**ECN 140 Final Project**

Cost of Living Index Report

Yutian Yang

915327218

# **I. Introduction:**

* 1. **Background**

Cost of living index is meant to compare the expenses an average person can expect to incur to acquire food, shelter, transportation, energy, clothing, education, healthcare, childcare, and entertainment in different regions. A cost of living index is also used to track how much the costs of basic expenses rise over a period (<https://www.investopedia.com/ask/answers/100214/how-cost-living-index-calculated.asp>). Nowadays, people are eager to find the most suitable place for them to live in. Different people have different criteria to judge depending on their own interests. Therefore, the cost of living index plays a role of indicator to show people the price of living in that particular city.

* 1. **Research goal**

In this research, I am interested in finding the relationship between the variables that affect the daily living price and cost of living index. In the meantime, I would also like to find the cities and countries that have a higher cost of living. Different techniques would be applied under a computationally efficient procedure that performs simultaneous variable and model selection. I would pick the best model to specify and predict the relationship between variables of different categories that affect the daily living price and cost of living index.

# **II. Dataset:**

**2.1 Data Summary**

The cost of living index is relative to New York City (NYC) which means that for New York City, each index should be 100. If another city has, for example, rent index of 120, it means that on an average in that city rents are 20% more expensive than in New York City. If a city has rent index of 70, that means on an average in that city rents are 30% less expensive than in New York City (<https://www.kaggle.com/debdutta/cost-of-living-index-by-country>). The Cost of Living Index by Cities data describes 536 cities and 7 features of different categories of cost of living index of each cities, all obtained from “Numbeo” (<https://www.numbeo.com/cost-of-living/cpi_explained.jsp>). I will take the 5 different features of daily living prices as our input and choose the cost of living index of each city as our output for our model and all of them are numerical.

**Attributes:**

|  |  |
| --- | --- |
| **Inputs** | **Descriptions** |
| City | City name and country name, state name included if city is inside the United States. |
| Local.Purchasing.Power.Index | It shows relative purchasing power in buying goods and services in a given city for the average wage in that city. |
| Rent.Index | An estimation of prices of renting apartments in the city compared to New York City. |
| Cost.of.Living.Plus.Rent.Index | An estimation of grocery prices in the city compared to New York City. |
| Groceries.Index | A comparison of prices of meals and drinks in restaurants and bars compared to NYC. |
| Restaurant.Price.Index | An estimation of consumer goods prices including rent comparing to New York City. |

Output:

|  |  |
| --- | --- |
| Cost.of.Living.Index | A relative indicator of consumer goods prices, including groceries, restaurants, transportation and utilities. Cost of Living Index doesn't include accommodation expenses such as rent or mortgage. |

**2.2 Data Visualization**

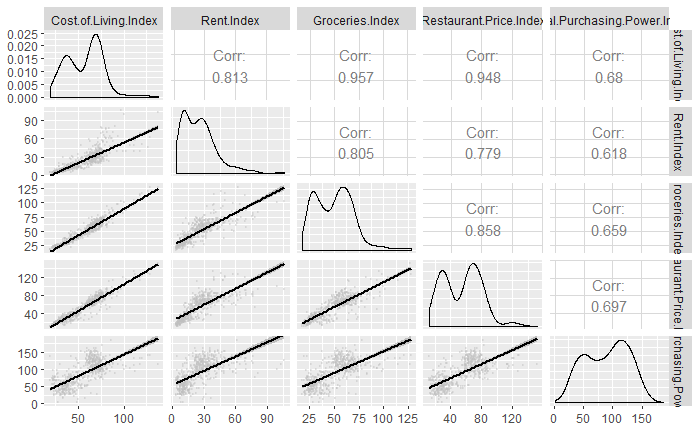
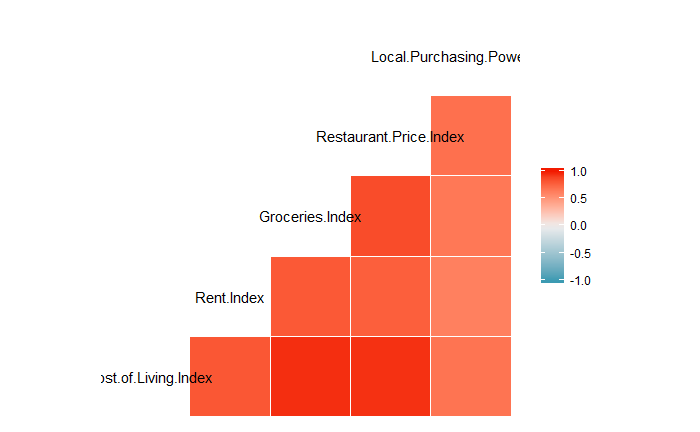


Figure 1: Corrogram Figure 2: Pairwise Correlations

In this step, I was interested in finding the relationship between each feature. By creating the Corrogram for 5 observation variables, see Figure 1, I was able to identify their correlation value via opacity of the color. Thus, we found that all correlation among these variables are positive. The cost of living index vs. groceries index, the cost of living index vs. restaurant price index, and the groceries index vs. restaurant price index have the largest correlation (show high opacity of red color). Thus, I assumed that there may be an interaction between cost of living index vs. groceries index, cost of living index vs. restaurant price index, and groceries index vs. restaurant price index to be assessed in our regression part.

To take a further look at detailed pairwise correlations among these three variables, I created a Pairwise Correlations graph. From Figure 2, I observed that all correlations among these five variables, variables that are related to daily living cost (groceries index and restaurant price index) are near to 1, which shows strong positive linear relationships.

# **III. Model Analysis:**

**3.1 Regression Model with Individual Variables**

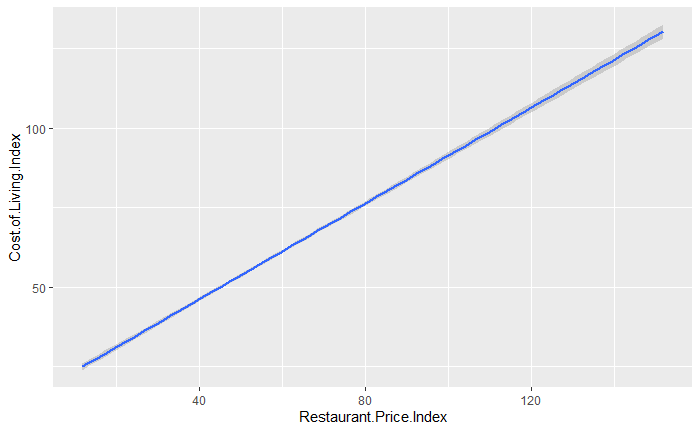
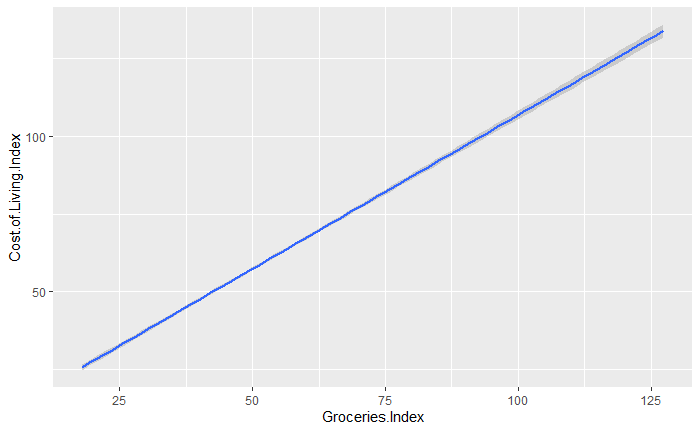


Figure 3: Groceries Index vs. Cost Figure 4: Restaurant Price Index vs. Cost

of Living Index of Living Index

At this part, I analyze and find the most important variable affecting cost of living index. I first applied each variable individually into the linear regression model to find their relationship with cost of living index. From linear regression models with single variable, I conclude that both Groceries Index and Restaurant Price Index are the most important variable for cost of living index because they both give very high r square value (Groceries: 0.9151, Restaurant: 0.8992).

Cost.of.Living.Index = β0 + β1Groceries.Index + u

This model gives R square 0. 9151.

Cost.of.Living.Index = β0 + β1Restaurant.Price.Index + u

This model gives R square 0.8992.

Based on that, I applied all variable without adding interaction into the model, I found out that rent index is not significant, so I eliminated them from the regression model.

**3.2 Model Selection**

Then, we manually implemented the Forward and Backward Selection. I start with an empty model by adding one variable to the model and record the R square results. I then keep adding until the R square results no longer decreases and keep the model. Then I start with a full model by deleting one variable from the model and also stop until the R square results no longer decreases and keep the model. Finally, I choose the same linear regression model drawn from both model selection results.

Cost.of.Living.Index = β0 + β1Local.Purchasing.Power.Index + β2 Groceries.Index + β3Restaurant.Price.Index + u

For this model, we have R square equals to 0.9773, and Adjusted R-square equals to 0.9771. For each parameter, we have β0 = 8.654063, β1 = -0.015296, β2 = 0.567363, and β3 = 0.396296.

**3.3 Interaction Model**

According to the corrogram (Figure 1), I noticed that there are correlations between the cost of living index and groceries index, the cost of living index and restaurant price index, and the groceries index and restaurant, which suggest that groceries index and restaurant price index are significant factors and there is large positive correlation between these two factors, so I added them as the interaction term in the model.

Cost.of.Living.Index = β0 + β1Local.Purchasing.Power.Index + β2 Groceries.Index + β3Restaurant.Price.Index + β4Groceries.Index \* Restaurant.Price.Index + u

Under this model, we have R square equals to 0.9788, and Adjusted R-square equals to 0.9786. For each parameter, we have β0 = 5.3991840, β1 = -0.0248045, β2 = 0.6534650, β3 = 0.4647246, and β4 = -0.0012501.

**3.4 Hypothesis Testing**

**3.4.1 Updated Dataset**

For hypothesis testing, I am dividing the data into two groups. Group 1 contains the cities that are outside of the United States, and group 2 contains the cities that are within the United States. In this part, I am going to look at the average cost of living index by cities and test on average if cities have the same cost of living index inside and outside of the United States.

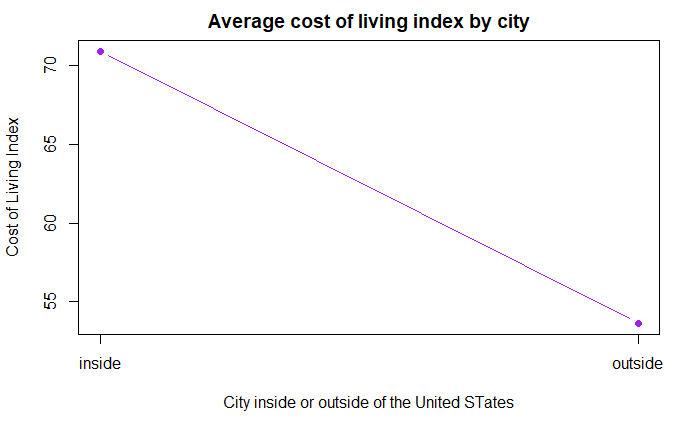
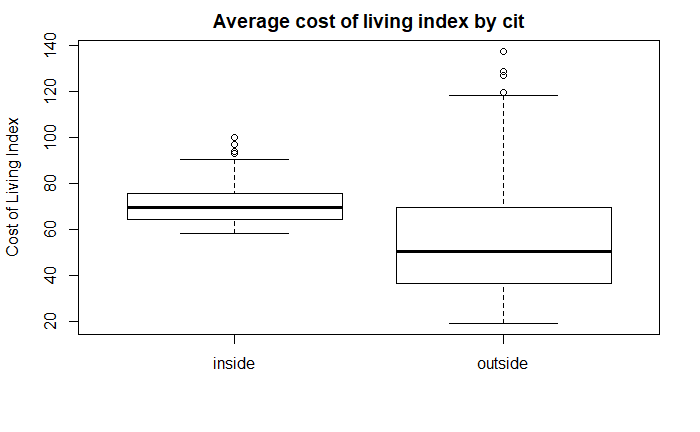


Figure 5: Average cost of living Figure 6: box plot of cost

index by cities of living index by cities

From figure 5 and figure 6, we can make an assumption which should suggest that there is a significant difference between the average cost of living index inside the United States and outside of the United States. To verify that statistically, I made a hypothesis test below:

H­0: µinside = µ­outside vs. H1: µinside ≠ µ­outside

Our null hypothesis is saying that there is no significant difference between the average cost of living index inside of the United States and the average cost of living index outside of the United States. Our alternative hypothesis is suggesting that the difference between the average cost of living index inside of the United States and the average cost of living index outside of the United States is indeed significant.

|  |  |  |
| --- | --- | --- |
|  | Inside of the United States | Outside of the United States |
| Means | 70.9065 | 53.6451 |
| Standard Deviation | 8.8109 | 20.5322 |
| Sample Size | 110 | 426 |

From the table we can see that the difference in mean and standard deviation is relatively large, so it is reasonable for us to expect that the result of the hypothesis test is to reject the null hypothesis which means there is significant difference between the average cost of living index inside of the United States and the average cost of living index outside of the United States.

**3.4.2 Two Sample t-test Result**

The Two Sample t-test provides following result: The t statistics of the Two Sample t-test is 13.257. The p-value of the Two Sample t-test approximately equals to 2.2e-16 (< 0.01), which means that if in reality there is no difference between the average cost of living index inside of the United States and the average cost of living index outside of the United States (H0 is true), we will observe our data or more extreme case with probability 2.2e-16 (<0.01). Since the p-value is smaller than α = 0.01, we reject H0 and conclude that there is a significant difference between the average cost of living index inside of the United States and the average cost of living index outside of the United States.

To further analysis this question, I conduct a 95% confidence interval to have a deeper understanding of the difference between the average cost of living index inside of the United States and the average cost of living index outside of the United States. The 95 percent confidence interval result is (14.70193 19.82070), which means we are overall 95% confident that the true average cost of living index inside of the United States is greater than the average cost of living index outside of the United States by between 14.70193 and 19.82070.

**3.5 Logistic Regression**

For this part, I am interested in testing what factors affect the odds of having high cost of living index in that city, so I divided the dataset into two groups: high cost of living index cities and low cost of living index cities, which separated by the mean of cost of living index = 57.18757 (https://www.statisticssolutions.com/binary-logistic-regression/). I first added all individual factors into the model, but the model suggests that rent index and local purchase power are not significant factors, therefore I eliminated both from the model. So, we have final model as below:

Cost.of.Living.Index = β0 + β1Groceries.Index + β2Restaurant.Price.Index + u

Under this model, for each parameter, we have β0 = -39.9455, β1 = 0.4897, β2 = 0.3192.

The intercept (β0) = -39.9455 means the log odds of a city with a groceries index and restaurant price index of zero having a high cost of living index.

The coefficient for groceries index (β1) is 0.4897 which means the expected change in log odds for a one-unit increase in the groceries index. The odds ratio can be calculated by exponentiating this value to get 1.631827 which means we expect to see about 63.2% increase in the odds of a city having a high cost of living index, for a one-unit increase in groceries index.

The coefficient for restaurant price index (β2) is 0.3192 which means the expected change in log odds for a one-unit increase in the restaurant price index. The odds ratio can be calculated by exponentiating this value to get 1.376027 which means we expect to see about 37.6% increase in the odds of a city having a high cost of living index, for a one-unit increase in restaurant price index (<https://stats.idre.ucla.edu/other/mult-pkg/faq/general/faq-how-do-i-interpret-odds-ratios-in-logistic-regression/>).

# **IV. Conclusion:**

This study provides an effective regression model with multiple important variables to predict the cost of living index in different cities. To analyze the most important variable that affects the cost of living index, I concluded that both groceries index and restaurant price index are the most important variables in the model. By adding the rest of the variables into the model, I found out that rent index and local purchasing power are not significant, so I removed them from the regression model. After looking at the corrogram (figure 1), it suggests that groceries index and restaurant price index are highly correlated, so we added them as interaction terms in the regression model.

I am also interested in the relationship between the average cost of living index inside and outside of the United States, so I conducted a hypothesis test and calculated a 95% confidence interval of it. From the hypothesis test, we rejected H­0 which means that the difference between the average cost of living index inside of the United States and the average cost of living index outside of the United States is indeed significant. From the 95% confident intervals, I concluded that we are overall 95% confident that the true average cost of living index inside of the United States is greater than the average cost of living index outside of the United States.

I then conducted a logistic regression to analysis the odds of having high cost of living index in different cities. Similar as linear regression, I found out that groceries index and restaurant price index are the most important factors in the model, and I eliminated rent index and local purchasing power since they are not significant. After analyzing the model, I concluded that one unit of increase of groceries index can affect the odds of a city having a high cost of living index the most.

The results of the study may offer a practical inference for people choosing cities to live. Nowadays, the decisions of people choosing cities to live in are mostly based on their subjective judgement, which might be biased varying among people. The models in this study provides the possibility of improving the objectivity of the decision. From economy perspective, it may solve the problem of mindlessly moving for some people and improve the economic efficiency for people to choose where to live in.

# **V. References:**

1. Binary Logistic Regression. (2019, October 2). Retrieved from https://www.statisticssolutions.com/binary-logistic-regression/
2. Cost of Living. (n.d.). Retrieved from https://www.numbeo.com/cost-of-living/cpi\_explained.jsp
3. Cost of Living Indexes and Social Security's Cost of Living Increases. (2020, February 5). Retrieved from https://www.investopedia.com/ask/answers/100214/how-cost-living-index-calculated.asp
4. Debd. (2018, November 14). Cost of Living Index by Cities. Retrieved from https://www.kaggle.com/debdutta/cost-of-living-index-by-country
5. How do I Interpret Odds Ratios in Logistics Regression? (n.d.). Retrieved from https://stats.idre.ucla.edu/other/mult-pkg/faq/general/faq-how-do-i-interpret-odds-ratios-in-logistic-regression/