Project Guideine

Ozan Sonmez

11/27/2017

The purpose of this document is to provide a guideline for the time series project that is due on Dec 1st, Friday. I will try to give a template of how your project should look like and give a step by step guideline on how to present your project. **Please note that you are allowed to use other formats to display your findings, this is just a template!**

How are you going to be graded

The grade of the project will be mainly distributed on the flow of your analysis and how you communicate your statistical findings to a non technical person. Here are some general rules to consider before you turn in your project:

- The project should be written using one of the followings: Latex, R markdown, Word. We are not going to accept hand written projects!
- The format of your project should include the following sections:
  - Title of your project, name and student ID
  - Introduction
  - Data Description
  - Data Analysis
  - Discussion
  - Conclusion
  - Appendix: Your R Code (well documented)
- Be organized about how you display your findings
- DO NOT PROVIDE EVERY R OUTPUT! You should only output the relevant R output that you think is important to your project. If you are displaying tables, please put them in the table format. If you are displaying figures, make sure that the figure is visible and they have uniform size throughout your project.
- Please write clear and understandable! An important job for statistician is to be able to communicate the findings, a major part of how you will be graded is based on how you communicate these findings, so be clear!
- If something seems wrong, talk about why! It is ok to have something that doesn't look right even after you did everything correctly. Knowing what might be causing these issues is also crucial and as valuable as a correct analysis.
- Please label your figures and tables, and refer them in your writing
- Finally, HAVE SOME FUN! This is just a project!

Introduction

- A little background about why is it interesting to you to look at that data (be creative!), what potential questions (non statistical) one might wonder about it.
- What is the proposed solution (in a non technical way)?
- How is the rest of the write-up organized?

Data description

This is the part where you discuss:

- What kind of data you are using
- Where did you get the data from?
- Characteristics of your data, including
  - Sample size
  - The time range of your data (ex: range of years the data is collected)
  - How data is recorded (ex: daily, monthly , weekly, yearly)
- Did you do any pre-calculations on your data? For example if your data is daily, did you convert that into monthly data by simply taking the average and etc,...
- Provide a plot of your data
- Provide summary statistics
- What are the interesting features of the data? (any trend, seasonality, ...)

Data Analysis

This is the part where you conduct your statistical analysis and display your results:

a) exploratory analysis

- Do you need to transform you data in order to stabilize the variance and etc ...
- remove the trend (discuss how you estimate the trend component)
- remove seasonality (discuss how you estimate the seasonal component)
- Get the rough part of your time series. If you have a time series $(X_t, t \in \mathbb{N})$ with a decomposition $X_t = m_t + s_t + Y_t$, then you should estimate the smooth components $\hat{m}_t$, $\hat{s}_t$, then get the random component by $\hat{Y}_t = X_t - \hat{m}_t - \hat{s}_t$
- Explore $\hat{Y}_t$: Is it stationary? Is there any dependence?

b) picking a model

One you are convinced that $\hat{Y}_t$ is stationary, next step is to choose models:

- Look at acf and pacf of $\hat{Y}_t$, and propose an ARMA model
- You can also use the function auto.arima(), but this might pick a model that is more general than ARMA (something that is not covered in the class, which is ok!)
- You can use a sequence of p and q models and do a for loop in R to fit ARMA(p, q) for varying p and q and pick the model with the smallest AIC or BIC.
- if the above three models vary among themselves, you can pursue them simultaneously, and the pick the best one at the end based on their predictive ability.
- Need to analyze the residuals after the model fit to make sure that they ensemble white noise:
  - Look at the acf and the pacf of your residuals
  - You can perform more statistical test to see if there is no auto-correlations in the residuals of your model fit by Ljung–Box test, The Portmanteau test or The rank test
  - Can also check if the residuals are Gaussian (Normal) white noise.

c) prediction

Once you have picked your model in b, you can do predictions using forecast() function in R defining the lag of the prediction.

Discussion

Explain the findings of your data analysis. Try to be critical of what you have done and make sure you don't hush over potential shortcomings of your analysis. It is better to be aware of something that is not ideal than to try covering it up. (Importantly, this is not being viewed negatively in the evaluation of your time series analysis. It can easily happen that you face an issue you cannot deal with based on the methods you learned in STA 137. When this happens try to use your best judgement to proceed in a reasonable way.)

Conclusion

This is the part where you tell your story to a non-statistician, for example to your product manager if you work in a advertisement company. You should avoid using technical terms here and convert your statistical findings into business solutions.

Appendix: R Code

Please put your relevant R code in the appendix. Make sure you document these by commenting on the functions you are using (ex: what are you doing with that specific line of code). You can also put other relevant results that you dint use in your main body of your project. (Do not put every R output, only the relevant part!)

Code of Conduct

Do not copy from the web and existing papers. You can and should use other resources, but they have to be clearly identified. Line-by-line copying from existing contributions will be considered plagiarism.