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

Problem Set 1

Professor: Christopher Adolph, Political Science and CSSS

Spring Quarter 2020

Due by email to `inhwanko@uw.edu` on Tuesday 21 April 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 grouped 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 on the first page of your assignment.

Problem 0: R refresher (optional)

[0 points.] If you are new to R, I recommend you work through the R practice exercises below. This optional problem carries no credit but will be corrected if submitted.

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 "."

Variable	Description
`COUNTRY`	numerical code for each country
`CTYNAME`	name of each country
`REGION`	name of region containing country
`YEAR`	year of observation
`GDPW`	GDP per capita in real international prices
`EDT`	average years of education
`ELF60`	ethnolinguistic fractionalization
`MOSLEM`	percentage of Muslims in country
`CATH`	percentage of Catholics 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)
`CIVLIB`	degree of civil liberties (1–7 scale, rising in civil liberties)
`REG`	presence of democracy (0=non-democracy, 1=democracy)

Table 1. Codebook for Problem 0. Data are in `democracy.csv`, and are taken from data from Przeworksi, *et al*, *Democracy and Development: Political Insitutions and Well-being in the World, 1950–1990.*

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 you are writing more than a line or two of code for each unstarred problem, look for a simpler way. And though it may not be obvious at first how, the starred problems should be answered with no more than two to six lines of code, rather than laborious looking up of data.

It is essential you show your work and all code.

a. [0 pts.] Load the Democracy dataset into memory as a dataframe.

b. [0 pts.] Attach it so that each variable in the dataset is accessible by name.

c. [0 pts.] ⋆Check whether `CTYNAME` has been read as a character variable (that is, as unique names for each case), or as factor variable (that is, a categorical variable that takes on a named value out of a menu of options). If it is a factor, convert it to character. (This is necessary for parts m., o., and q. to go smoothly.)

d. [0 pts.] Report summary statistics (means and standard deviations, at least) for all variables.

e. [0 pts.] eport a correlation matrix of all the variables in the dataset. Watch out for missing values.

f. [0 pts.] Create a histograms for political liberties.

g. [0 pts.] Create a histogram for GDP per capita.

h. [0 pts.] Create a scatterplot of political liberties against GDP per captia.

i. [0 pts.] Create a boxplot of GDP per capita for oil producing and non-oil producing nations.

j. [0 pts.] 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?

k. [0 pts.] 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?

l. [0 pts.] What was the median average years of education in 1985 for all countries?

m. [0 pts.] ⋆Which country was (or countries were) closest to the median years of education in 1985 among all countries?

n. [0 pts.] What was the median average years of education in 1985 for democracies?

o. [0 pts.] ⋆Which democracy was (or democracies were) closest to the median years of education in 1985 among all democracies?

p. [0 pts.] What were the 25th and 75th percentiles of ethnolinguistic fractionalization for new countries?

q. [0 pts.] Which country-years were nearest to the 75th percentile of ethnolinguistic fractionalization for new countries?

Problem 1: Identifying unknown stationary time series processes

[54 points.] 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

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

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.

Problem 2: Identifying unknown, possibly nonstationary time series processes

[15 points.] In this problem, we focus on the problem of identifying nonstationarity. In the file `mysterytsUW2.csv`, you will find five additional time series, **s.** to **w.** Once again, each column is an independent time series generated to have a particular structure. All of these series follow some AR(p) process. No moving average processes, deterministic trends, or seasonal variation are present in these time series. However, these five time series are not guaranteed to be covariance stationary. For each time series:

(*i.*) Subset the first 20 observations. Plot them against time and plot the ACFs and PACFs. (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 autoregressive parameters.

(*ii.*) Subset the first 100 observations and repeat the analysis. Did any of your conclusions change or become more or less certain?

(*iii.*) Conduct your analysis on all 1000 observations. Did any of your conclusions change? If so, what implications does this have for assessing stationarity in time series in your field?

Problem 3: Student project checkpoint

If you are working with other students in the class, please identify them; you should jointly write this section of the homework. Please submit this problem separately *from the problems above: one copy per group! Email your group's response to this problem to* `cadolph@uw.edu`.

[31 points.] Provide a brief (2–3 paragraph) summary of your proposed research design. The key questions to answer:

- What is the outcome studied in your analysis? What range of values can it take on?

- What is the unit of analysis? In particular, what is the number of cross-sectional units *N* and the number of periods *T*? Is there anything unusual about the selection or observation of units or time periods?

- Is there any pattern of missing data and/or selection of data?

- In what ways do you think different observations over time may be related (dependent) in your data?

- Do you suspect any Gauss-Markov assumptions would violated if these data were analyzed with a simple linear regression?

- Based on what you have learned so far, what methods might be appropriate for analyzing your data? If you have the data in hand, you should provide histograms and/or density plots of the dependent variable, as well as any other relevant diagnostics to support your answers.