Predicting poverty using regression

Luka Urbanč
Jožef Štefan Institute
Ljubljana, Slovenija
urbancluka3@gmail.com

Marko Grobelnik
Jožef Stefan Institute
Ljubljana, Slovenija
marko.grobelnik@ijs.si

Joao Pita Costa
IRCAI, Quintelligence
Ljubljana, Slovenija
joao.pitacosta@quintelligence.com

Luis Rei
Jožef Stefan Institute
Ljubljana, Slovenija
luis.rei@ijs.si

Abstract

Poverty reduction is the first Sustainable Development Goal set by the United Nations to be achieved by 2030, but current data indicates that the progress is insufficient. The diverse factors influencing poverty across different nations pose a challenge in developing effective predictive models. This paper evaluates the use of various regression models to predict poverty rates using a comprehensive dataset of 111 variables from sources such as the UN and the World Bank. The data, spanning multiple domains like political stability, education, and economic conditions, was preprocessed and transformed to create auxiliary features and interactions. Among the models, Ridge regression yielded the best results, achieving a Root Mean Square Error (RMSE) of 3.6, indicating high predictive accuracy on a global scale. This study highlights the importance of addressing multicollinearity and incorporating a wide range of features to improve the generalizability of poverty prediction models. Future research should explore more complex methods, such as neural networks, and refine model hyperparameters for enhanced performance.

Keywords

poverty, linear regression, lasso regression, ridge regression, elastic net regression, sustainable development goals

1 Introduction

The need to eradicate poverty has been a long standing issue, which was globally recognized numerous times, most importantly in the United Nations (UN) Sustainable Development Goals (SDGs), being given the number one spot of SDG1: "End poverty in all its forms everywhere", which should be achieved by 2030. The latest UN report on the progress made in achieving SDG1 indicates Poverty has returned to pre-pandemic levels in middle- and high-income countries, with poverty in low income countries still a fraction above those reported in 2019. While the trends seem to be going in the right direction, the UN warns that the current pace of improvement is insufficient to reach the agreed goals before 2030. This raises the question of what impacts poverty rates the most and how countries can most effectively reduce poverty levels.

To fully understand and address the issue of poverty, one must navigate several definitions, which can often lead to confusion. The baseline definition used in this paper is the poverty line as is defined by each country individually, recognizing that different countries have different measures of, e.g., what life conditions and how much income makes an individual reach a "poor" status, as well as how we can normalise this to better compare these relative indicators between countries. We are still missing a clear theory in poverty research, despite the issue existing for a number of decades [2]. With that being said, some authors have already explored the causes of poverty. For instance, corruption, political instability, ineffective local governance, government polices, gender inequality and short-term wage replacement policies, such as maternity leave benefits and sickness pay, impact relative poverty [6, 7]. When assessing what people believe causes poverty some geographical differences emerge. For example, the United States are mostly of the thought that an individuals traits are responsible for poverty, while countries in Europe have a blend of individualistic, fatalistic and structural beliefs such as lack of will, bad luck and social injustice respectively [4].

Machine learning (ML) has also been used in academic research to identify trends and analyze data in most fields, including poverty research. Although a number of papers have already been published on the use of ML to predict poverty [1, 10, 12, 5, 3, 8] (for more see [11]) including combining satellite images and neural networks to help predict poverty in five African countries [5], most take a limited number of variables. Usmanova's literature review found 22 papers published between 2016 and March 2022, with a total of 57 AI methods applied, the most popular being random forest, used in more than half of all papers reviewed. It also found most papers focus only on African and South Asian countries, a finding consistent with our own [11].

In this paper we focus on the following research questions: (i) can regression be useful to identify the most influential features, from a large amount of global indicators; and (ii) can direct and indirect causality relations be identified that signal new indicators relevant to the Poverty-related issues?

2 Data

To address the research questions, we utilized 111 primary variables from sources such as the UN and the World Bank, aggregated through the Our World in Data portal. These variables span diverse domains, including political stability, policies, education, healthcare, economic conditions, and inequality. We prioritized features that prior research has identified as significant, while also incorporating some factors that are less intuitively linked to poverty. The dataset was then used to train various models aimed at predicting poverty rates across countries. This task is particularly challenging because countries respond differently to the same variables. For instance, GDP growth tends to have a more significant impact on poverty reduction in developing nations compared to developed ones. Additionally, many variables

Information Society 2024, 7–11 October 2024, Ljubljana, Slovenia

https://doi.org/https://doi.org/10.70314/is.2024.sikdd.20

are strongly correlated, making it difficult for linear regression models to capture their relationships accurately.

As previously mentioned, most of the data used in this paper was sourced from *ourworldindata.com* (OWiD), with some additional data coming from *fao.org*—including variables such as foreign direct investment inflows and outflows, and the added value of agriculture, among others. Data on the transatlantic slave trade and colonial rule was obtained from *www.slavevoyages.org*. All datasets were preprocessed before being merged, following a series of steps.

The first preprocessing step involved light modifications, such as removing irrelevant columns, renaming columns, and excluding data from before 1987 and after 2023 due to gaps and incomplete data. Despite increased reporting in recent years, many countries still omit certain indicators, complicating model training. To address this, missing features with more than n data points for a given country were interpolated, with the edges filled using backward fill (bfill) and forward fill (ffill). Those with less than n data points used the mean of the country's income group for the given year as a filler value. The number n was intuitively chosen to be five and the methods bfill and ffill were chosen to prevent the use of unrealistic data. The World Bank classifies countries into income groups: low (less than 1,045 USD), lower-middle (1,046 USD to 4,095 USD), upper-middle (4,096 USD to 12,695 USD), and high income (12,696 USD or more). However, it is important to note that the data generated using the aforementioned methods somewhat reduces overall robustness.

The next step involved generating auxiliary columns, specifically lagged columns and changes in value for relevant parameters. For instance, the row corresponding to Niger in 2013 would also include the GDP per capita for 2012, 2011, and earlier years, in addition to the value for 2013. This approach reflects the fact that poverty trends often manifest in response to changes over time, rather than immediately. The default number of years for lagged data was set to five. Similarly, we incorporated changes in value over the same five-year period to capture more explicit data on unusual events, such as the onset of wars or significant political changes.

Next each primary parameter was also used as an argument for a number of mathematical functions in an effort to see if any correlations aren't linear but perhaps squared, cubed or another elementary function. The functions used were: x^2, x^3, $\ln x$, $\sin x$, $\cos x$, $\tan x$, $\arcsin x$, $\arccos x$, $\arctan x$ to try and capture any elementary nonlinear dependence within the model.

The last step was to create all possible products with the available primary parameters, as creating all possible products with all auxiliary parameters included would have been computationally inefficient. After all these steps were made, the individual columns were fused together. This method of preprocessing increases the possible variables included, making the model even more general and retaining as many rows of data as possible.

The function responsible for preprocessing, generating and merging the data has a few parameters: *basic_parameters_only*, *combinations* and *math*. *basic_parameters_only* determines, if the model will only contain data obtained from various online databases, or if the model should include generated data: the change in value and values for previous years. *combinations* determines, if the model should create all possible combinations with the primary parameters and *math* determines if mathematical columns are included in an attempt to gain a deeper insight into the features' relationships. The parameters are marked with B, C and M. For instance, B+M would mean the file contains the basic parameters in addition to the mathematically derived columns.

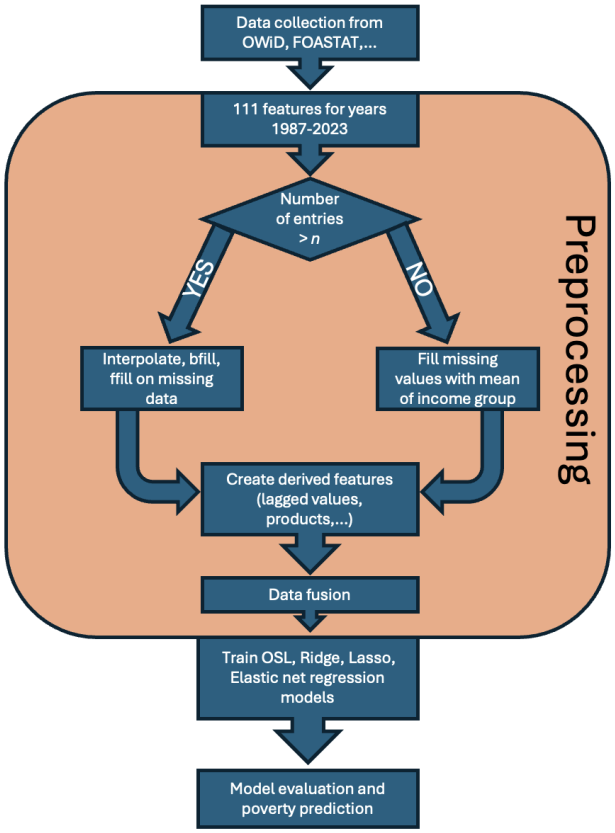

Figure 1: Scheme of adopted methodology

3 Methodology

In order to predict worldwide poverty levels, we have used different linear regression models and compared their accuracies. With this we aimed to ease the interpretability of the models, which is harder to obtain with more complex methods such a neural networks. To perform the research work that is the base of this paper, we have selected ordinary linear regression, lasso regression, ridge regression and elastic net regression as the models to compare. OLS regression struggles with multicollinearity, where predictor variables are highly correlated, leading to unstable estimates of the coefficients. Ridge regression addresses this by adding an L2 regularization term, which penalizes large coefficients and helps to stabilize the estimates in the presence of multicollinearity. By shrinking the coefficients, ridge regression reduces the sensitivity of the model to colinear predictors, ensuring more reliable and generalizable results. Unlike lasso, ridge retains all predictors, making it particularly useful when multicollinearity is a key concern but feature selection is not the goal. We use the implementation of these linear regression algorithms in scikit-learn [9].

The datasets were split into training and test sets using the sklearn function `train_test_split`, with 80% for training and 20% for testing. The training set was used to train four regression variants (`LinearRegression, Lasso, Ridge, ElasticNet`), all with a random state seed of 42. while the test set was used to

determine the mean squared error (MSE) and R^2 value using the functions `mean_squared_error` and `r2_score` from [9], both common metrics used to assess models accuracy. All models except OLS regression also had the data standardized before training. The hyperparameter α for the models was sensibly chosen as 0,1. The results, seen in Table 1 are color coded: red for poor performance, yellow for intermediate, and green for the best. The variation in the number of rows is due to the exclusion of rows with insufficient yearly data, which were dropped when calculating differences from previous years.

After identifying the most successful model, we proceeded to compare its performance between high-income and low-income countries. This comparison aimed to assess how the accuracy and frequency of reported data influence the model's performance. These two income groups were chosen because low-income countries typically report less data with lower accuracy, while high-income countries provide more precise reports. We selected all high- and low-income countries from the dataset that were not used during the model's training. From the 20% of data reserved for evaluation, 444 rows (30%) belonged to high-income countries, and 368 rows (24%) belonged to low-income countries.

We used the trained model to predict poverty levels for these groups and evaluated its performance using the MSE metric to analyze differences between income groups. Additionally, we calculated the maximum error to determine if the average performance was skewed by outliers. A similar evaluation was conducted on the data from Slovenia and Somalia, which were part of the split. Slovenia had 8 rows of data, and Somalia had 6, allowing us to explore how missing data impacts the model's performance, as Somalia had significantly fewer data points overall.

4 Main Results

The file configuration plays a critical role in the model's performance. The results show that C+M, C, and B+C are the best configurations. The C+M file includes all basic features, lagged values, changes in value, mathematical columns, and all possible combinations of basic parameters, totaling 8,236 parameters. Configuration C contains all basic features, combinations, and lagged and difference columns. Lastly, B+C includes only the basic parameters and their combinations. All top-performing models were trained on these datasets.

The results in Table 1 show considerable variation. Models trained with ordinary least squares regression performed poorly, with the best model reaching an RMSE just under 10.15 and an R^2 of 0.50. In contrast, lasso and elastic net regression achieved better results, with RMSEs around 7 and R^2 values close to 0.80. Ridge regression also struggled, except for configuration B+C, which provided the best results with an RMSE of 3.6 and an R^2 of 0.94. However, caution is advised when interpreting models using configuration C+M or C, due to the high number of features relative to the dataset size, which could affect their real-world reliability.

The model weights reveal that only products are present among the top ten most important factors. These products include data on population, population density, agriculture, equality, healthcare, and education. The largest weights show the biggest differences, gradually decreasing in magnitude. The top ten weights range from just over 10 to 7, with the highest weights involving combinations such as population and population density, meadows and pastures with the global peace index, and

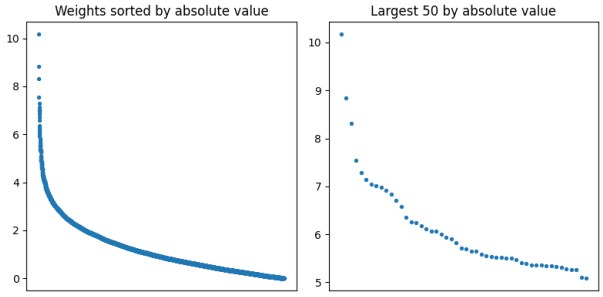

Figure 2: Visual representation of model weights

population with urban and rural population share. Other notable combinations include secondary school completion with women's civil liberties, internet usage with sanitation access, and military spending with wealth distribution. The weights also reflect factors like infant mortality, years colonized, and agricultural employment. Figure 2 further illustrates the decline in the absolute value of these weights.

The model performed better on high-income countries, with an MSE of 6.60, significantly below the overall MSE. In contrast, the MSE for low-income countries was 20.68. The maximum error was also lower for high-income countries (22.1) compared to low-income ones (34.4).

The difference in the model's performance on Slovenia and Somalia was notable. For Slovenia, the MSE was 0.78 with a maximum error of 1.54, far below the overall metrics. Somalia, however, had a much higher MSE of 95.7 and a maximum error of 18.7, likely due to less reliable and extreme poverty data, which skews the model's performance on extreme cases.

5 Discussion

Firstly, the fact that ordinary least squares linear regression couldn't produce an accurate model confirms the fact that the parameters are indeed correlated. This is probably also the reason why the ridge regression model performed the best: ridge regression is used to address the issue of multicollinearity and the features included are mostly strongly correlated, as stated in the introduction. Furthermore, the correlation between parameters is obviously drastically increased by generating all possible products of basic parameters.

Secondly, the impact of mathematical columns needs to be considered. Of the first four models, two have mathematical columns and two don't. Of the eight models generated, three of them perform worse if mathematical data is present, while 5 performed better with mathematical data included. This might indicate some deeper connection, which would be interesting to try and understand. Furthermore, lasso regression handles mathematical columns much better compared to the other models used due to its ability to exclude features.

The impact of product combinations of basic features stands out, with all better-performing models having the *combinations* parameter set to `True`, suggesting deeper relationships between variables. Exploring these connections further, perhaps by training a neural network on the basic parameters and comparing it to linear regression models, could be insightful. If the neural network performs better, further investigation into these correlations would be needed.

Structure	Linear MSE	Linear R^2	Lasso MSE	Lasso R^2	Ridge MSE	Ridge R^2	Elastic net MSE	Elastic net R^2	Shape of X
M	203	0.031	74	0.65	-	-	-	-	(7653, 2131)
None	-	-	109	0.48	163	0.22	108	0.49	(7653, 1221)
C+M	198	0.054	45	0.78	-	-	40	0.81	(7653, 8236)
C	-	-	50	0.76	-	-	45	0.79	(7653, 7326)
B	103	0.50	110	0.47	103	0.50	111	0.46	(7661, 111)
B+C	-	-	48	0.77	13.3	0.94	43	0.79	(7661, 6216)

Table 1: MSE and R-squared values for different regression models and dataset configurations. The presence of B, C or M signals the presence of basic basic parameters only (B), combinations (C) and mathematically (M) derived columns in the dataset. A dash is used to label non-converging models with a negative R-squared value.

The dataset used spans from 1987 to 2023, which is relatively short, given that poverty often has deep historical roots. Although data becomes scarcer in earlier years, those points could still be crucial for improving model accuracy. Moreover, most hyper parameters in this paper were chosen sensibly due to time and computational constraints. Different values for the number of lagged years, years of differences, hyperparameters in the training of models and the minimum number of data points required to interpolate missing data could all lead to interesting discoveries and improvements of the generated models. Our result here shows it is possible to achieve this degree of accuracy, but it doesn't limit what the best model could be. The elastic net, especially, should benefit from such a tuning.

As stated in [11], the recent literature mostly uses the random forest model and, in fact, ordinary linear regression wasn't even in the top ten most common methods. An interesting thing to explore would also be the performance of random forest using the best configuration, B+C. The models may struggle to capture correlations between variables due to differing impacts across countries, as mentioned in the introduction. A potential solution is to split the countries into k groups and train separate models for each group. While this could improve predictions, it raises two challenges: how to split countries without bias and how to ensure enough data for training.

The weights in the model further emphasize the issue of multicollinearity among the parameters, with only product terms emerging as the most influential. However, this does not reveal the true importance of individual parameters, as they may enhance the impact of another factor within the product term. Additional research is needed to better determine the true significance of these parameters and gain a clearer understanding of what drives poverty rates up or down. It can be seen in Figure 2, the models weights occupy a wide range. It is clear that some features are more important, based on their weights and further work is being done to understand which features stand out and why.

The model also performed better in predicting poverty levels in high-income countries compared to low-income countries. This discrepancy can likely be attributed to the fact that high-income countries report more data with greater accuracy, allowing the model to identify underlying patterns more effectively. In contrast, much of the data for low-income countries had to be interpolated, which reduced variability between countries and negatively impacted the model's performance.

6 Conclusion

In this paper, we have shown that a general model exists, based on linear regression methodologies, which can predict poverty with a relatively high accuracy (RMSE of 3.6). This was achieved through testing of numerous linear regression models using open data, with the best model being created by using ridge linear regression trained on data which also included all possible combinations of the basic features included in the dataset. The basic parameters included consist of 111 different parameters describing countries across 36 years. Better models could possibly be generated using more complex methods such as neural nets or random forest, gaining in accuracy but compromising the explainability of the model. The models could also benefit from hyperparameter tuning during the whole process to improve results and find the optimal values. We will be addressing this in further research.

7 Acknowledgements

This research was partially funded by the Future of Life Institute under the project "An AI-driven Observatory Against Poverty", and the European Commission's projects under grant agreement 101135800 (RAIDO) and 101120237 (ELIAS).

References

[1] Gianni Betti, Antonella D'Agostino, and Laura Neri. 2002. Panel regression models for measuring multidimensional poverty dynamics. *Statistical methods and applications*, 11, 359–369.

[2] David Brady. 2019. Theories of the causes of poverty. *Annual Review of Sociology*, 45, 1, 155–175.

[3] Muse A.H. Hassan A.A. and Chesneau C. 2024. Machine learning study using 2020 sdhs data to determine poverty determinants in somalia. *IEEE Transactions on Radiation and Plasma Medical Sciences*, 14, 1, 5956.

[4] Dariush Hayati and Ezatollah Karami. 2005. Typology of causes of poverty: the perception of iranian farmers. *Journal of Economic psychology*, 26, 6, 884–901.

[5] Neal Jean, Marshall Burke, Michael Xie, W Matthew Davis, David B Lobell, and Stefano Ermon. 2016. Combining satellite imagery and machine learning to predict poverty. *Science*, 353, 6301, 790–794.

[6] AH Ng, Abdul Ghani Farinda, Fock Kui Kan, Ai Ling Lim, and Teo Ming Ting. 2013. Poverty: its causes and solutions. *International Journal of Humanities and Social Sciences*, 7, 8, 2471–2479.

[7] Rense Nieuwenhuis, Teresa Munzi, Jörg Neugschwender, Heba Omar, and Flaviana Palmisano. 2019. Gender equality and poverty are intrinsically linked: A contribution to the continued monitoring of selected sustainable development goals. Tech. rep. LIS Working Paper Series.

[8] Shah O. and Tallam K. 2023. Novel machine learning approach for predicting poverty using temperature and remote sensing data in ethiopia. *IEEE Transactions on Radiation and Plasma Medical Sciences*, 5, 6, 2302.14835.

[9] F. Pedregosa et al. 2011. Scikit-learn: machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.

[10] Mubaraq Dele Sulaimon. 2020. Multidimensional poverty and its determinants: empirical evidence from nigeria.

[11] Aziza Usmanova, Ahmed Aziz, Dilshodjon Rakhmonov, and Walid Osamy. 2022. Utilities of artificial intelligence in poverty prediction: a review. *Sustainability*, 14, 21, 14238.

[12] Huang Zixi. 2021. Poverty prediction through machine learning. In *2021 2nd International Conference on E-Commerce and Internet Technology (ECIT)*. IEEE, 314–324.