynamics in comparative data

> *NB: A more involved analysis of single time-series data is available in Bonus Problems B and C. You may choose to work Bonus Problems B and C instead of problem 6 – they will give you practice estimating ARMA models, choosing ARMA models based on goodness of fit, and simulating from those models – but they will take considerably longer to complete.*

For this problem, you should use the TSCS data you have obtained for your own research. If your data have small T (e.g., less than 15 or so), use the `democracy` dataset instead.

a. Choose a continuous variable from the dataset, and choose three different cross-sectional units. Create in new variables the corresponding time series for each of your units, so that you end up with three time series variables without any panel structure.

b. For each of your three time series, construct (and show in your write-up) a plot of the time series against time and correlograms of the ACF and PACF. Assume that the series given is either AR(p) or MA(q), and guess the data generating process, order, and indicate whether you believe it to be covariance-stationary.

c. Difference each of your three time series, and place the results in three new vectors. *Hint:* For non-panel data, the differenced version of `x` is given by

```
diffx <- diff(x,1)
```

Now repeat the analysis of part b. on the *differenced* series. What do you find?

Problem 7: Assessing potentially non-stationary dynamics in comparative data

For this problem, you should use the TSCS data you have obtained for your own research. If your data have small T (e.g., less than 15 or so), use the `democracy` dataset instead.

a. Choose a continuous variable y_t from the dataset, and choose one cross-sectional unit that you suspect may have a unit root. Create in a new variable the corresponding time series, so that you end up with a single time series variable without any panel structure. Also select any covariates x_t you think may influence y_t, and select their values for the cross-sectional unit of interest. Make sure to include at least one continuous x_t.

b. Construct (and show in your write-up) a plot of the time series y_t against time and correlograms of the ACF and PACF. Assume that the series given is either AR(p) or MA(q), and guess the data generating process, order, and indicate whether you believe it to be covariance-stationary. (This may be a repetition of your last exercise.)

c. Perform an augmented Dickey-Fuller unit root test on y_t and report and explain your results, including any limitations of the test.

d. Estimate a model of your time series using an appropriate ARIMA(p,d,q) model, including any covariates you think are appropriate.

e. Using the simulation techniques of your choice, show would happen to y_t if starting at some point in the middle of the time series, one of your covariates took on values $x_{t,\text{new}}$ other than the historically observed levels (i.e., calculate $\text{E}(y|x_{t,\text{new}})$. Include confidence intervals on your plot or table.

f. Contrast the short and long run predictions from part e. Which do you trust more?

g. Repeat part e. using the actual, observed covariates x_t. Does the model predict values of $\text{E}(y_t|x_t)$ that comport reasonably with the actual values?

h. Examine the ACF and PACF plots for an continuous covariates in your model. Note any x_t's that may have a unit root, and perform ADF tests on those x_t's.

i. Estimate an error correction model of y_t on x_t. Be sure to report the estimate of the cointegrating vector, and the second stage least squares estimates of the ECM itself. Interpret the coefficient values you have obtained.

j. Contrast the results of your ECM and ARIMA models of y_t. How are they the same? How are they different?

Problem 8: Analyzing a Panel Data Set

> *NB: A more involved analysis of panel data is available in Bonus Problem D. Because problem D is much longer than Problem 8, I recommend saving it for after the course.*

For this problem, you should use the TSCS data you have obtained for your own research. You may use any statistical package you wish, *so long as you can perform the appropriate analyses from the course*. The examination for this course will ask essentially the same questions of a provided data set with large T, and you will be allowed to adapt your code from this assignment to that dataset.

a. Choose a continuous variable y_{it} from the dataset, and specify a model of this variable in terms of some covariates $\mathbf{x}_{it}$. Throughout this assignment, you may find it necessary to go back and choose a new specification of x_{it} with a better fit. You need only write up your final specification, but you should mention any specifications you tried and discarded, with a justification for your choice.

b. If your panel data have sufficient time periods, check the time series properties of y_{it} using all the relevant methods we have learned (time series plots, ACFs, PACFs, unit root tests, AIC tests on estimated ARIMA models), and specify an ARIMA(p,d,q) specification for y_{it}. For modeling purposes, you may assume that all units share the same ARIMA(p,d,q) process, but note if this assumption seems faulty based on your data exploration. If your data are very short in time periods, you may skip this step, though it would still be good to do some exploratory data analysis.

c. Choose and justify either a random effects, fixed effects, or mixed effects intercept for the model. If you choose fixed effects, select an appropriate estimator based on the number of units and periods. Indicate whether you think one- or two-way effects are appropriate.

d. Estimate your panel model with unit heterogeneity, and present the parameter estimates, standard errors, and fit statistics nicely in a table, as you would for a paper. If T is sufficiently large, your model might be a panel ARIMA(p,d,q) model or a lagged dependent variable model (be sure to check for serial correlation and consider alternatives). If T is small, your model might be a panel GMM, in which case you should assess the sensitivity of results to model assumptions.

e. For each covariate in x_{it}, interpret the short-run and long-run conditional expectations of y_{it} given hypothetical values of that covariate, with all other covariates held constant. You may calculate either an "expected value" like $\mathrm{E}(y_{\mathrm{hyp},t}|\mathrm{x}_{\mathrm{hyp},t})$ for some hypothetical $\mathrm{x}_{\mathrm{hyp},t}$ or a "first difference" $\mathrm{E}(y_{\mathrm{post},t} - y_{\mathrm{pre},t}|\mathrm{x}_{\mathrm{pre},t}, \mathrm{x}_{\mathrm{post},t})$ for two different hypothetical values, $\mathrm{x}_{\mathrm{pre},t}$ and $\mathrm{x}_{\mathrm{post},t}$. Present these results nicely in graphics, tables, or sentences, as you would for a paper. Explain whether and why you trust or distrust the long run estimates for your model.

Bonus Problem A: Identifying unknown stationary time series processes

In the file `mysterytsUW.csv`, you will find 18 columns of time series data. Each column is an independent time series generated by your instructor to have a particular structure. That structure might include a deterministic time trend, seasonal effects, AR(p) processes, and/or MA(q) process. All of these time series can be assumed to be covariance stationary.

For each series, your task is to make your best guess of the data generating process (DGP) which produced the data. Thus, for each time series **a.** to **r.**, you should indicate whether you suspect the time series DGP includes any of the following four components:

(*i.*) deterministic trend If you suspect a deterministic trend, indicate your evidence for that trend, describe it (e.g., with an estimate of the monthly increase or decrease) and then remove it from the time series to yield a detrended time series for further analysis.

(*ii.*) seasonality Assume the data are monthly, so that any seasonality should show up on a 12 observation cycle. In this case, assume that any seasonality present is additive.[1] If you suspect seasonality, describe the seasonal cycle, then remove the seasonal means from your data to yield a seasonally-adjusted time series.

(*iii.*) autoregression If you suspect autoregression is present, describe the order of autoregression and the likely signs and magnitudes of terms. Be sure to use detrended and/or seasonally-adjusted data if you found either a time trend or seasonality.

(*iv.*) moving averages If you suspect moving average components are present in the error term, describe the order of the moving average process and the likely signs and magnitudes of terms. Be sure to use detrended and/or seasonally-adjusted data if you found either a time trend or seasonality.

The first 12 time series – **a.** to **l.** – contain at most one of the four components above. The remaining six time series – **m.** to **r.** – may contain more than one component.

Use any tools you know, including graphs of the time series, correlograms of autocorrelations (ACFs), and correlograms of partial autocorrelations (PACFs). It may be useful to apply these tools to the original, detrended, and/or seasonally ajusted time series. It is not necessary to show every graph you make; often a sentence summarize a plot will be sufficient, but if you are in doubt, show and describe the plot.

You will be marked on the basis of your choice and use of appropriate diagnostic tools. You will not be penalized for failure to guess time series processes correctly, unless this reveals deficiencies in your understanding of diagnostic tools.

Bonus Problem B: Analyzing US House seat shares using ARMA

Since 1963, the US House of Representatives has consisted of 435 elected voting members serving two-year terms. Every seat in the House is up for election in November of even-numbered years to seat the Congress that will serve in the following two years. Thus, the 2016 election determined the 435 members of the House for the 115th Congress, serving from 2017–2018.

1 Multiplicative seasonality is more common but is not used in this example.

Variable	Description
`Congress`	session of Congress (effectively a time index)
`StartYear`	the first year of each two-year session
`DemSenateSeats`	the number of Democrats (and independents caucusing with the Democratic Party) elected to the Senate in this session of Congress
`DemSenateMaj`	the size of the Democratic Senate Majority, or `DemSenateSeats` minus 50
`DemHouseSeats`	the number of Democrats (and independents caucusing with the Democratic Party) elected to the House in this session of Congress
`DemHouseMaj`	the size of the Democratic Senate Majority, or `DemHouseSeats` minus 217
`Midterm`	whether this session was elected in a midterm election (1) or a presidential election (0)
`DemPresident`	whether the president on the last election day was a Democrat (1) or a Republican (0)
`Unemployment`	the monthly unemployment rate at the time of the election of this session of Congress
`UnemDeviation`	the difference between pre-election unemployment and mean unemployment, 1963–2016 (which was 6.075%)
`Coattails`	1 if the presidency shifted to the Democrats on election day, −1 if the presidency shifted to the Republicans, and 0 if the party of the president was unchanged
`PartisanMidterm`	1 for midterms in which the Democrats hold the presidency, −1 for midterms in which there is a Republican president, and 0 in presidential elections
`PartisanUnem`	equal to `UnemDeviation` when a Democrat is president, and to −1× `UnemDeviation` when a Republican is president
`Pre1994`	1 if this Congress was elected before 1994, 0 otherwise

Table 1. Codebook for Congressional Seats data. Data are in `congress.csv`, and are taken from the Bureau of Labor Statistics (unemployment) and Wikipedia (all other raw variables), or constructed from these data by your instructor.

We will study the evolution of the time series of the number of seats won by Democrats (or by independents who caucus with the Democratic Party) in each election held from 1963 to 2016 (a total of 28 observations). As substantive interest focuses on the party in control and the size of their majority, we will focus our analysis on these outcomes, where positive values indicate the size of a Democratic majority and negative values the size of a Republican majority.

We will also consider three possible influences on the size of the Democratic majority. First, when one party sweeps the other party out of the presidency, they tend to bring in a wave of co-partisans to Congress who are "clinging to the president's coattails." Second, the party of the president tends to do badly in midterm elections (those that do not involve a presidential election; e.g., 2010, 2014, and 2018). The usual explanation is that voters frustrated with the president cannot replace him, but can only offset his power with opposition in Congress.[2] Third, in all elections, voters tend to attribute economic performance to the party of the president. For example, keeping unemployment below its long term average should help the president's party at the expense of the opposition, and *vice versa* when unemployment is higher than usual.

These three explanations leave a lot out: for example, changes in the use of redistricting for partisan advantage, in the partisan composition of the electorate, and especially the transition of the Southern Democrats to the Republican Party. Because incumbency provides a strong advantage to sitting members of Congress, arguably many of these changes did not act gradually, but with a "bang" when a sudden shock caused many incumbents to lose or retire. The shock in question is the 1994 midterm election; to account for the possibility it reflects a "structural break" in the level of the time series, we will also consider a control for whether our observations come before or after this watershed.

In the file `congress.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:

a. Plot the time series `DemHouseMaj` and plot its ACF and PACF. Perform augmented Dickey-Fuller and Phillips-Peron tests for unit roots. Describe your findings, being sure to describe what kind(s) of time series process may be at work. Now "demean" the data by period, removing the pre-1994 mean from cases before 1994, and the post-1994 mean from cases after 1994. Make new time series, ACF, and PACF plots. If 1994 represents a "structural break" in the level of the Democratic majority, what effect does that have on your diagnosis of the behavior of the time series?

b. Fit an AR(0) regression to the time series `DemHouseMaj` controlling for the covariates `PartisanMidterm`, `PartisanUnem`, and `Coattails`, which test the three theories mentioned above. Also control for `Pre1994` to allow for a structural break. Present the results in a table, being sure to note the coefficients, their standard errors, the AIC for the entire model, the standard error of the regression, and the number of observations. Format the table nicely, as if for a paper, and describe what you have found substantively as well as you can.

2 More subtle theories exist; see Alberto Alesina and Howard Rosenthal, 1995, *Partisan Politics, Divided Government, and the Economy*, Cambridge University Press.

c. Now fit the following additional models and add them to the table you made in part **b**: (*i.*) an AR(1) model; (*ii.*) an AR(2) model; (*iii.*) an MA(1) model; (*iv.*) an ARMA(1,1) model. Make sure to include the same four controls as in part **b.** Discuss the substantive and statistical similarities and differences across all five fitted models.

d. Perform a rolling-windows cross-validation of all five models using a window of 20 periods and forecasting forward 3 periods. Place in a table the following goodness of fit statistics for all five models: AIC, in-sample root mean squared error, and the cross-validation mean absolute error (MAE) up to 1, 2, and 3 periods ahead, respectively, as well as the average of these three cross-validation MAEs. Based on these statistics, select a final "best" model.

e. Using the model you selected in part **d.**, forecast what will happen to the size of the Democratic majority in the US House in the 2018, 2020, and 2022 elections for three scenarios. For all three scenarios, assume the Democrats recapture the presidency in 2020 and compute appropriate counterfactual values of `PartisanMidterm` and `Coattails`, and set `Pre1994` to 0. For unemployment, assume the following:

Scenario	Counterfactual
1	unemployment stays at 4.6% for all three elections
2	unemployment falls to 3.6% for all three elections
3	unemployment rises to 5.6% for all three elections

For each scenario, report or graph the predicted Democratic majority and its 95% confidence (or predictive) interval for the 2018, 2020, and 2022 elections. 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.

NB: As a check on your work, for each scenario and year also 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, you may use either `predict()` or the `simcf` library's `ldvsimev()`.

Bonus Problem C: Analyzing US Senate seat shares using ARMA

Since 1963, the US Senate has consisted of 100 elected voting members serving staggered six-year terms. Roughly one-third of the seats in the Senate are up for election in each even-numbered year. As a result, the Senate has three "classes" of seats. For example, the class of 2012 is up for re-election in 2018. Because 2012 was a good year for Democrats (due to Barack Obama's coattails, among other factors), that means the Democrats have many seats to defend in 2018, putting them at a disadvantage relative to 2016.

a. Plot the time series `DemSenateMaj` and plot its ACF and PACF. Perform augmented Dickey-Fuller and Phillips-Peron tests for unit roots. Now "demean" the data by period, removing the pre-1994 mean from cases before 1994, and the post-1994 mean from

cases after 1994. Make new time series, ACF, and PACF plots and compare your results. Diagnosis the time series, accounting for the possibility of a structural break.

b. Repurpose your code from Problem 1 to model the time series `DemSenateMaj`. In particular, control for `PartisanMidterm`, `PartisanUnem`, `Coattails`, and `Pre1994`, and consider five models: an AR(0) model, an AR(1) model, an AR(2) model, an MA(1) model, and an ARMA(1,1) model. Recreate the two tables you made in Problem 1 (the table of coefficients and the table of goodness of fit statistics) for the Senate data. How do the substantive results compare to the House models?

c. Now estimate a sixth model: an AR(1)AR(1)$_3$. Add this model to the two tables you made in part **b.** How well does the new model do? What model is best overall? Can you provide a substantive rationale for using either an MA(1) or an AR(1)AR(1)$_3$ model to model the US Senate, but not the House?

Bonus Problem D: Analyzing partisan seat shares in US state legislatures

In the last two bonus problems, 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

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 2. 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.

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.

Throughout, 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.[3] In the second specification, called M2, we control for all these variables and region specific trends.[4]

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.

a. 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.[5]

b. 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.

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

4 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?).

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

c. 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.[6] 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. 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. 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.[7] 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. 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. 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).

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.

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

7 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.

h. 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...