An Empirical Investigation into the Determinants of Poverty for Brazil (1985-2018)

Stellenbosch
UNIVERSITY
IYUNIVESITHI
UNIVERSITEIT

forward together
sonke siya phambili
saam vorentoe

Department of Economics

Data Science 871

Tiago Baltazar

19776209

July 2021

Contents

1 Introduction

The elimination of extreme poverty was one of the main pillars of the United Nations Millennium Development Goals, and, to a large extent, nations have been successful in eliminating the most extreme forms of deprivation. Whilst significant strides have been made in reducing the levels of absolute income poverty over the last decades, relative income inequality has risen (Alvaredo, & Gasparini, 2013). Brazil is one of the most unequal countries in the world, along with South Africa, and has even been dubbed a 'Belindia' by Beghin (2008); with a large segment of poor coexisting with a small enclave of rich. Since the mid 1990's, however, Brazil has undergone a period of rapid poverty reduction, which Ferreira De Souza (2012) attributes to more effective social policies and a consumer-led economic boom. The goal of this paper is to determine what the most significant determinants of the poverty headcount ratio in Brazil were for the period 1985-2018 by employing a LASSO procedure for variable selection among a potential list of 35 predictors included in the final dataset. The LASSO will be used for three models, specifying different values of the tuning parameter (λ), and the best model of this cohort will be selected by evaluating which one results in the highest prediction accuracy.

2 Literature Review

The use of a LASSO procedure for variable selection has generally been used more in the field of financial economics rather than for developmental studies. Chan-Lau (2017) provides a summary on the use of LASSO regularization in finance, economics, and financial networks. The use of these models in this field have been particularly useful due to the, generally, high dimensionality of financial and economic data, as well as the potential for the predictor variables to be highly collinear.

In the field of development economics, Dutt & Tsetlin (2021) use a LASSO to examine the relative importance of different poverty metrics for predicting development outcomes (schooling, institutional outcome, and p/capita income). The results from fitting their LASSO models suggest that, out of 37 income distribution measures, the poverty headcount ratio is the only relevant factor in predicting the aforementioned outcomes. Afzal, Hersh, & Newhouse (2015) use various variable selection methods to predict the most important variables in predicting poverty in Pakistan and Sri Lanka, they find that the LASSO procedure outperforms both

discretionary and stepwise model selection where the number of potential predictors is large. Parameter selection using LASSO models were found to be more appropriate than regular OLS estimation by Baxter & Hersh (2015) in their study on the determinants of bilateral trade flows between countries. From their LASSO regression, they found that using this method of model selection resulted in fewer significant variables when compared with the null hypothesis rejection methodology (for insignificant predictors) when using OLS. In their article, Zixi (2021) use different machine learning methods in order to predict poverty using a dataset with 59 variables and 12600 observations. A LASSO regression is used to address the problem of multicollinearity in their sample, they find that the LASSO was successful in addressing the multicollinearity problem, and led to an improvement in their forecast accuracy. The LASSO procedure was also used by Ofori (2021) in his study on the drivers of inclusive growth in Sub-Saharan Africa. Here, the author use the LASSO (among other machine learning algorithms) to identify the most relevant predictors driving inclusive growth for a panel of 43 countries in the region, using 97 covariates.

3 Exploratory Analysis

The data used for this analysis was sourced from the World Bank (WB) Poverty and Equity database. This contains various indicators relating to countries income distribution, shared prosperity, poverty rates, and the population. When compiling the sample for Brazil, many variables that could be relevant in predicting the poverty headcount ratio had a large number of missing values. In total, 30 variables had more than 26% of observations missing and were therefore excluded. In order to conduct research, data from the WB World Development Indicators was used to supplement the original sample. This contains statistics about global development and includes variables for topics ranging from health and education, to economic policy and debt. Summary statistics for the variables included in the sample can be found in Tables 9-14 in the appendix.

The poverty headcount ratio measures the proportion of the population living below a specified poverty line. For the purpose of this study, the most stringent poverty line was used, and the evolution of this variable over the sample period can be seen in Figure 1 below, with Figure 2 showing the poverty headcount ratio using different poverty lines.

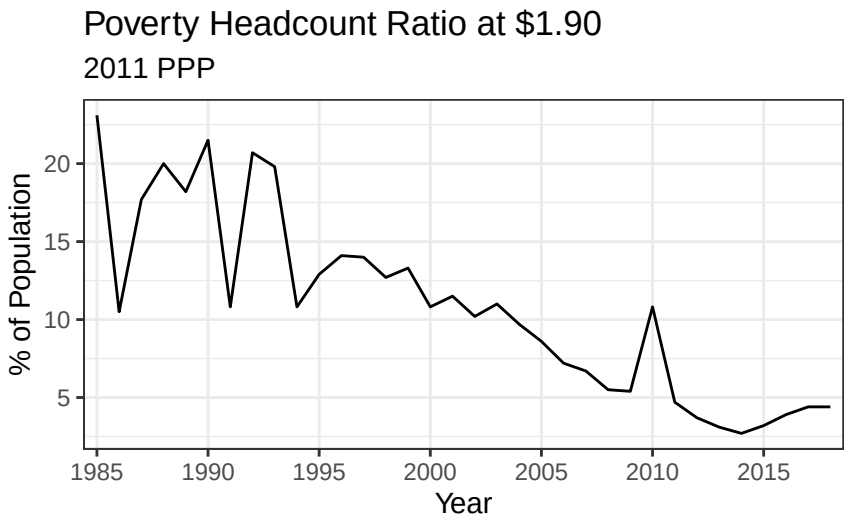

Figure 1: Poverty Headcount Ratio (2011 PPP)

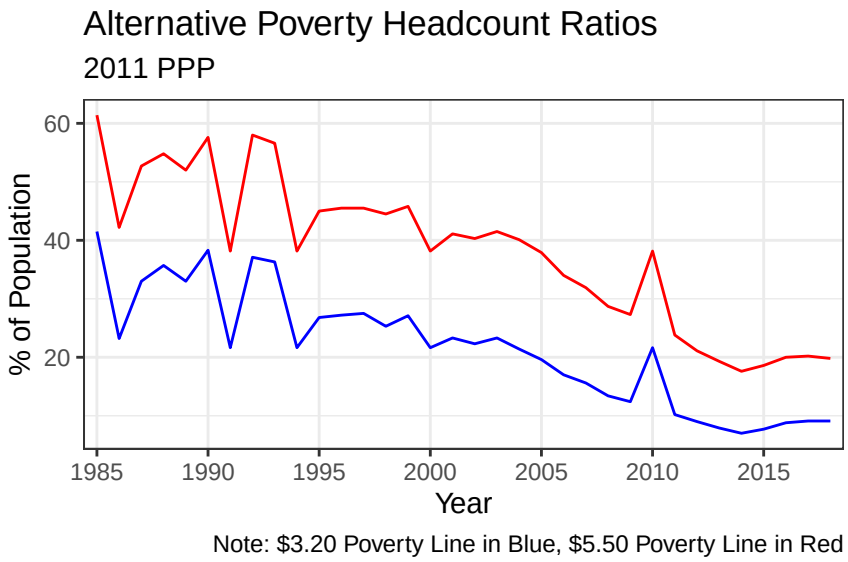

Figure 2: Poverty Headcount Ratio: Alternative Poverty Lines (2011 PPP)

The GINI coefficient was chosen as the measure of income inequality. Figure 3 shows how inequality has evolved over the sample period.

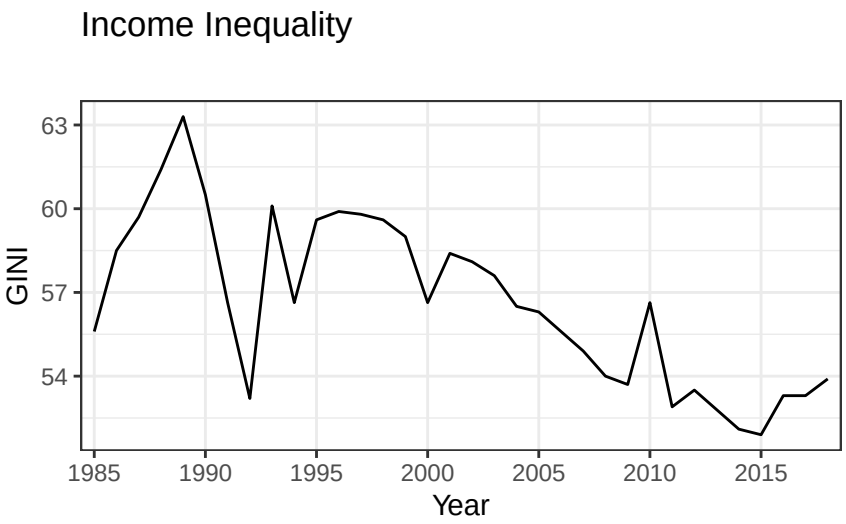

Figure 3: GINI Coefficient

Whilst poverty has generally been decreasing, inequality appears to be more variable and seems to have been increasing (along with poverty) since 2015. Over the sample period, developmental outcomes appear to have generally been improving in Brazil. Life expectancy (Figure 4) has increased significantly since 1985, crude birth rates (Figure 5) have been decreasing, however the discrepancy between male and female employment rates (Figure 6) remains significant. And, in addition to this, labor force participation among the young (Figure 7) has been decreasing, which could be indicative of a lack of employment opportunities for young people.

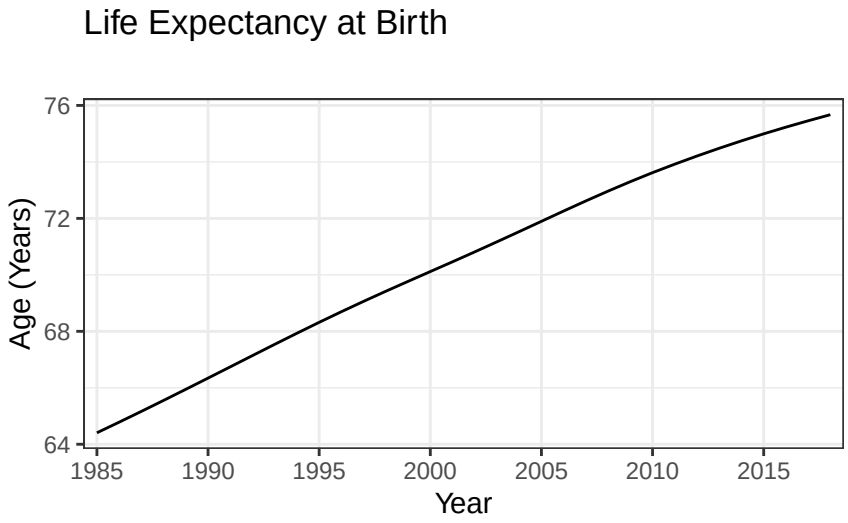

Figure 4: Life Expectancy

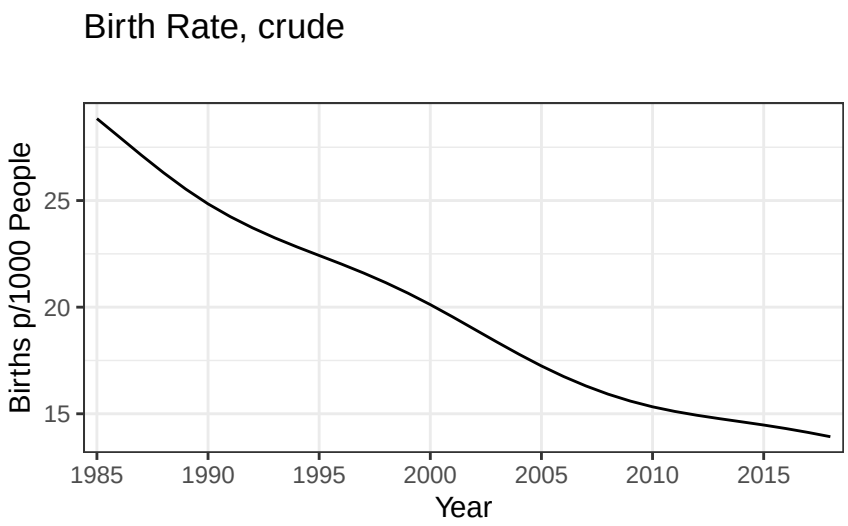

Figure 5: Birth Rate

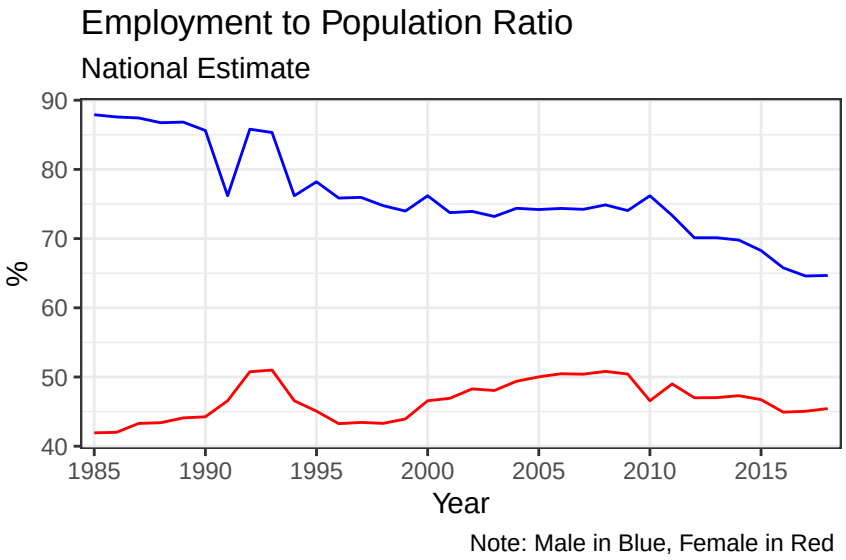

Figure 6: Employment to Population Ratio: Males and Females

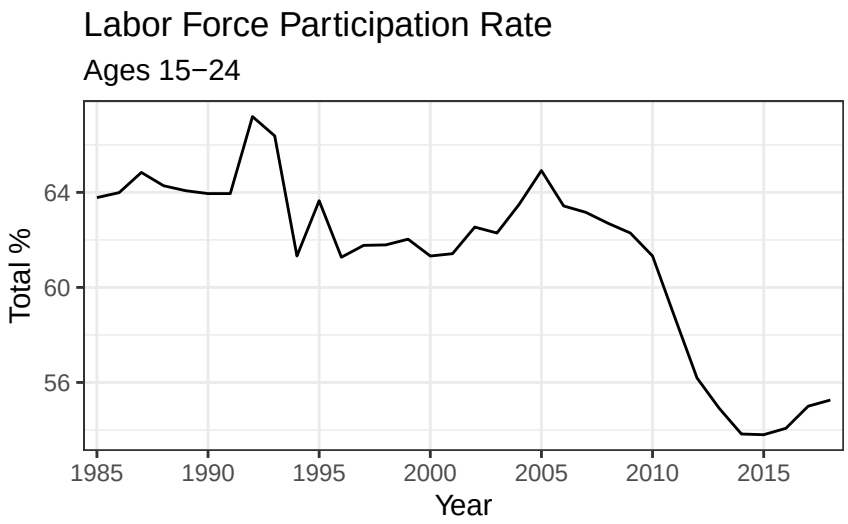

Figure 7: Youth Labor Force Participation

At the same time, the Age-Dependency ratio (Figure 8) has decreased significantly. This ratio relates the number of children (0-14 years) to old people (65+) to the working age population (World Bank, 2021), with a lower value indicating that the working age population is getting larger, relative to the "dependent" population. Finally, from (Figure 9) education expenditure (as a % of GNI) began to increase significantly from 2000 onward, with the expansion in educational expenditure corresponding to the period of decline in the $1.90 poverty headcount ratio.

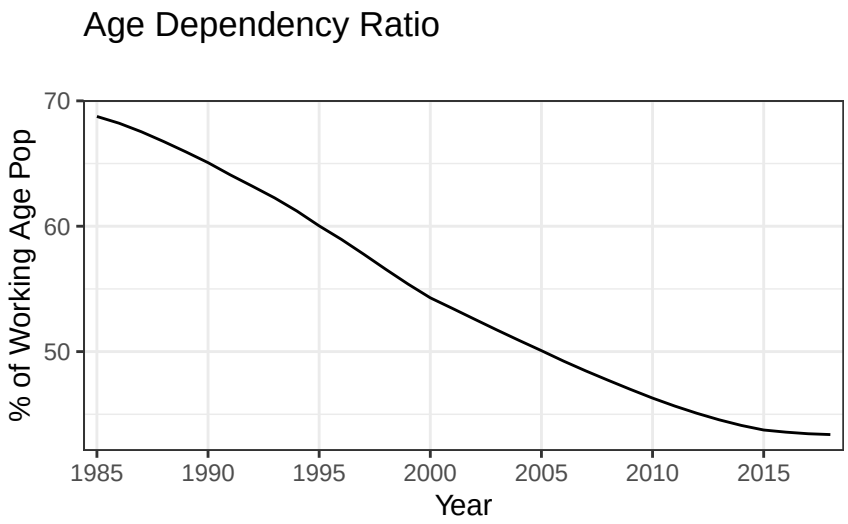

Figure 8: Age Dependency Ratio

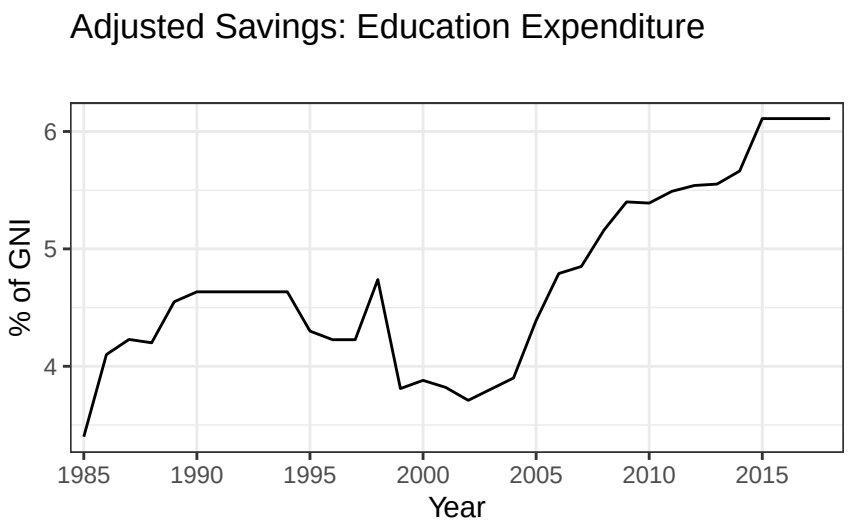

Figure 9: Education Expenditure

Figure 18 displays the correlation plot for all of the variables included in the final dataset. The variables Y1, Y2, and Y3 represent the poverty headcount ratio at $1.90, $3.20, and $5.50 poverty lines respectively. The variable of interest for this study is Y1. From this plot, it can be seen that Y1 is significantly positively correlated with the male employment ratio, the proportion of the population living in rural areas, the birth rate, fertility and mortality rates, and the GINI coefficient. Whereas it is significantly negatively correlated with the urban population, the total life expectancy, CPI inflation, food production, and the GNI.

From this exploratory analysis, one should notice two things. Firstly, the number of predictors is large relative to the sample size of the partitioned dataset. And,in addition, many of the predictors appear to be significantly correlated with eachother. In the presence of high collinearity among regressors, and a large number of potential predictors, relative to the sample size, there will be infinite solutions to the OLS' SSE minimization. Therefore, a method more appropriate to this problem must be used.

4 Empirical Analysis

4.1 Methodology

In order to determine the most relevant predictors of the $1.90 poverty headcount ratio in Brazil, a LASSO regression model will be used for variable selection. LASSO belongs to a class of estimators that involve a penalized regression Varian (2014), and can be used to perform variable selection as it can force some coefficients to equal zero. Regularized regression models are most appropriate when there is a large number of potential covariates, where they are possibly highly correlated, and where the number of predictors is potentially larger than the number of observations (Dutt & Tsetlin, 2021). For the, training, sample data, there are 23 observations for 34 variables (p>n), and, as can be seen in Figure 18 in the appendix, many of the predictors are significantly correlated with eachother. Tibshirani (1996) suggests two reasons why mtehods other than OLS would be preferred in this case. Firstly the prediction accuracy of estimates can be improved by shrinking some coefficients to zero; reducing the variance of the predicted values and thereby improving overall prediction accuracy. And secondly for interpretability, the authors argue that it is often desirable to determine a smaller subset of predictors that have the strongest effect on the target variable (Sparcity-Principle).

Regularization involves adding a component to the objective function which penalizes the inclusion of additional variables. With the, LASSO, objective function now taking the form:

$min(SSE + \lambda \sum_{j=1}^{p} |\beta_j|)$

With λ the tuning parameter determining the severity of the penalty for including additional predictors. Optimizing the above objective function will yield both zero and non-zero variables, as in Afzal, Hersh, & Newhouse (2015), a variable will be considered selected by the LASSO estimator if it is still non-zero after minimizing the objective function. This should yield more parsimonious models where only a subset of variables will exhibit large non-zero coefficients for the target variable. The performance of the LASSO modelsd fit will then be measured by calculating the root mean square error (RMSE), which can be calculated as $RMSE = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n}}$, with the e_i representing the residuals.

4.2 Results

For the purpose of this analysis, three LASSO regression models were fit with varying values of the tuning parameter λ. Firstly, a LASSO model with cross-validation will be fit to determine the optimal value of λ (minimizing the MSE). Then a model will be fit after using a function to determine the λ that yields the lowest RMSE, and finally a LASSO model that gives the smallest cross validation error (minimum λ) will be fit. This will provide three distinct lists of coefficients which were significant in determining the \$1.90 poverty headcount ratio over the sample period, the model yielding the lowest RMSE will thereafter be deemed to have the most credible performance.

4.2.1 Cross-Validation LASSO

The first LASSO model is fit using a 5-fold cross validation. Prior to discussing the dynamics of the LASSO coefficients, a RIDGE regression model was fit to better visualize the distinction between these two different types of regularization methods.

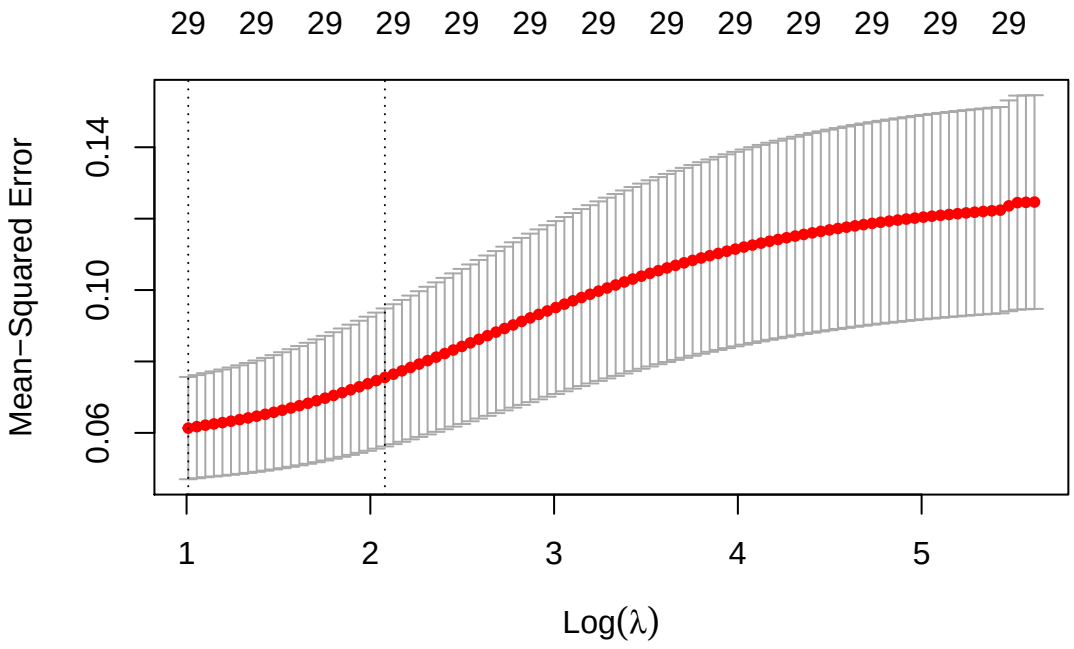

Figure 10: RIDGE Penalty Dynamics

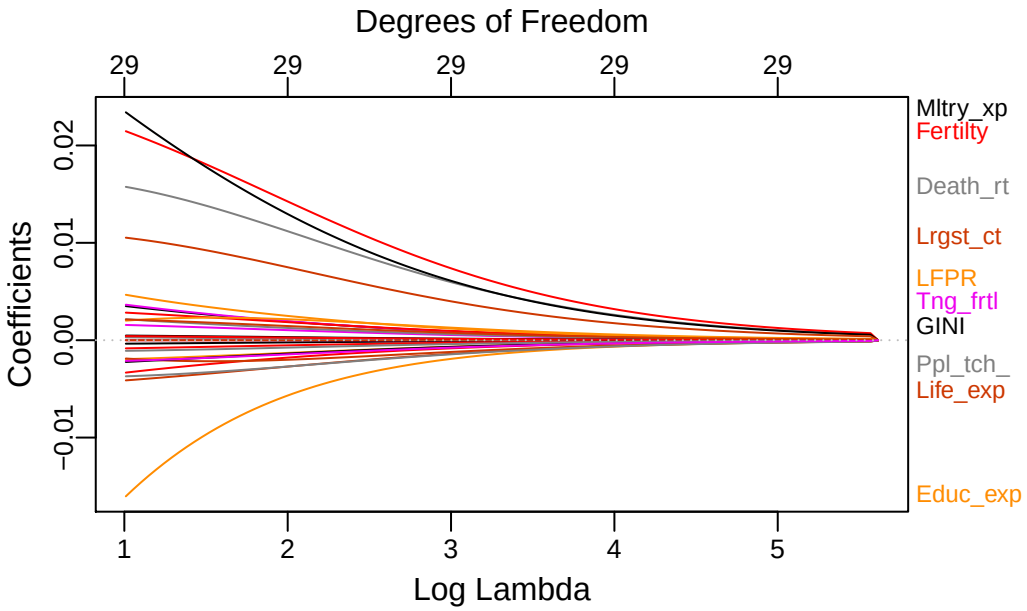

Figure 11: RIDGE Penalty Dynamics

From this, one can see the main distinctions between LASSO and RIDGE models. Firstly, the RIDGE model has 30 non-zero coefficients, from Figure 11, one can see that the coefficient magnitudes get pushed to zero as $\lambda \to \infty$, but not all the way. As can be seen from the variable importance plot for the ridge regression below, many variables are still considered significant predictors of the $1.90 poverty headcount ratio.

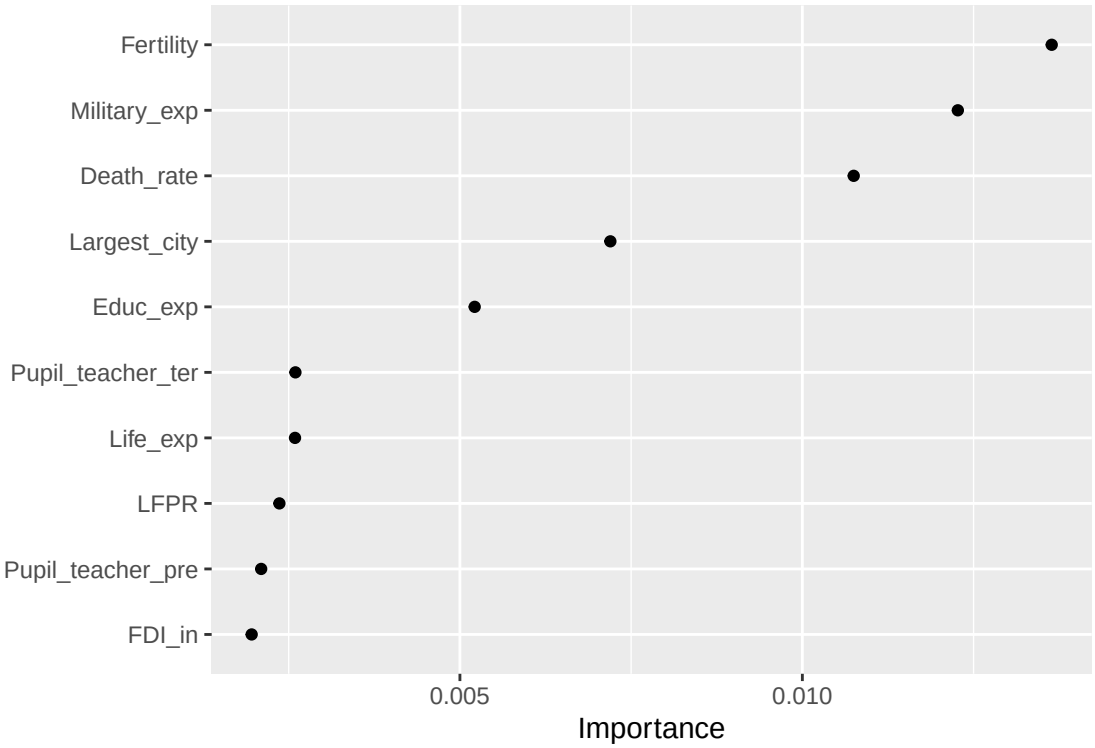

Figure 12: Variable Importance (RIDGE)

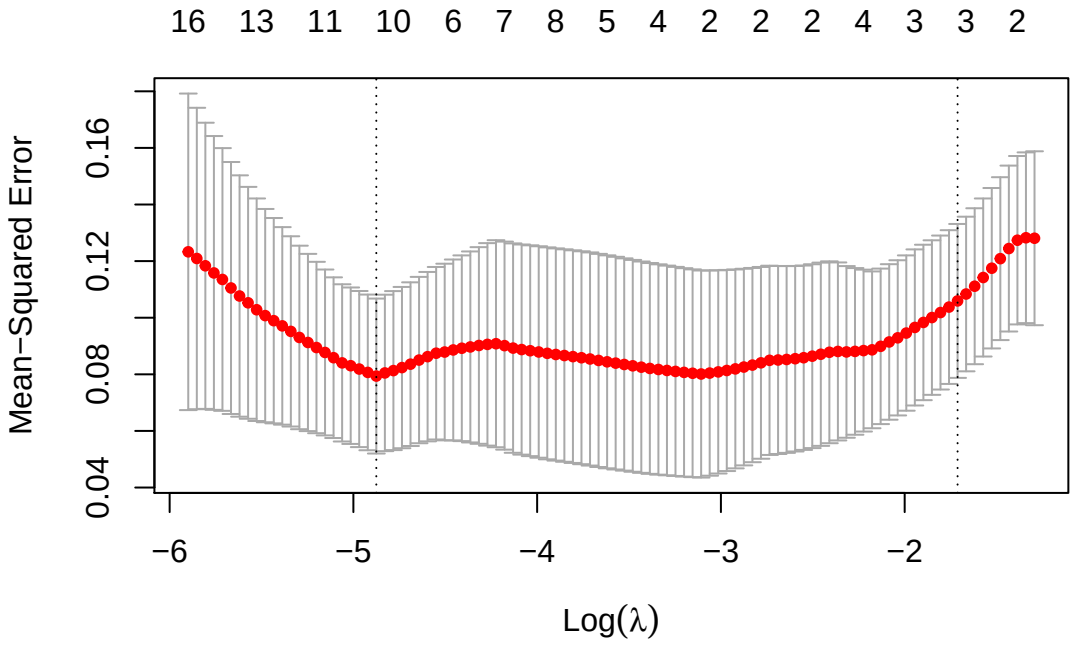

Figure 13: LASSO Penalty Dynamics

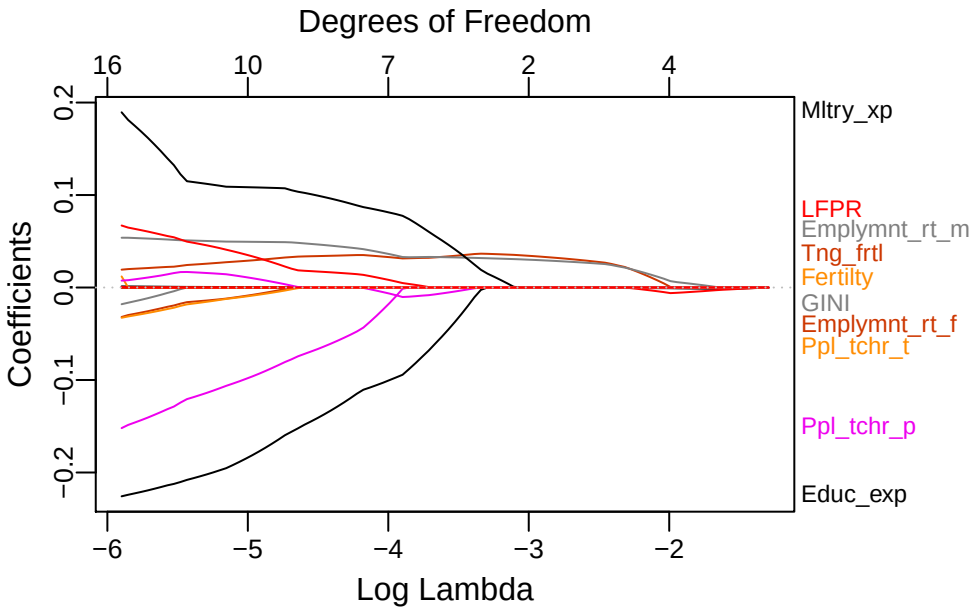

Figure 14: LASSO Penalty Dynamics

From the above figure one can see that the LASSO penalty forced 26 of the coefficients to zero. Now there are only 4 non-zero predictors. The variable importance plot and list of non-zero coefficients are presented below:

Table 1: Coefficient List (1)

var	val
(Intercept)	2.6627720
Food_prod	-0.0034058
Employment_ratio_mal	0.0015436
CPI	-0.0013230

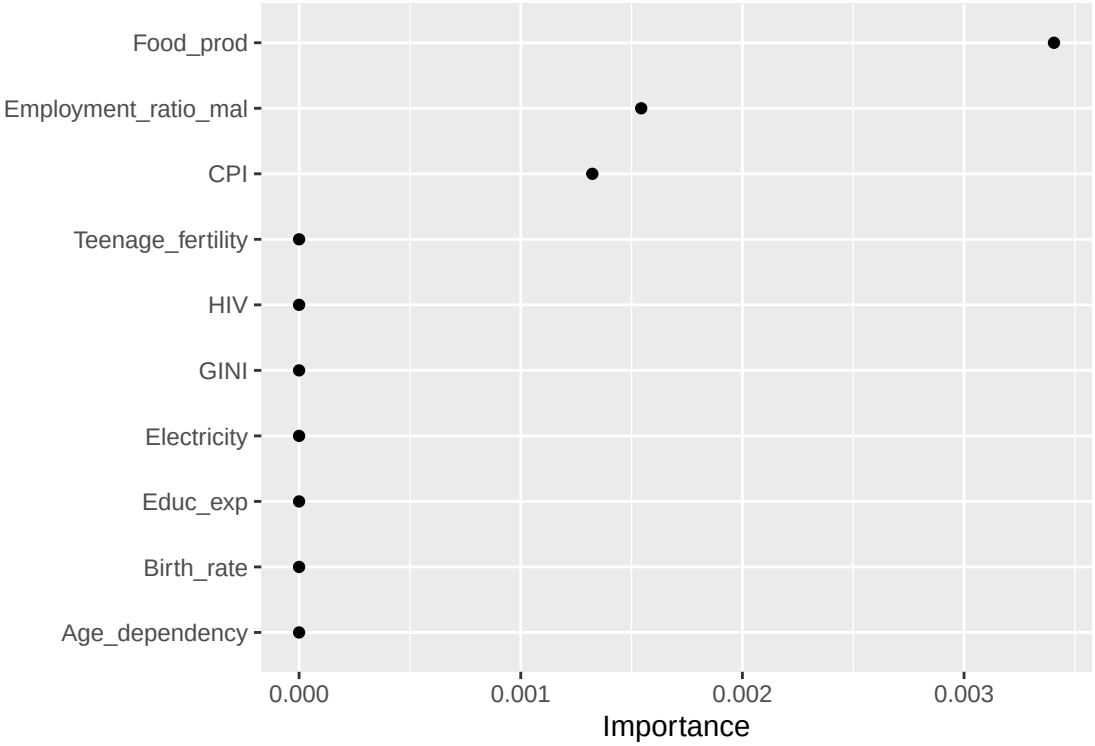

Figure 15: Variable Importance (LASSO 1)

From the cross-validation LASSO model, only the coefficients for the intercept, male employment ratio, food production, and CPI are significantly different from 0. The food production index appears to have the greatest effect on the $1.90 poverty headcount ratio, followed by the male employment ratio, and the CPI. Excluding the intercept term, only 3 of the original 29 variables included in the model were not forced to zero by the LASSO. This model produced an RMSE of 3.8659.

4.2.2 Minimum RMSE

In order to determine the model producing the lowest RMSE, a function was run producing the following output:

Table 2: Optimal RMSE Model

alpha	lambda	RMSE	Rsquared	MAE	RMSESD	RsquaredSD	MAESD
1	0.2390732	3.267434	0.5852761	2.49779	1.69936	0.3450918	1.197031

The model that minimizes the RMSE for estimating the poverty headcount ratio therefore uses an $\alpha = 1$, and $\lambda = 0.2391$ to yield an RMSE of 3.5625 and $R^2 = 0.5715$. Fitting the LASSO model following these specifications yields the coefficient list in Table 3 and plot in Figure 16, with an RMSE of 4.4084.

Table 3: Coefficient List (2)

var	val
(Intercept)	2.6340578
Food_prod	-0.0008815
CPI	-0.0007562

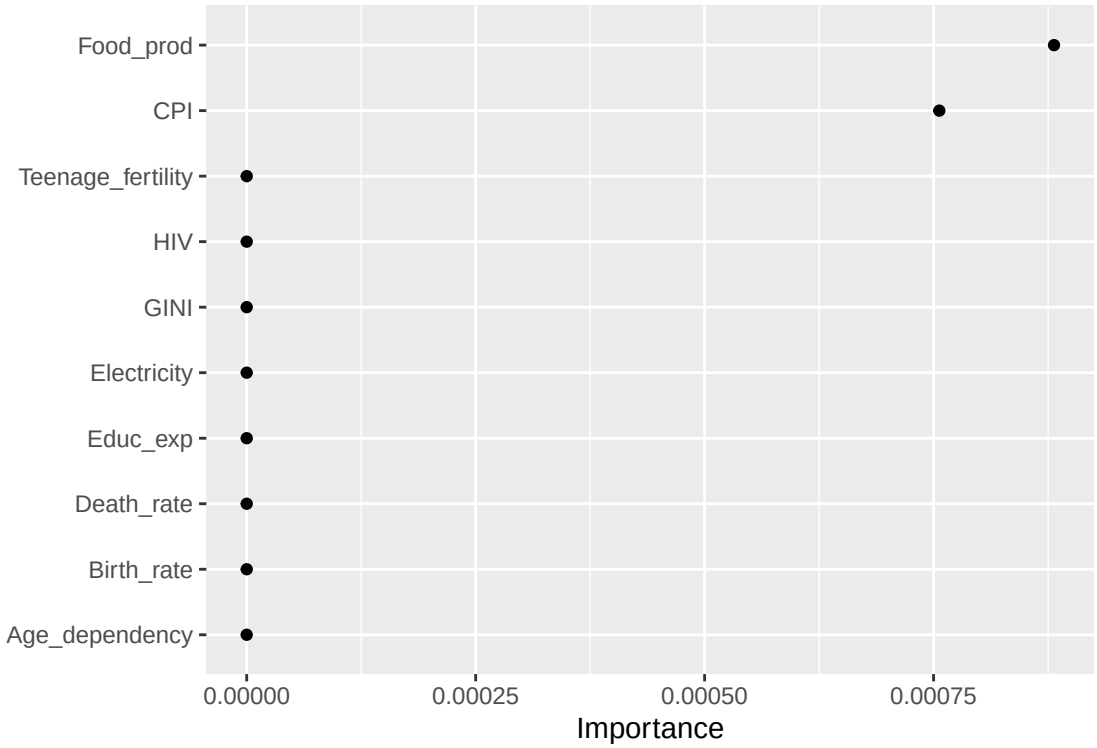

Figure 16: Variable Importance (LASSO 2)

Using the λ value implied by the RMSE minimizing model gives the result that only the food production index and the CPI (inflation) were significant determinants of the $1.90 poverty headcount ratio.

4.2.3 Minimum Lambda

The final model fit to investigate the most important determinants of the poverty headcount ratio in Brazil over the sample period involved specifying the tuning parameter $\lambda = 0.05385$. This implies a less severe penalty relative to the previous model, one can therefore expect that fewer coefficients will be forced to zero by the LASSO.

The list of non-zero coefficients and the variable importance plot can be found below.

Table 4: Coefficient List (3)

var	val
(Intercept)	-2.9020811
Educ_exp	-0.1739682
Military_exp	0.1080122
Pupil_teacher_pre	-0.0907605
Employment_ratio_mal	0.0491918
Teenage_fertility	0.0303611
LFPR	0.0301721
FDI_in	0.0078137
Pupil_teacher_ter	-0.0068447
Employment_ratio_fem	-0.0059772
Credit_GDP	0.0002218

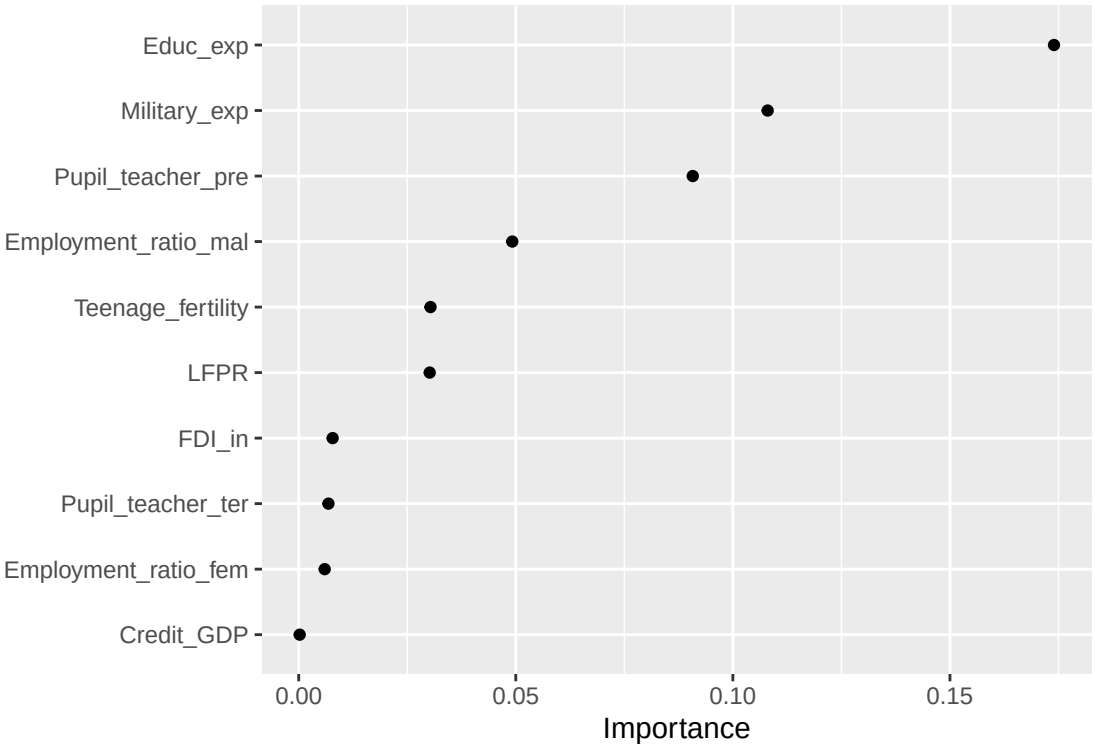

Figure 17: Variable Importance (LASSO 3)

After fitting the LASSO using the minimum λ from the cross-validation procedure, one can see that there are now 11 non-zero coefficients (including the intercept) and 19 coefficients that were forced to zero. This model produces an RMSE=1.6909, the lowest of all three models; implying that the variation in the prediction errors, produced by this model, is minimized. Table 4 gives the list of coefficients in descending order of their (standardized) absolute values. From this, one can see, that education expenditure (% of GNI) is the variable that had the most significant impact on the $1.90 poverty headcount ratio for Brazil over the sample period. Followed by military expenditure (% of GDP), the pupil-teacher ratio at pre-primary level, the male employment ratio (which now has a bigger impact), the adolescent fertility rate, and the youth labor force participation rate. Whilst net inflows of Foreign Direct Investment (% of GDP), the pupil-teacher ratio at tertiary level, the female employment ratio, and the total domestic credit to the private sector (% of GDP) appear to have the least impact on the poverty headcount ratio.

5 Robustness Checks

In this section, the models for the 5-fold cross-validation LASSO, the LASSO using the RMSE minimizing parameters, and the LASSO using the minimum λ were fit on the testing data over the period 2008-2018. The non-zero coefficients for each model will be presented and their performance will be evaluated using the RMSE, and the Mean Absolute Error (MAE), these will then be compared to the results obtained when using the training data.

5.1 Applying Models to Testing Data

5.1.1 Model 1 (Cross-Validation LASSO)

Table 5: Coefficient List Test (1)

var	val
(Intercept)	-8.5386690
Military_exp	0.3776163
GINI	0.1642633
LFPR	0.0122226

For the cross-validation LASSO, there are now 4 non-zero predictors, and 26 coefficients that were forced to zero. Excluding the intercept term, military expenditure (% of GDP), the GINI

coefficient, and the youth LFPR are now the only relevant predictors of the $1.90 poverty headcount ratio for Brazil.

5.1.2 Model 2 (Minimum RMSE)

Table 6: Coefficient List Test (2)

var	val
(Intercept)	-2.776363
GINI	0.079553

Specifying a LASSO ($\alpha = 1$) with, the tuning parameter, $\lambda = 0.2391$ yields the coefficient list above. The GINI coefficient is now the only non-zero coefficient relevant for predicting the poverty headcount ratio.

5.1.3 Model 3 (Minimum Lambda)

Table 7: Coefficient List Test (3)

var	val
(Intercept)	-9.8281808
Military_exp	0.5252077
GINI	0.1786843
LFPR	0.0176360

The LASSO using the minimum lambda (=0.05385), obtained from the cross-validation, yields the same coefficients as that of the first model, however their magnitudes are slightly larger in this case. None of the variables selected by the LASSO models as relevant to predicting the $1.90 poverty headcount ratio in the training set (1985-2007) are relevant predictors when using the test data (2008-2018).

5.2 Alternative Evaluation Metric

As an additional robustness check for the relative performance of the three models, the MAE's were calculated for the respective models using $MAE = \frac{\sum_{i=1}^{n} |e_i|}{n}$. The values of the MAE for for the cross-validation, minimum RMSE, and minimum lambda models are 3.0988, 3.5505, and 1.2185 respectively. Using this alternative metric confirms that the minimum lambda model is the most accurate, yielding the lowest prediction errors and hence best model performance. The RMSE and MAE values for the three models fit using the testing data are presented below:

Model	RMSE	MAE
1	0.8933	0.484
2	1.7207	1.0758
3	0.6878	0.4601

Even when using the testing data set, model 3 (with $\alpha = 1$ and $\lambda = 0.05385$) has the best performance, in terms of prediction accuracy, relative to the other models. Although the relative performance of each of the models, measured by the RMSE and MAE, appears to be the same when these are applied to the hold-out (testing) data, the lists of non-zero predictors are significantly different. This implies a poor out-of-sample performance of the models specified in Section 4. This could be due to a misspecification in the original model, or as a result of vastly different country characteristics for the 2008-2018 period, relative to the training set period, which would cause the determinants of the poverty headcount ratio to change significantly, with the youth LFPR being the only variable that is still significant in both the training and testing data for the best performing (minimum λ) model.

6 Conclusion

This paper attempted to investigate the determinants of the $1.90 poverty headcount ratio for Brazil over the period 1985-2018 by using LASSO regressions to perform variable selection, and determine the best set of, non-zero, predictors. Based on the training data, the model which minimizes the prediction errors, using both RMSE and MAE, specifies a tuning parameter (λ) value of 0.05385, and selects the list of variables seen in Table 4. According to this model, the

most significant determinant of poverty over the training period (1985-2007) was the level of government education expenditure (% of GNI). Fitting the same model to the hold-out data, covering the period 2008-2018, still resulted in the best out-of-sample relative to the other models fit on the testing data, but led to a different list of non-zero predictors. The only variable that remained significant across both models was the youth LFPR. The robust determinants of the poverty headcount ratio for Brazil were determined by using a LASSO procedure to perform variable selection. Drawing any, significant, inference from these variables regarding the magnitude of their impacts on the poverty headcount ratio is not possible using the methodology followed in this study. Belloni & Chernozhukov (2013) suggest a post-LASSO procedure in order to conduct inference which entails two steps. Firstly applying a LASSO regularization to determine the list of non-zero variables, and then estimating coefficients on the remaining variables via OLS and using only the variables with non-zero coefficients. This, however, lies beyond the scope of this investigation.

7 Bibliography

- Afzal, M., Hersh, J., & Newhouse, D. (2015). Building a better model: Variable selection to predict poverty in Pakistan and Sri Lanka. *World Bank Research Working Paper*.
- Alvaredo, Facundo; Gasparini, Leonardo (2013) : Recent Trends in Inequality and Poverty in Developing Countries, Documento de Trabajo, No. 151, Universidad Nacional de La Plata, Centro de Estudios Distributivos, Laborales y Sociales (CEDLAS), La Plata
- Baxter, M., & Hersh, J. (2015, May). Robust determinants of bilateral trade. Working paper
- Beghin, N., 2008. Notes on Inequality and Poverty in Brazil: Current Situation and Challenges. s.l.:Oxfam.
- Belloni, A. & Chernozhukov, V., 2013. Least squares after model selection in high-dimensional sparse models. *Bernoulli*, 19(2), pp. 521-547.
- Chan-Lau, J. A., 2017. Lasso Regressions and Forecasting Models in Applied Stress Testin. s.l.:International Monetary Fund.
- Dutt, P. & Tsetlin, I., 2021. Income Distribution and Economic Development: Insights from Machine Learning. *Economics and Politics*, 33(1), pp. 1-36.
- Ferreira De Souza, P. H., 2012. Poverty, inequality and social policies in Brazil, 1995-2009. s.l.:Institute for Applied Economic Research.
- Friedman J, Hastie T, Tibshirani R (2010). "Regularization Paths for Generalized Linear Models via Coordinate Descent." *Journal of Statistical Software*, 33(1), 1–22. https://www.jstatsoft.org/v33/i01/.
- Greenwell BM, Boehmke BC (2020). "Variable Importance Plots—An Introduction to the vip Package." *The R Journal*, 12(1), 343–366. https://doi.org/10.32614/RJ-2020-013.
- Ofori, Isaac Kwesi (2021) : Catching The Drivers of Inclusive Growth in Sub-Saharan Africa: An Application of Machine Learning, EXCAS Working Paper, No. 21/044, ZBW - Leibniz Information Centre for Economics, Kiel, Hamburg

- Tibshirani, R., 1996. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1), pp. 267-288.
- Varian, H. R., 2014. Big Data: New Tricks for Econometrics. *Journal of Economic Perspectives*, 28(2), pp. 3-28.
- Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, Grolemund G, Hayes A, Henry L, Hester J, Kuhn M, Pedersen TL, Miller E, Bache SM, Müller K, Ooms J, Robinson D, Seidel DP, Spinu V, Takahashi K, Vaughan D, Wilke C, Woo K, Yutani H (2019). "Welcome to the tidyverse." *Journal of Open Source Software*, 4(43), 1686. doi: 10.21105/joss.01686.
- World Bank, 2021. World Development Indicators. [Online] Available at: https://datacatalog.worldbank.org/dataset/world-development-indicators [Accessed June 2021].
- World Bank, 2021. Poverty and Equity Database. [Online] Available at: https://datacatalog.worldbank.org/dataset/poverty-and-equity-database [Accessed June 2021].
- Zixi, H., 2021. Poverty Prediction Through Machine Learning. Changshu, China, International Conference on E-Commerce and Internet Technology.

8 Appendix

8.1 Descriptive Statistics

Table 9: Descriptive Statistics (1)

X	Y1	Y2	Y3	GINI	Electricity	Educ_exp	Teenage_fertility
Min	2.70	7.0	17.6	51.90	87.50	3.400	57.90
1st Q	5.43	12.7	27.7	53.80	94.30	4.210	67.30
Median	10.80	21.6	39.1	56.60	96.30	4.630	80.90
Mean	10.80	21.6	38.2	56.60	95.90	4.730	75.40
3rd Q	13.80	27.2	45.5	59.50	99.00	5.400	82.90
Max	23.10	41.5	61.4	63.30	100.00	6.110	84.30
Std. Dev	5.93	10.0	13.0	3.02	3.48	0.779	9.15

Table 10: Descriptive Statistics (2)

X	Age_Dependency	Birth_rate	CPI	Death_rate	Debt_service	HIV
Min	43.40	13.90	0.00	6.020	8.22e+09	35000
1st Q	46.50	15.40	6.28	6.090	1.62e+10	42300
Median	53.00	19.30	58.50	6.270	5.08e+10	45100
Mean	54.30	19.70	63.60	6.440	4.62e+10	45100
3rd Q	62.00	23.20	98.80	6.610	6.15e+10	47800
Max	68.80	28.80	161.00	7.660	1.18e+11	51000
Std. Dev	8.57	4.57	51.40	0.463	2.92e+10	3790

Table 11: Descriptive Statistics (3)

X	Employment_ratio_fem	Employment_ratio_mal	Fertility	Food_prod	FDI_in
Min	41.90	64.60	1.730	33.9	0.129
1st Q	44.10	73.50	1.800	43.8	0.635
Median	46.60	74.60	2.200	61.2	2.470
Mean	46.60	76.20	2.310	66.1	2.310
3rd Q	48.80	77.70	2.680	87.9	3.670
Max	51.00	87.90	3.470	108.0	5.030
Std. Dev	2.77	6.73	0.532	24.2	1.560

Table 12: Descriptive Statistics (4)

X	Credit_GDP	LFPR	Life_exp	Military_exp	Mortality_fem	Mortality_mal
Min	27.7	53.80	64.40	1.220	91.4	189.0
1st Q	36.7	61.30	67.60	1.430	103.0	208.0
Median	46.6	62.30	70.60	1.530	120.0	240.0
Mean	54.0	61.30	70.50	1.650	126.0	243.0
3rd Q	62.4	63.90	73.50	1.830	148.0	281.0
Max	134.0	67.20	75.70	2.690	174.0	295.0
Std. Dev	25.7	3.77	3.49	0.319	26.5	37.9

Table 13: Descriptive Statistics (5)

X	IR	GNI	Homicide	Largest_city	Rural_pop	Urban_pop
Min	6.87	1.01e+12	16.80	11.900	13.40	69.90
1st Q	11.20	1.28e+12	22.30	11.900	15.70	76.40
Median	18.50	1.54e+12	23.80	11.900	18.30	81.70
Mean	831.00	1.66e+12	23.80	12.400	19.80	80.20
3rd Q	248.00	2.11e+12	25.30	12.900	23.60	84.30
Max	9390.00	2.38e+12	30.80	14.200	30.10	86.60
Std. Dev	2070.00	4.47e+11	3.25	0.746	5.04	5.04

Table 14: Descriptive Statistics (6)

X	Pupil_teacher_pre	Pupil_teacher_ter	LFPR_fem_to_mal
Min	16.60	11.40	48.30
1st Q	20.20	12.20	59.50
Median	20.20	15.10	65.30
Mean	20.20	15.10	63.80
3rd Q	20.50	17.40	70.50
Max	24.20	20.50	73.00
Std. Dev	1.88	2.88	8.02

8.2 Correlations

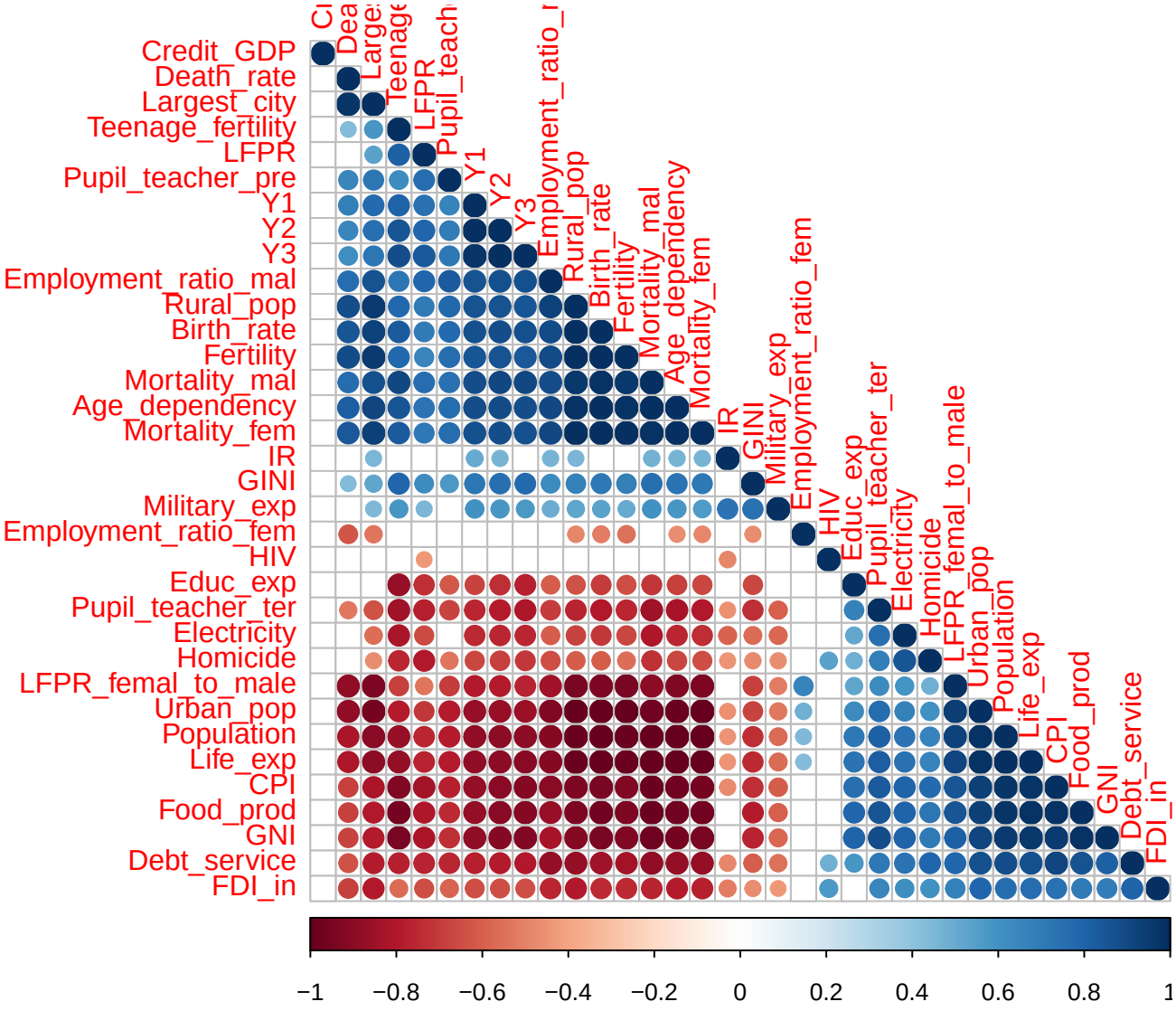

Figure 18: Correlation Plot (excluding insignificant)

8.3 Optimal RMSE model

```
glmnet

23 samples
29 predictors

No pre-processing
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 19, 18, 18, 18, 19
Resampling results across tuning parameters:

  alpha  lambda     RMSE      Rsquared   MAE
  0.10   0.2390732  4.065218  0.5149462  3.045389
  0.10   0.7560159  4.011972  0.4827998  3.043811
  0.10   2.3907322  3.662336  0.5095076  2.816396
  0.55   0.2390732  3.562534  0.5715391  2.746192
  0.55   0.7560159  3.387248  0.5655313  2.647624
  0.55   2.3907322  3.493706  0.5918523  2.844507
  1.00   0.2390732  3.267434  0.5852761  2.497790
  1.00   0.7560159  3.292453  0.5811922  2.588158
  1.00   2.3907322  3.809604  0.6118617  3.241214

RMSE was used to select the optimal model using the smallest value.
The final values used for the model were alpha = 1 and lambda = 0.2390732.
```

8.4 Predictors for Testing Data

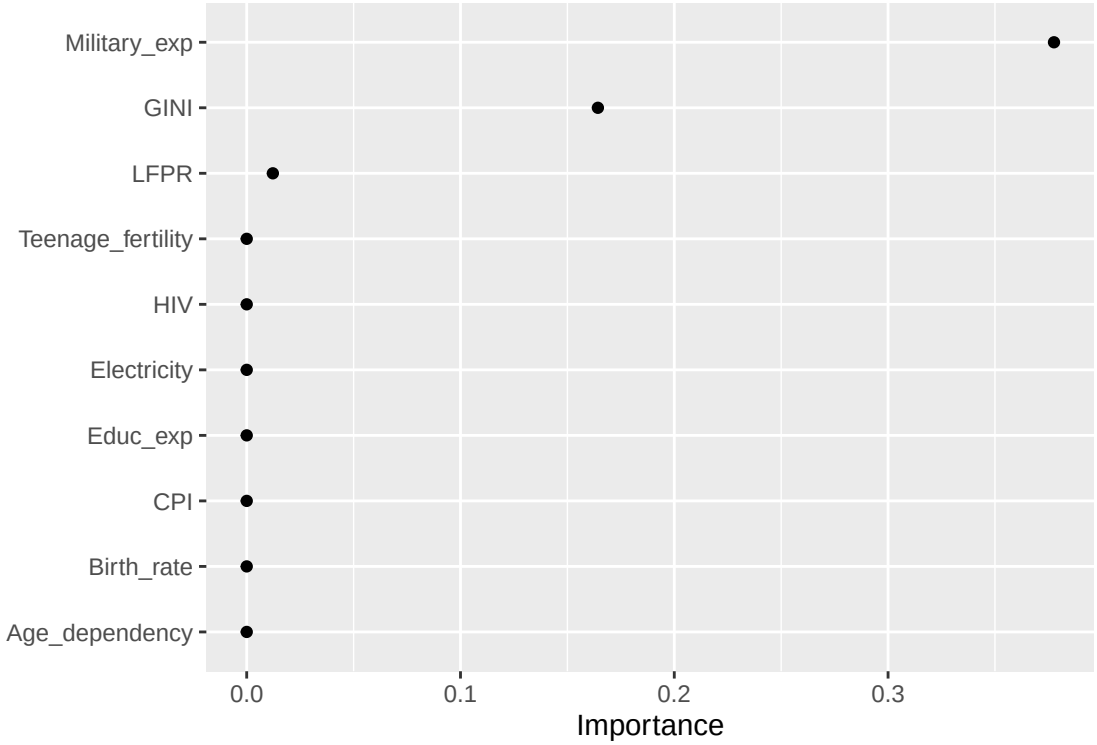

Figure 19: Variable Importance Test (LASSO 1)

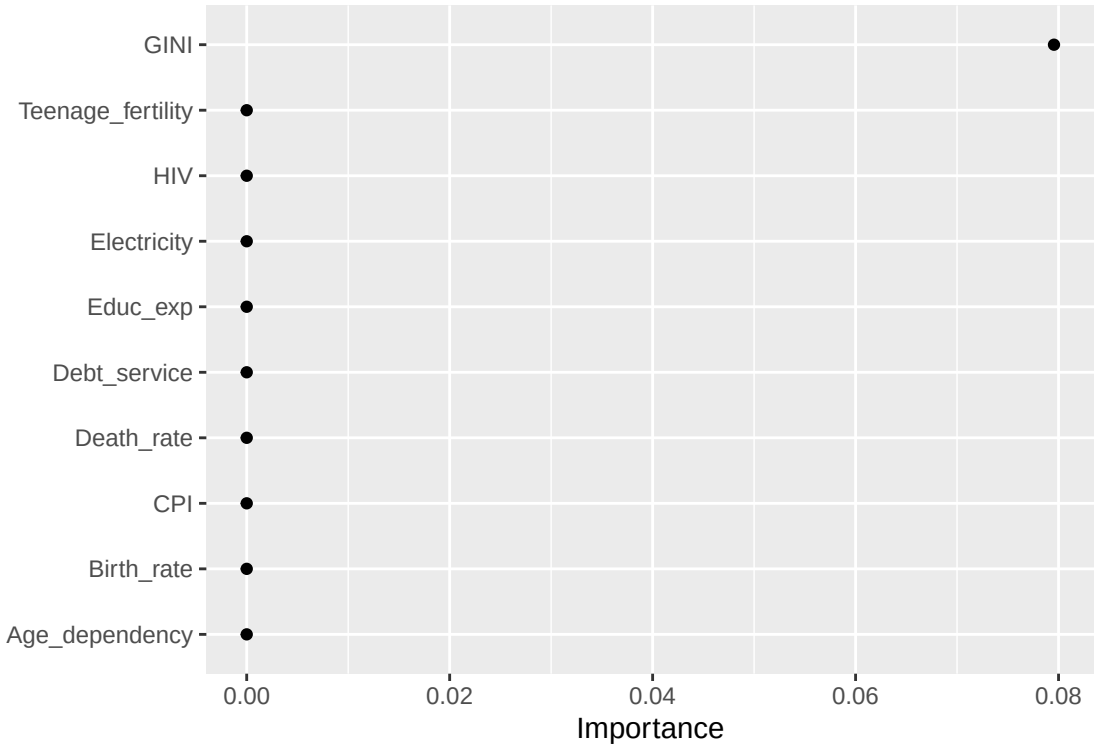

Figure 20: Variable Importance Test (LASSO 2)

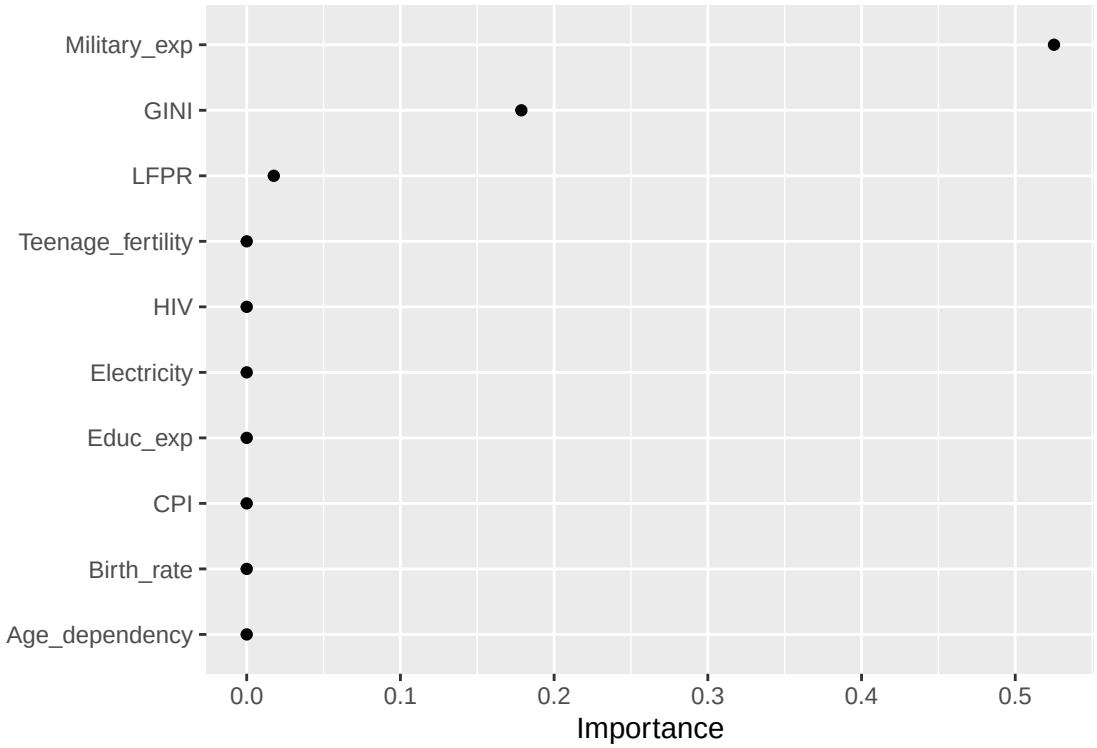

Figure 21: Variable Importance Test (LASSO 3)