The scenario for this project is as follows: you are interested in forecasting a particular variable. Call this your target variable. That is you want to use past values of the series to make predictions for the future of this target variable. In order to do that you need to model the process generating this time series. The model for this time series must be validated before you use it for forecasting. So you collect data on the time series of length n. Then use n-12 observations to fit the model and keep 12 to validate the forecasting performance of the model.

The project answers the following research questions for a specific area of application: (a) Do multivariate atheoretical (vector autoregressive) models do a better job at forecasting _____(your target variable)______than univariate models ? Do causal univariate models do a better job at forecasting than models that depend only on past values of the variables?

For the first installment of the project you must search for:

1) More than one time series. Three to four variables needed, where one of these variables is the one you want to forecast, and the others are believed to be related to the target variable. The variables must be observed at the same frequency (e.g., they all must be monthly or quarterly or aggregated to monthly or quarterly if the frequency is higher than 12), they must be observed during the same period of time. For example, variables x, y, z, observed between 1900 and 2000 monthly. So the same number of observations for each variable is needed during that same period.

2) Research question needed. How do you think x affects y and z, etc.

3) Large sample size, more than 300 monthly observations in each series.

4) The time series can not be seasonally adjusted and not detrended,.

5) Reference, web site, where you get the data. This web site must also say explicitly that the data have not been processed (neither seasonally adjusted nor detrended). If your data can not be referenced, i.e., if I can not go and see it in its source location and double check that, then you can not use it.

6) No missing observations. So if for example, January 1901 is missing, the data is not acceptable. There can not be missing data in any of the variables.

7) To turn in: Wednesday, April 19th, before 11 pm, upload three files: (a) a no longer than 5 pages typed document describing the data. Use material from chapter 1 of the textbook. Use the format indicated in the next page. The file uploaded must be a pdf file. I will not accept under any circumstances files emailed to me. So if you miss the deadline for whatever reason and you do not upload much earlier than 11 PM on Wednesday the 19th, you will get 0 for this part of the project. Your file should have name :LASTNAME-ID.pdf .Name and ID must be on the top right hand side of the file.

(b) A .R script file containing your name and ID on the top and the R code used to analyze your data. This filename will be LASTNAME-ID.R

(c) The whole data set in .csv format. This file name will be LASTNAME-ID.csv

Double check code and data to make sure it all works.

8) No two students can use the same data sets.

9) The rancdom component of your target variable must have some correlation structure that is stationary in it. If you get white noise you can not use these data.

10) Below is an example of format of your document:

Front page (not counted in the 5 required pages).

Name: J…. ID ----

Course

Title for your project. For example: Forecasting number of miles run monthly by Prof. Sanchez.

Next pages (example)

I. Introduction.

Introduce your research question. For example: In this paper, we forecast the number of hours run monthly by professor Sanchez. I hypothesize that there is a negative relationship between the number of hours teaching per month, and the number of miles run monthly by professor Sanchez. However, I expect that the more calories per month eaten, the more she runs. Having more money also increases the number of hours running per month, as she likes to wear nice running clothes when she runs and that can only happen if she can afford the clothes. Thus probably she runs more in the summer, when she makes extra money doing summer teaching. A seasonal effect is expected in the summer months. As she gets older, it is anticipated that she runs less per month, thus a decreasing trend is also expected.

In order to model running and its relation with the other variables, I collected data from (website address). In this (address web site), we can see that it says that the data are not seasonally adjusted. The following section describes the variables selected to answer the research question.

(So I have four variables: miles run per month, number of hours teaching per month, number of calories per month, and monthly income.)

II. The data

Here, I describe the variables used for the analysis. I give a short name to each of the variables used. The variable **running** measures the number of miles per month. It was obtained from the archive of UCLA teachers that run….. (source). The length of the series is n=600. The variable **calories** measures the number of calories consumed per month. Etc…

If you need to aggregate your data because the frequency is larger than 12, explain what you do to aggregate.

The following table contains a summary that makes it easier to see the variable description and short name.

Table 1

Write here a table with variable short name (column 1), variable description, ie what it is, and units (column 2) and length (column 3).

III. Description of the data (chapters 1 and 2 methods must be used).

Plot 1 contains four separate time plots, one for each time series. (using the same scale for the horizontal axis in all the plots). Describe the plots, what they say about the data. Do they provide indication that there is need to transform the data to make the series variance stationary? (log, square root)? That is necessary before you look at anything else.

PLOT 1 HERE

Example. We can see that there is an upward trend in the running data . Then there is a seasonal. I used window to determine the month and observe the time data more closely. The seasonal in running increases in amplitude overtime, suggesting that multiplicative decomposition would be more appropriate. The seasonal in running is notvregular, hard to pinpoint its nature.

Plot 2 shows the additive decomposition of then target variable, plot the the additive decomposition of the running series. You must describe what you see in those decomposition plots. Plot 3 shows the seasonal plots of the target variable. Describe what they say. . Explain what you see in those plots.

PLOTS 2, 3 HERE

Then plots 5, 6, contain de multiplicative decomposition of the same variable. And then seasonal plot. Explain what you see in those plots.

PLOTS 5,6 HERE

III.1 Autocorrelation structure of the random component of the running data.

Once you decide whether multiplicative or additive decomposition is more appropriate, you can proceed to look at the ACF of the random component. If it has stationary correlation structure, you can describe what this structure is (which r's are significant, etc.). IF it does not, investigate whether you did the right decomposition, the right pre-transformation to stabilize the variance before you decomposed, etc. If the random error term is noise, you have to start all over.

Plot 7, 8 here.

References section:

[1]   where the data come from.

[2] Textbook reference.