| WHO Life Expectancy Project Report

| by Chrisitan Hardin

DATASET OVERVIEW

This dataset was taken from Kaggle, and is originally from the World Health Organization (WHO). The main groups of data contained in the dataset are mortality rates, economic indicators, disease rates, and population health indicators. It contains only quantitative variables, in the form of numbers and percentages.

The research question I set out to answer using SLR, MLR, and Logistic Regression is:

- ❖ Which variable alone has the greatest power to predict life expectancy? (SLR)
- ❖ What model can give us the greatest predictive power without using excessive variables? (MLR)
- ❖ What is the probability someone will live longer than the median life expectancy based on schooling? (LR)

SIMPLE LINEAR REGRESSION

- ❖ Equation: Life Expectancy = 35.97 + 51.06(income composition of resources)
- ❖ Slope of the above model has a strong positive correlation. There were initial outliers from missing data.
- ❖ R-Squared = 79%. P-Value = 0.000.
- ❖ The Assumptions for SLR are normal distribution, independent observations, and constant variance. This model meets all of the above assumptions.

MULTIPLE LINEAR REGRESSION

- ❖ Equation: Life Expectancy = 38.478 – 0.2259(Alcohol) + 0.2837(Total Expenditure) – 0.4752(HIV/AIDS) + 53.1(ICR) – 0.3096(Schooling).
- ❖ Used Stepwise regression to select variables on a sigma of .01. Most variables had significant p-values but were eliminated due to insignificant coefficients compared to the ones included.
- ❖ R-Squared = 87.43% and Adj. R-Squared = 87.4%.

LOGISTIC REGRESSION

- ❖ Created a column with 1 = life expectancy >= 72.2, and 0 if not. This model was based on schooling.
- ❖ The odds ratio is 2.4185, meaning the odds increase by 136% each time there is an increase of 1.
- ❖ The deviance R-Squared is only 44.05% which is not as high as desired, however the area under the ROC curve is 90.03% which is satisfactorily close to 1.

KEY TAKEAWAYS & ANALYSES

SLR: It was surprising to see that income composition of resources was the single best predictor of life expectancy over another economic indicator like GDP or total expenditure on health. Perhaps it just goes to show that ICR is just a more wholistic account of a nation's economic position.

MLR: Here, I was happy with the model created. While it made sense for Alcohol and HIV/AIDS to have a negative effect, I was surprised to see that schooling did as well. Perhaps when also considering a nations ICR, more schooling indicates a less balanced use of its resources.

LR: With what was found in MLR about schooling, I wanted to see how schooling increased the probability that someone would live longer than the median age of 72.2 in the data provided. When I made a model, more schooling significantly increased the probability that someone would be live past 72.2, which may merely be an indicator of a nation's economic health; with better economic health comes more education opportunity.

Overall, I was very surprised to see disease data not having nearly as much predictive power as economic data.