Olivya Pastis

Professor Levkoff

ECON 494

18 November 2020

Project II Written Report

Through RStudio, I opened my TIDY data from Project I and renamed it “Project II,” and then I commenced the process of preparing my data in order to apply predictive modeling to observe and describe the varying linear relationships among the variables of income, child mortality, population, median age, life expectancy, and the employment rate for individuals 15+ within the year 2019 for each country. I started organizing my data by creating more concise names for the variables to use within the models, and then I incorporated nonlinear transformations to utilize within the models I would build as well. For partitioning the data, I determined that I would use 70% of the sample for the training partition, and established the number of observations while simultaneously ensuring the number would round down to the nearest integer with the floor() function. In the next step, I set the seed to make the partition reproducible, and then created a vector with the shuffled row numbers of Project II. As a result, I confirmed my training and testing partitioned data, and proceeded to create the linear regression models.

From the hypothesized relationships I found in Project I, I observed how income and child mortality play crucial roles as independent variables in determining the outcomes of the other variables. Despite a lack of significant correlation among a majority of the variables, I noticed that these two variables greatly contributed to the result of life expectancy, so I decided to choose life expectancy as the consistent y-variable within my linear regression models. Through the utilization of income and child mortality as my x-variables, I referenced the directly proportional relationships I discovered in Project I to strategically implement polynomial transformations as well as optimal combinations of the variables and their transformations within different linear regression models. Then, I proceeded to use five different models to interpret the relationships by comparing the summary statistics and analyzing the differing model complexities.

* Linear Regression Models: y = B0 + B1x1 + B2x2 + B3x3

1. Life Expectancy ~ Income
   * y = 71.12 + 0.0001084x
2. Life Expectancy ~ Income + Income2
   * y = 68.56 + 0.0004021x + -4.034e-9x
3. Life Expectancy ~ Income + Income2 + Income3
   * y = 67.10 + 0.0006826x + -1.268e-8x + 6.126e-14x
4. Life Expectancy ~ Child Mortality
   * y = 75.44 + -0.08411x
5. Life Expectancy ~ Income + Child Mortality
   * y = 73.40 + -8.735e-5x + -7.053e-2x

![Chart, scatter chart

Description automatically generated]()

For my first linear regression model, I started off by creating a plot to interpret the impact of income on life expectancy by using income as my sole regressor. The linear regression equation contains a positive coefficient, illustrating a directly proportional relationship. Income and life expectancy proved to be the most strongly correlated among the variables, in which RStudio depicted the high significance value between the two variables with a three-star code. Given that the p-value is .0000131 compared to the significance test of 0.05, the model fit is very significant and will significantly increase with the additions of the nonlinear transformations. However, the Jarque-Bera test displayed that the residuals do not follow a normal distribution, given that the histogram of the residuals is strongly negatively skewed and the p-value is .00843.

![Chart, scatter chart

Description automatically generated]()

In the second prediction model I created, I used the first graph as a baseline and included a quadratic polynomial transformation of income, which I depicted in green on the plot. Since the value of the quadratic transformation is negative, the green line displays a negative, parabola-like curve. The p-value decreased even further to 2.47e-11, portraying the high significance of the model, in which the plot conveys how the curve fits the data better visually too. Additionally, the F-test illustrates the joint significance of the mode, given that it’s a high non-negative number along with a small p-value. The adjusted R2 increased from 0.094 to 0.227 as well, indicating the addition of an important independent variable.

Likewise, for my third model, I added a cubic polynomial transformation to the foundational equations of the first and second models. Similar to the previous models, the significance codes illustrate the high significance of the model, where the p-value was 4.168-12 . The Jarque-Bera test also demonstrated that the data follows a normal distribution, since the calculated p-value is 0.2643. Compared to the other models, I believe this one performs the best on the training data due to the multiple polynomial transformations.

For my fourth model, I changed the independent variable from income to child mortality in order to compare the difference in linear relation by changing the x-variable. To my surprise, the significance value was higher for these two variables than I initially anticipated, given that the p-variable is .000002543 and is smaller than life expectancy compared to income. However, the multiple R2 value is 0.1142, indicating a low percentage in which the variance in the dependent variable of life expectancy can be explained by the independent variable of child mortality. A histogram of the residuals also illustrates a negative skewness of the variables, in which the p-value is 0.1004.

Within the fifth model, I used both child mortality and income to compare against life expectancy. The p-value is 2.383e-8 and the F-statistic is a high non-negative number, depicting high overall significance for the model. Both independent variables share the same three-star code of significance within how they’re related to life expectancy as well. Income and Child mortality also share low standard errors, through which they exceed passing the significance test and prove the accuracy within their respective sample distributions.

Since root-mean-square-error is measured in the same units as the output that’s being predicted, it was the most optimal unit of measurement for predicting the error. Through the use of RMSE, I observed a clear decline in the metric from the first model to the third model. Given that I included additional transformations from the first to the third model, the data showcased the principle of how adding nonlinear transformations of one of the independent variables improves a linear regression model’s fit. The baseline graph of y = life expectancy, x = income resulted in an RMSE out-of-sample sample of 8.552, which then decreased to 8.549 with the inclusion of income squared. The addition of a cubic transformation caused the RMSE in-sample data to drop even further, causing it to be the smallest error.

Ultimately, although they were all relatively close, the third model had the smallest out-of-sample sample error in comparison to the other five models with an RMSE out-of-sample error of 8.548. As I expected, the fourth model had the highest RMSE of 8.5531, which was the model comparing life expectancy to the sole independent variable of child mortality. Given the inclusion of a quadratic and cubic transformation along with the original variable of income as the independent variables, it makes sense that the third model is the best model to predict the variable of life expectancy.