CSE 544 Project, Spring 2023

Due **May 12th 11:59pm** via Brightspace (submission instructions similar to assignments)

1 Submission

Submit your project solution on Brightspace by May 12th 11:59pm (no extensions). The submission instructions are similar to assignment submissions: make sure all groups members are listed in the submission page 1, zip/tar all files into one archive and upload, include all plots, code, explanations, observations, etc., as asked in each task. For example, if you are asked to find MME or MLE of a distribution's parameters, then please show your work as to how those are derived. If these were already derived in class or in an assignment, you can simply refer to those without redoing the work, but do mention the reference in your submission. Please ensure all plots are legible, as in assignments. If you have any doubts, please post on piazza in a timely manner. If you are asked to make assumptions or considerations or if you use outside sources, please say so clearly and cite references as needed.

The Consumer Price Index (CPI) is a widely used measure of inflation in the United States. It measures the average change over time in the prices paid by urban consumers for a basket of goods and services, such as food, housing, transportation, and medical care. The CPI is published monthly by the Bureau of Labor Statistics (BLS), which is a part of the U.S. Department of Labor.

The BLS collects data for the CPI through a variety of methods, including surveys of households and businesses, as well as data from government agencies and other sources. The data is then used to construct price indexes for different categories of goods and services, which are weighted together to produce the overall CPI.

The CPI dataset contains time-series data for different categories of goods and services, such as housing, energy, and transportation. This data is used by policymakers, economists, and businesses to monitor inflation and make decisions related to monetary policy, wage adjustments, and investment strategies.

There are 28 datasets based on 14 US urban centers. Each urban center gives rise to two monthly time-series datasets, one from 1970-1996 (27 years) and the other from 1997-2022 (26 years). Each panel dataset should have enough datapoints for you to perform robust statistical analysis. Each group must work with one time-series dataset. The google drive download link will be shared in google sheet with all the groups in piazza. Please write the name of the group members, in front of whichever dataset your group downloaded. Also write the name of one group nominee who will be my point of contact in case any dataset disputes arise. For grading purposes all the datasets

are same and your grades will be based on your implementation and answers and not depend on the dataset you choose.

2 Dataset

The Consumer Price Index (CPI) is a measure of the average change over time in the prices paid by urban consumers for a basket of goods and services. In this assignment, we will be working with a dataset containing the CPI, Housing, Rent, Energy, and Purchasing power of the Dollar for all urban centers for a period of 25-30 years. The goal of this assignment is to use the concepts learned in the class on the given dataset to gain insights into the relationship between these variables. For a better understanding of the variables here is brief description of the columns:

1. Consumer Price Index for All Urban Consumers - All Items: This column measures the average change over time in the prices paid by urban consumers for a basket of goods and services. It includes all items in the US city average, such as food, clothing, housing, transportation, and medical care. (Column name : CPI_all_items)

2. Consumer Price Index for All Urban Consumers - Rent of Primary Residence: This column measures the average change over time in the prices paid by urban consumers for rent of their primary residence. It is one of the components of the CPI for housing. (Column name : Rent_of_Primary_residence)

3. Consumer Price Index for All Urban Consumers - Housing in the city: This column measures the average change over time in the prices paid by urban consumers for housing-related expenses, including rent, utilities, furniture, and other household items. (Column name : Monthly_Housing_Cost)

4. Consumer Price Index for All Urban Consumers - Energy in the city: This column measures the average change over time in the prices paid by urban consumers for energy-related expenses, including electricity, gas, and fuel oil. (Column name : CPU_Energy)

5. Consumer Price Index for All Urban Consumers - Purchasing Power of the Consumer Dollar: This column measures the value of a dollar in terms of what it can buy. It reflects the average change over time in the prices paid by urban consumers for goods and services relative to the value of a dollar. A decrease in the purchasing power of the dollar means that consumers can buy fewer goods and services with the same amount of money. (Column name : US_Dollar_Purchasing_power)

3 Project ToDos

3.1 Outlier Detection

The first task is to perform outlier detection for all 5 columns in your dataset. Use Tukey's rule from class with $\alpha = 1.5$ for this task. The rule should be applied across all values (and all years) for each column separately. Note the number of outliers detected for each column in your submission. Then, delete the values of these outliers in your dataset. We will replace them with more reasonable values in the next task.

3.2 Perform linear interpolation of missing data-points

U.S. statistics of labor bureau keep changing the frequency at which they poll the urban centers for CPI statistics, as such you might encounter some columns which have missing data. Also, because of the outlier removal from the previous task, you may have some blank entries. For this task you, will use linear interpolation to fill the missing data points.

Linear interpolation is a method of estimating values that are missing in a dataset. It involves using the known data points above and below the missing value to estimate what the value would be if it were present. The process of linear interpolation involves the following steps:

1. Identify the missing data points in the dataset.
2. Find the known data points that are immediately above and below the missing data point.
3. Calculate the slope of the line that connects the two known data points. This can be done using the formula:

$$slope = (y_2 - y_1)/(x_2 - x_1) \tag{1}$$

   where x_1 and y_1 are the coordinates of the first known data point, and x_2 and y_2 are the coordinates of the second known data point. Note that in your case y_1 and y_2 are the column values above and below the range of missing data and $x_2 - x_1$ gives the number of days of missing data + 1
4. Use the slope to estimate the value of the missing data point. This can be done using the formula:

$$missing\_value = y_1 + slope * (missing\_value\_x - x_1) \tag{2}$$

5. Replace the missing data point with the estimated value.

To fill in multiple missing data points, repeat the above steps for each missing data point. You are tasked with filling in the missing values using linear interpolation. Specifically, you are asked to take the data point above and below the

missing data point and take the average to fill in the missing data. You will need to write a Python program that reads in the dataset, identifies the missing values, performs the linear interpolation, and writes out the completed dataset. Here is an example of a missing data point and what should your result be after linear interpolation:

Table 1: Missing Dataset

Date	**Data**
1990-01-01	5
1990-01-02	8
1990-01-03	
1990-01-04	
1990-01-05	9

After linear interpolation, your code should generate the following dataset.

Table 2: Completed Dataset

Date	**Data**
1990-01-01	5
1990-01-02	8
1990-01-03	8.33
1990-01-04	8.67
1990-01-05	9

In case either during the outlier detection step or in the base dataset the first and last values are missing, please use the following values to perform the linear interpolation: If your dataset starts in 1970-01-01 and ends on 1996-12-01, use the following first and last values of the columns, depending on whichever value is missing (**Do NOT replace any existing values if they already exist**):

Table 3: Missing Values for linear interpolation

Date	**CPI All items**	**Rent_of Primary residence**	**Monthly Housing Cost**	**CPI Energy**	**US_Dollar Purchasing Power**
1970-01-01	37.900	45.600	35.500	25.100	265.2
1996-12-01	159.100	164.000	154.000	113.900	63.1

If your dataset starts in 1997-01-01 and ends on 2022-12-01, use the following first and last values of the columns, depending on whichever value is missing (**Do NOT replace any existing values if they already exist**):

Table 4: Missing Values for linear interpolation

Date	CPI All items	Rent_of Primary residence	Monthly Housing Cost	CPI Energy	US_Dollar Purchasing Power
1997-01-01	159.400	164.400	155.100	115.200	62.8
2022-12-01	298.900	385.649	310.725	287.176	33.7

Before proceeding to the next task, make sure all your columns are complete. Please do not proceed without getting a complete dataset.

3.3 Performing Wald's test, Z-test and t-test

In this step, we want to check how the mean of monthly stats has changed between 2020 and 2021 (if your dataset is from 1997-2022) and 1994 and 1995 (if your dataset is from 1970-1996). Apply the Wald's test, Z-test, and t-test (assume all are applicable) to check whether the mean of Consumer Price Index for All Urban Consumers: Rent of Primary Residence in 'assigned urban center' and Consumer Price Index for All Urban Consumers: Energy in 'assigned urban center' are different for given years in the urban center. Do this separately for both columns, i.e., you have to compare mean of monthly stats from year 1 with mean of monthly stats from year 2 separately for both columns. Use MLE for Wald's test as the estimator; assume for Wald's estimator purposes that daily data is Poisson distributed.

3.4 Performing K-S test and Permutation test

Infer the equality of distributions between Consumer Price Index for All Urban Consumers: Housing in 'assigned urban center' and Consumer Price Index for All Urban Consumers: Rent of Primary Residence in 'assigned urban center' for the years 2018-2020 (if your dataset is from 1997-2022) of your dataset and 1992-1994 (if your dataset is from 1970-1996) using K-S test and Permutation test. For the K-S test, use both 1-sample and 2-sample tests. For the 1-sample test, try Poisson, Geometric, and Binomial. To obtain parameters of these distributions to check against in 1-sample KS, use MME on 2018-2020 (if your dataset is from 1997-2022) and 1992-1994 (if your dataset is from 1970-1996) Consumer Price Index for All Urban Consumers: Housing in 'assigned urban center' data to obtain parameters of the distribution, and then check whether the 2018-2020/1992-1994 data for Consumer Price Index for All Urban Consumers: Rent of Primary Residence in 'assigned urban center' in your dataset has the distribution with the obtained MME parameters. For the permutation test, use 1000 random permutations. Use a threshold of 0.05 for both K-S test and Permutation test.

3.5 Posterior Distributions, Normality and Linear Regression

For this task, you need to use the percent change values of all five columns as computed in the task above.

3.5.1 Getting the MAP of all posteriors on Gamma Assumption

For this task, you will be using the monthly stats for CPI_Energy in 'assigned urban center' from the urban center assigned to you. Assume the monthly stats follow a Gamma distribution with shape α and rate β. Find α and β using MME on the years 2015-2018/1989-1992 as the sample data. Now, use the year 2019/1993's data to obtain the posterior for α and β via Bayesian inference. Then, use the year 2020/1994's data to obtain the new posterior for α and β, using the prior as posterior after 2019/1993. Repeat till the end of 2022/1996 (that is, repeat till you have posterior after using 2022/1996's data). Plot all posterior distributions on one graph. Report the MAP for all posteriors.

Note, the next three sections (3.5.2, 3.5.3, 3.5.4) must be performed not on the raw values, but on normalized values. Specifically, you need to normalize the data by converting the columns to percentage change with respect to the previous year. To calculate the percentage change of a time series column, you need to compute the percentage difference between consecutive values as they occur over time.

Here are the steps to calculate the percentage change of a time series column:

1. Calculate the difference between the current value and the previous value in the column for each consecutive time period.
2. Divide the difference by the previous value.

Rename all five columns as follows (we will be using these column names for nomenclature purposes. This doesn't matter for the analysis as long as you understand what needs to be done):

1. CPI_All_items_per_change
2. Rent_of_Primary_residence_per_change
3. Monthly_Housing_Cost_per_change
4. CPI_Energy_per_change
5. US_Dollar_Purchasing_Power_per_change

3.5.2 Getting the MAP of all posteriors on Normal Distribution

For this task, you will be using the monthly stats for Consumer Price Index for CPI_Energy_per_change in 'assigned urban center' from the urban center assigned to you. Assume the monthly stats are normally distributed with mean

μ and standard deviation σ. Assume a Normal prior (with mean m and standard deviation s) on μ. Assume μ and σ are found using MME on the years 2015-2018/1989-1992 as the sample data. Now, use the year 2019/1993's data to obtain the posterior for μ and σ via Bayesian inference. Then, use the year 2020/1994's data to obtain the new posterior for μ and σ, using the prior as posterior after 2019/1993. Repeat till the end of 2022/1996 (that is, repeat till you have posterior after using 2022/1996's data). Plot all posterior distributions on one graph. Report the MAP for all posteriors.

3.5.3 Checking Normality

For this task, you will be using the monthly stats for Consumer Price Index for CPI_Energy_per_change and CPI_All_items_per_change in 'assigned urban center' from the urban center assigned to you. If your dataset lies between 1970-1996, take a subset of the dataset between 1990-1995. If your dataset lies between 1997-2022, take a subset of the dataset between 2016-2021. Plot 2 Q-Q plots for the monthly stats of the two columns. Also perform Shapiro-Wilk's test on both columns and report the Shapiro-Wilk's W statistic and the corresponding p-value. Comment on the behavior of both the time-series, whether they are Normal or not based on your experiments. Please give clear and concise reasons.

3.5.4 Linear Regression

Using the dataset given to you, perform 4 linear regression experiments (with the β_0 intercept) to predict CPI_All_items_per_change, by iteratively adding the columns in the following order:

1. US_Dollar_Purchasing_Power_per_change
   (CPI_All_items_per_change vs US_Dollar_Purchasing_Power_per_change)

2. CPI_Energy_per_change
   (CPI_All_items_per_change vs [US_Dollar_Purchasing_Power_per_change, CPI_Energy_per_change])

3. Monthly_Housing_Cost_per_change
   (CPI_All_items_per_change vs [US_Dollar_Purchasing_Power_per_change, CPI_Energy_per_change, Monthly_Housing_Cost_per_change])

4. Rent_of_Primary_residence_per_change
   (CPI_All_items_per_change vs [US_Dollar_Purchasing_Power_per_change, CPI_Energy_per_change, Monthly_Housing_Cost_per_change, Rent_of_Primary_residence_per_change])

In each step, we repeat the experiment with one extra column added. Plot the original data and the regression fit, and report the SSE in each linear regression experiment. Comment on which variables are most relevant in predicting

the CPI_All_items_per_change based on the linear regression experiments performed. (You can perform additional experiments using different combinations of columns if you want and report your findings, however you must mandatorily report your findings for the aforementioned 4 combinations mentioned.) Also please refrain from using any prebuilt python libraries for implementation of linear regression. This should be done from scratch.

3.6 Time-series Prediction

In this task, we want to predict the Consumer Price Index for All Urban Consumers: Rent of Primary Residence in your 'assigned urban center'. Use the dataset to predict the Consumer Price Index for All Urban Consumers: Rent of Primary Residence in 'assigned urban center' (CBSA) for the year of 2021/1995 using data from 2018-2020/1992-1994. You are supposed to predict for each month of 2021/1995. Also to predict next month data, assume you have data from past 3 years plus all monthly data for that year until that month. For example, to predict for March 2021, use data from Jan 2018 - Dec 2020 and Jan 2021 and Feb 2021. Use the following four prediction techniques:

1. AR(3)
2. AR(5)
3. EWMA with $\alpha = 0.5$
4. EWMA with $\alpha = 0.8$

Report the accuracy (MAPE as a % and MSE) of your predictions using the actual data.

3.7 Chi-Square Test

For the Urban Center you are working on, for columns Monthly Housing and All Items in the Urban Center do the following operations: 1) If the data is missing for any month, fill it with the average value for that year as the value for that month. 2) Categorize the data into high and low using the median value as threshold. Now use Chi Square Analysis to find out whether the Average Monthly Housing Cost and All Items in the Urban Center are dependent or not. Note : Use the data from the year 1972-1996 (if your dataset is from 1970-1996) and 1997-2021 (if your dataset is from 1997-2022) for this analysis.

4 Conclusion

Please submit a report with the names of all your group members along with all the codes/programs etc. Please attach all the plots/graphs and report all the answers/conclusions in your report as well.