Exploring Factors Affecting Life Expectancy Using Linear Techniques

Spencer Slaton, Leon Weingartner

*Abstract*—This research paper aims to explore the factors that affect life expectancy using linear models. The study utilizes a real-world dataset to derive insights and conclusions based on the techniques learned in class. The analysis goes beyond interpreting coefficient results from a multilinear regression model and provides an in-depth study of the factors that impact life expectancy.

# INTRODUCTION

T

his study begins by exploring the data using descriptive statistics and visualization techniques. Then, various linear models are developed to examine the relationship between life expectancy and several independent variables. The models are evaluated using diagnostic tools such as residual analysis, Cook's distance, and variance inflation factor (VIF) to ensure their validity and reliability.

In addition to traditional linear models, the study also employs more advanced techniques such as heteroskedastic coefficient estimation, interaction effects, and regularization methods to gain deeper insights into the factors that affect life expectancy. The results of the analysis provide a comprehensive understanding of the complex relationship between life expectancy and various socio-economic, demographic, and health-related factors.

This research paper showcases the importance of using advanced linear modeling techniques to conduct in-depth analyses of real-world data. The study contributes to the existing literature on factors that impact life expectancy and provides valuable insights for policymakers and public health practitioners.Project Goals

## Data (Imputation techniques, transformations

We use a dataset that is a combination of life expectancy data provided by GHO (Global Health Observatory) and UNESCO (United Nations Educational Scientific and Culture Organization). The data has life expectancy information about every country for every year from 2000 until 2015. There are around 30 potential predictors. Our target variable “life\_expect” is the life expectancy at birth in years. We use a variety of imputation techniques to obtain 3 datasets. With some initial EDA we inspect the distributions and correlations of our feature variables to understand which features should be left out or transformed logarithmically.

## Analysis Topics (OLS, model selection, feature selection, aic, bic, lasso, bootstrapped lasso, EHW)

## Model Diagnostics

We will utilize leverage scores, outliers, Cook’s distance, variance inflation factor (VIF) and residual analysis to examine each proposed model. We want to identify influential observations in the dataset that may be driving the results of the analysis. VIF is an indicator of multicollinearity. Visualizing the residuals in a QQplot allows us to check that the residuals are normally distributed.

## Final Regression Model

## Additional Work

# Conclusion

Statistical significance testing in machine learning is an underused model comparison tool that can lead to more reliable model selection and strengthen ML research. We’ve

References

1. D. Ulmer, C. Hardmeier, J. Frellsen “Easy and Meaningful Statistical Significance Testing in the Age of Neural Networks” URL https://arxiv.org/abs/2204.06815
2. Nils Reimers and Iryna Gurevych. Why comparing single performance scores does not allow to draw conclusions about machine learning approaches. arXiv preprint arXiv:1803.09578, 2018.
3. Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. The hitchhiker’s guide to testing statistical significance in natural language processing. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 1383–1392, 2018
4. Alessio Benavoli, Giorgio Corani, Janez Demšar, and Marco Zaffalon. Time for a change: a tutorial for comparing multiple classifiers through bayesian analysis. The Journal of Machine Learning Research, 18(1):2653–2688, 2017.
5. Rotem Dror, Segev Shlomov, and Roi Reichart. Deep dominance - how to properly compare deep neural models. In Anna Korhonen, David R. Traum, and Lluís Màrquez (eds.), Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers, pp. 2773–2785. Association for Computational Linguistics, 2019. doi: 10.18653/v1/p19-1266. URL <https://doi.org/10.18653/v1/p19-1266>.
6. Taylor Berg-Kirkpatrick, David Burkett, Dan Klein, An Emperical Investigation of Statistical Significance in NLP. URL <https://aclanthology.org/D12-1091/>
7. **Project Overview Video URL:** [**https://www.youtube.com/watch?v=5Sb3s7QCiV4**](https://www.youtube.com/watch?v=5Sb3s7QCiV4)
8. Github code files URL: https://github.com/sbslaton/230-254-Final-Projects/tree/main/254