Using Stepwise Regression and K Fold Cross Validation to Build a Model that Predicts Unemployment Level of NJ

**Abstract**

The current project sought to identify a model that best predicted the effect of seven economic factors (Income, Minimum Wage, GDP, Population, Degrees Earned, Labor Force, and Rental Vacancies) on unemployment in New Jersey using data sourced through *FRED*. Analyses yielded two different models: one with (modelnew2) and one without the Population the factor (modelnew1). K fold Cross validation was used to compare the models. Results showed that modelnew1 (Y~GDP+Min.wage+Labor.force) was the best model for predicting unemployment because it had the lowest RMSE (0.384609) and MAE (0.3733439).

**Introduction**

The current project sought to identify a model that best predicted the effect of economic factors on unemployment in New Jersey. These factors included: Income, Minimum Wage, GDP, Population, Degrees Earned, Labor Force, and Rental Vacancies. When we first started to build the model, we had no specific hypotheses about what we were going to find. Thus, this project was exploratory in nature and is intended to describe a process.

Data from a larger dataset were assessed using scatterplots and regression analyses. Analyses yielded two different models (one with the factor population and one without), which were then compared using K fold cross validation.

**Methods**

            All of the data used in the project come from https://fred.stlouisfed.org/, which is a reliable economic source, which included data collected over an eleven-year period (Jan 2009 to Jan 2019). Specific data collection methods and materials used to collect data are unknown. But for the purposes of this paper, it is assumed that all standards have been met.

Data were cleaned and organized to report changes on the first of each year and scatterplots were then created to visually inspect data. Alpha = .05 was used for the entirety of this project. Scatterplots indicated that unemployment was negatively correlated with education and GDP, respectively. In contrast, unemployment was shown to be positively correlated with rental vacancies. See Figure 1.

Calendar

Description automatically generated

Figure 1: Scatterplot

Additionally, a base model was used to assess all factors together. An F test was then performed on the base model, which showed a significant difference between our model and the intercept-only model. However, there were no significant factors identified from t-tests. The ANOVA of the base model showed that Income, Population, and GDP significantly contributed to unemployment. Furthermore, all possible regressions, both and forward stepwise, and best subset were conducted in order to improve upon the base model, as well as to remove some of the noise created by extraneous factors.

These subsequent findings led us to two models. Modelnew1 was chosen by all four processes and modelnew2 was included for its higher R squared adjusted. Both appeared to be a better fit than the full model. Thus, in order to assess the primary concern of this paper, K fold cross validation was used to examine which of the resulting models better predicted unemployment. Results from the K fold cross validation are presented in the next section.

**Results**

From the results of the K fold cross validation showed that modelnew1 (Y~GDP+Min.wage+Labor.force) was the best model for predicting unemployment because it had the lowest RMSE (0.5179042) and MAE (0.4808447). *See appendix for output.*

**Discussion**

Scatterplots revealed that unemployment was negatively correlated with economic conditions that are generally considered “good”, such as level of education. For example, as the number of degrees earned, as well as GDP increased, Unemployment generally trended down. (See Figure 1)

The residuals were not satisfactory because they violated the assumption of normality, but for the purpose of this project they were ignored. In a real-life scenario, we would investigate transformations or try to fit higher order models.

Because analyses were limited to a small dataset, K fold cross validation was appropriate because there was not enough data to split and create a training set and test set. The use of K fold cross validation may also have helped to avoid overfitting. Thus, while modnew2 had a higher R sq adjusted value compared to modnew1, the simpler model is the preferred model. This is supported by findings from Nested f test, which showed that no extra factors were significant beyond the three (GDP, minimum wage, labor force) included in modnew1.

Since only some of the factors included in the dataset go beyond 2019, the economic conditions included in this project do not reflect the potential economic impacts caused by COVID-19. As such, I am curious about how this model would be different in the coming years. Will 2020 and 2021 be outliers or will we learn about a new important factor for predicting unemployment in NJ? It would also be interesting to see how well this model could be used to predict unemployment in other states.

Appendix A

Output

Text, letter

Description automatically generated

Text, letter

Description automatically generated

Table

Description automatically generated

Text

Description automatically generated

Table

Description automatically generated

Text

Description automatically generated

Table

Description automatically generated

Table

Description automatically generated

A picture containing text, receipt, screenshot

Description automatically generatedTable

Description automatically generated