Introduction

The objective of this data analysis project was to gather data sets from multiple sources, clean and merge the data, and then perform statistical analysis. For a topic I chose to analyze the relationship between the number of deaths worldwide due to leading risk factors and the financial prosperity of these countries, as measured by their Gross National Income per capita. After collecting and cleaning data I created a first order linear regression model to predict GNI by risk factors. Prediction does not necessarily indicate causation, meaning that reducing a cause of death does not guarantee that GNI will increase. However, the model could be one source of information to help the developed world decide which risk factors to focus on to reduce human suffering and lift countries out of poverty. The final had an Adjusted R-squared value of .676 which measures the proportion of variance in the dependent variable (GNI) that can be explained by the independent variables (risk). For a simple linear model without higher order terms, this is a good starting point for crafting a more comprehensive model.

### Data provenance

I used three data sets for this project. The country populations and Gross National Income per Capita (GNI) came from the World Bank DataBank. The deaths by risk factors came from Our World in Data. All three are cited in the Sources section. The Death by Risk Factors included data from 1960 to 2019. The Population and GNI data included data from 1990 to 2022. I chose 2019 for my observations.

There were some entries listed which did not have country codes. On further examination I realized these were not official countries and therefore dropped them from the table. For example, “African Region”, and “East Asia and Pacific”. England, Scotland, Ireland, and Wales were also dropped as together they form Great Britain.

There were 19 countries without GNI numbers for 2019. Rather than mix and match data sources to fill in these numbers I chose to delete those counties. The death totals were in absolute numbers. When creating the columns of data for model fitting, I converted these numbers to deaths per 100,000 people. I made minor edits to column names to make sure it was clear that the data was for 2019.

Merging Data

I read each of the datasets into Python using Pandas read\_csv function. I only read in the country, country code, and 2019 columns for the GNI and population data sets as their data was in wide format. The risk factors data was in long format, so all the data was read in and then filtered for just 2019 rows. I successively merged the GNI, risk factors, and population data frames into a single data frame, performing an inner join on country code. One final data edit. I filtered the data to only select countries with less than the median GNI. I wanted the model to focus on how lower income countries might rise to become mid income. The higher income countries could possibly skew the data.

Create a Linear Regression Model

I used a stepwise regression technique called forward selection to create a first order linear model. This is known as a “greedy” algorithm because it takes the best option immediately available, rather than trying every possible combination. The algorithm starts by creating one variable models for each independent variable and choosing the variable with the lowest t-test p-value. In layman’s words, the variable that appears most likely to be correlated to the dependent variable. The process then creates a series of two variable models using the variable just selected plus each of the remaining variables. Again, the best of the second variable p-scores is kept and the process continues. I had my algorithm stop when the Adjusted R-squared value for the model stops increasing. R-squared is a calculated value that tells what percentage of the sample data variation is explained by the model. Adjusted means that the score can be reduced as you add more variables.

I created nested loops to perform the regression. The inner loop created an n+1 model with each of the remaining variables. The outer loop selects the one with the lowest p-value. If the R-squared adjust value increases the variable gets added to the model and the loop continues. The outer loop ends when either the Adjusted R-squared stops increasing, or all 28 variables are used.

Forward Stepwise Regression Results



The last line of the table shows that the highest Adjusted R-squared was .678, so roughly two thirds of the variation in GNI of low to median income countries is explained by our first order model.

Model Summary

The full summary of the final model is listed below.



The F-statistic and its p-value tell if the model is useful. A low p-value is good. 1.66e-18 means our model is useful. The R-squared, Adjusted R-squared, BIC, and AIC are all measures of how well the model accounts for the variation between sample data and model data. The constant, const, is the y intercept of our model. If all the deaths were set to zero our model predicts the GNI would be $2914. We do not have a country in our survey that has zero deaths in all categories so the intercept should not be interpreted or extrapolated on its own. The coefficient, coef, is the slope of each variable in the equation. A negative number means that an increase in that variable leads to a decrease in the GNI. For example, for each additional death due to a diet low in fruits, with all other variables staying constant, our model predicts that GNI will decrease by $45.08. The negative coefficients take a bit of explanation. GNI is predicted to increase $6.16, all other variables staying constant, for each additional death due to smoking. A possible explanation is that you need a certain level of wealth to be able to afford cigarettes. A similar explanation could be made for death due to alcohol consumption.

The t-test is a measure of how likely the slope of the variable is not zero. A t-test of zero equates to a slope of zero, which would mean there is no relationship between the independent and dependent variable. The p-value is the probability that the slope is not zero, so a low score is good. Some of our model individual p-values are higher than we would like. A typical cutoff is .05, meaning there is less than 5% chance that the slope is zero. A high p-value means we cannot conclude that the independent variable affects the dependent variable. However, since the overall model is useful, if the goal is to get an accurate model, then we can accept that the individual variable p-values are not all ideal. The Durbin-Watson score at the bottom of the table is 1.89. Scores for this test range from 0 to 4. It measures the autocorrelation of the data. Does one score influence the next score? A value of 2.0 indicates there is no autocorrelation. 1.89 indicates little autocorrelation.

Model Validation

Residual analysis involves investigating if the residuals (the error terms) have an expected value of zero, the variance (spread of values from the mean) is constant, the values are normally distributed, and the observations are independent.

I tested for variance by plotted residuals versus fitted model (figure on left). The data points expand into a cone shape as they move to the right. I tried a log model to flatten the variance (figure on right). The variance is improved somewhat, but not enough for me to go ahead with a log-linear model.

A graph with blue dots

Description automatically generatedA graph with blue dots

Description automatically generated

I also created a model using the entire sample data set, not just the low to mid incomes. The residual variances were even further skewed. A higher order model might be needed to capture the full shift from low to high GNI.

I tested for normal distribution using the qqplot function. A QQ plot can be used to plot residual quantiles versus ideal normal distribution quantiles. The straight line in my plot means the residual data matches an ideal normal distribution. There are a couple stray points at the tails, but overall, the residuals appear to be normally distributed.

A graph with blue dots

Description automatically generated

My final check is for data points that are influencers and high leverage points. The Cook’s Distance function accomplishes this. The large spheres to the right are influencers and to the top and bottom are leverage. The bubble on the far right is for Moldova. Looking at the data for Moldova they are ranked in the top four for multiple death per risk categories per 100,000.

A diagram of blue dots

Description automatically generated

Conclusion

I set out to combine a data manipulation exercise with statistical analysis using Python and a variety of data, visualization, and statistical packages. Three data sets from two main sources were merged. Data was removed where the country was not a “real” country, and a few countries were removed for lack of GNI data. The data was filtered to only include low to mid GNI countries. The program can be easily changed to include or exclude countries based on their GNI. After all the updates there were still 98 countries which was sufficient to create a model. The program performed a forward selection stepwise regression. The resulting model used 12 different predictors and was able to explain over two thirds of the variability in GNI per capita. This is encouraging, as much more can be done to build a more comprehensive model. Second order and interaction terms can be tested to see if they improve the score.

Sources

<https://ourworldindata.org/grapher/number-of-deaths-by-risk-factor>

<https://databank.worldbank.org/reports.aspx?source=2&series=NY.GNP.PCAP.CD&country=>

<https://databank.worldbank.org/reports.aspx?source=2&series=SP.POP.TOTL&country=>