Predicting Gender Wage Ratios Case Study Rubric

DS 4002 – Fall 2024 – Kristy Luk
Due: TBD
Submission format: Upload link of GitHub repository and pdf of the written portion, via Canvas

Individual Assignment

General Description: Submit a link on Canvas of your Github repository and written portion after completing this case study.

Why am I doing this? This case study allows you to hone in on your data science knowledge and writing skills and apply them to real-world problems. As you work through this assignment using time-series data to forecast gender wage ratios, you will delve deeper into the ways that data analysis tools can be used in practical senses, with potential implications for public policy.

What am I going to do? The GitHub repository for this case study can be found here: https://github.com/kristyluk/DS4002CS3. You will obtain data on historical wage ratios, as well as additional socioeconomic variables to minimize the potential of omitted variable bias. A cleaned dataset is included in the repository for ease. You will then use the data from just 1960-2000 to predict the wage ratios for 2001-2019 using the ARIMA forecasting model with Python based on these variables. The goal is to see if the predicted values are the same as the actual values, which can be seen in the full.data.csv file. After, you will test the accuracy and predictive power of the model by calculating the residuals and root mean squared error values.

Final deliverables include:

- A link of your GitHub repository, including all the data used, scripts, output, and additional materials/references
- A short written document in pdf format of your results, conclusions, and discussion

Tips for success:

- Read the reference materials and code comments. This will help you better understand each step as you work through this case study, especially when you work on the ARIMA portion of the code.
- Change code if needed. Don't feel like you are restricted to use the same exact code you see in the script. If you feel there are better techniques or parameters to use for ARIMA, please do so! Use the script as guidance.

How will I know I have Succeeded? You will meet expectations on this case study assignment when you follow the criteria in the rubric below.

Spec Category	Spec Details
Formatting	● One GitHub repository, submitted via link on Canvas ● Your repository should include the following: ○ A README.md file, including references ○ A DATA folder ○ A SCRIPTS folder ○ AN OUTPUT folder ● An about one page pdf document of your results, conclusions, and discussion
README.md	● Goal: This file serves as an orientation to everyone who comes to your repository, it should enable them to get their bearings ● Does not need to be super detailed, but should include enough information for someone with not much familiarity to read ○ Include brief background information, summary, research goal ● Use markdown headers to organize content ● Include references, if applicable, at the end of the README.md file ○ Use IEEE Documentation style
DATA folder	● Goal: This folder contains all of the data for this project ● You should AT LEAST include the initial data, and the final data analyzed after any cleaning performed ● If needed, the code in the SCRIPTS folder should be able to get you from the initial piece of data to the final one ● If your data fits in GitHub, place all of it here
SCRIPTS folder	● Goal: This folder contains all the source code for your project ● All script files should include header comments at the beginning of a script for organization purposes ● Include comments of what each code or sequence of code accomplishes and the purpose throughout your script
OUTPUT folder	● Goal: This folder contains all of the output generated by your project ● In a pdf document, include ALL plots, tables, figures, etc. generated ● Organize the file by putting them in order of how they appear in the script, labeling axes, and including titles
Written Portion	● Goal: Use your writing skills to display key findings in concise format ● About one page in pdf format

	● Discuss the key findings of the study ○ Be concise ○ Include only the most important images from the OUTPUT folder here and discuss the results from them ○ Discuss the implications of your findings ○ Discuss challenges, future improvements, and next steps

Acknowledgements: Thank you Professor Alonzi for providing guidance on creating this rubric!