Predicting Household Electricity Expenditure in South Africa using Lasso Regularization

Robert Toto (RVT245)

5/13/2021

Overview

The electrification of developing country households can expand education, reduce air pollution, and bring greater economic opportunity and income to low-income communities. In South Africa, over 90% of households have access to electricity—-the highest electrification rate in Sub-Saharan Africa-—but grid reliability is greatly impaired by chronic inability to meet electricity demand with supply. The government-owned grid-operator, ESKOM, must regularly shut off household electricity (through a "load-shedding" program) in order to balance limited electricity supply with electricity demand ("load"). Should grid load exceed grid capacity, South Africa could experience protracted blackouts, subjecting residents to months without power. As South Africa works to improve its electricity generation infrastructure (e.g., building more plants and improving grid hardware), ESKOM must use all available grid consumer data to minimize necessary load-shedding. To improve ESKOM's load forecasts, this analysis uses electricity spending and demographic data in South Africa to develop a model to predict electricity spending. More accurate spending forecasts will enable ESKOM to reduce load-shedding, which, in turn, improves household economic welfare and health.

Figure 1 shows the breakdown of monthly electricity expenditure per household by house size, income decile, and whether the household is located in a rural or urban area. Moderately-sized homes consume more electricity than very large homes, and urban homes consume generally greater electricity than rural homes. The households spending the most on electricity are typically in the highest income deciles, but there are some very poor households consuming up to 1,800 ZAR of electricity per month in rural areas—-well above the average expenditure. Incorporating home size, location, and income decile into the electricity spending predictions will add valuable information to ESKOM's load forecast. Figures 2 and 3 expose South Africa's large income inequality, which is an important aspect of electricity consumption. Figure 2 shows that households in the bottom 60% of the income scale spend less than 200 ZAR per month on electricity, while the top 10% of earners spend up to 575 ZAR on average per month, with urban households generally spending more than rural households. Figure 3 further shows that income inequality in South Africa is extreme: the lowest income decile earns less than 1000 ZAR per month, while the top decile earns nearly 30,000 ZAR on average. Furthermore, the lowest 70% of the income scale earns less than 7,500 ZAR each month.

The consequential financial burden on low-income households is shown in Figure 4—-an extremely important visualization. The lowest income decile spends about 15% of its income on electricity while the highest decile spends a negligible 1-2% of its income the same basic necessity. At the level of subsistence-living experienced by the lowest 10% of earners in South Africa, giving up 15% of wages to a basic utility imposes a massive economic burden. Nonetheless, as shown in Figure 5, most households own traditional electrical appliances. For instance, nearly 5,000 rural households in the sample own a television, compared to only about 1,270 rural households that do not. This indicates that even in rural locations connected to the grid, there is load demand related to modern electrical appliances. Interestingly, rural households with higher earnings own less televisions than low-income rural households. Given these consumption dynamics tied to the country's unique rural-urban outlay and large income inequality, it is essential to make use of the variation stored in these variables when creating an electricity spending prediction model.

Figure 1: Electricity Spending vs House Size

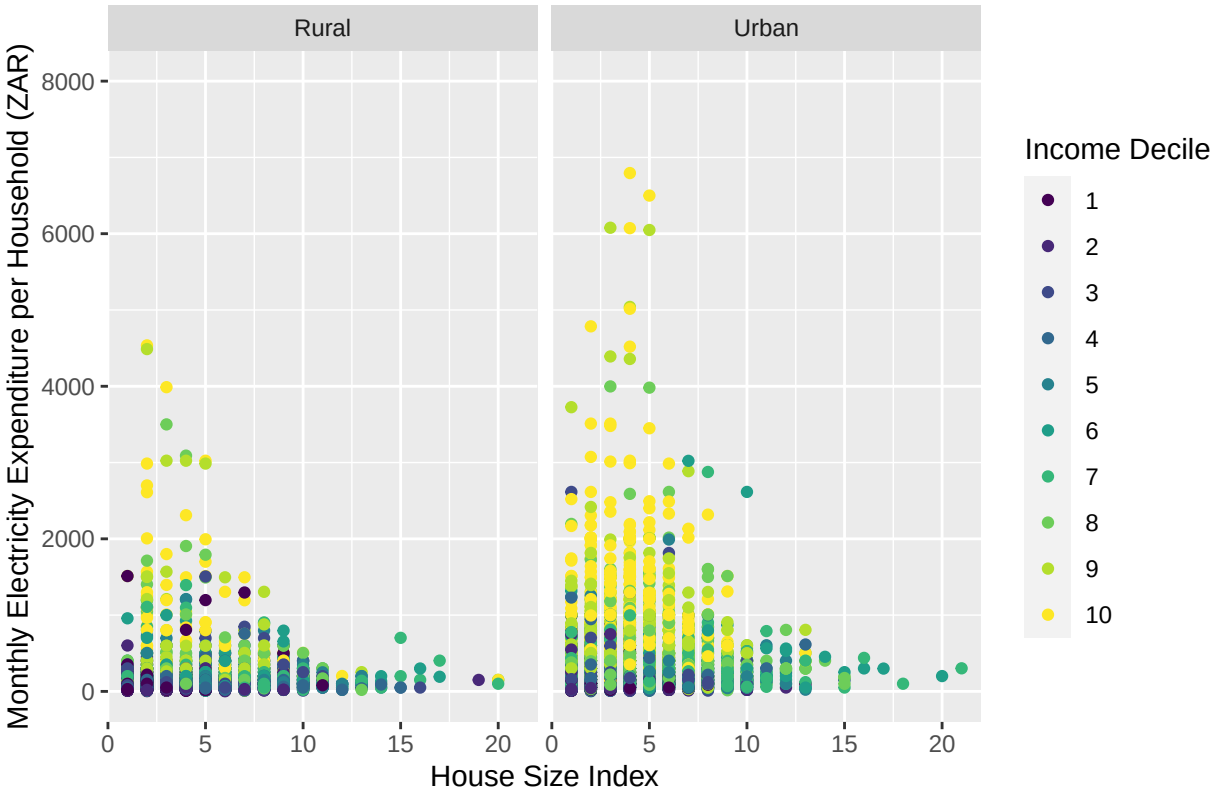


Figure 2: Electricity Spending vs Income Decile by Urban/Rural

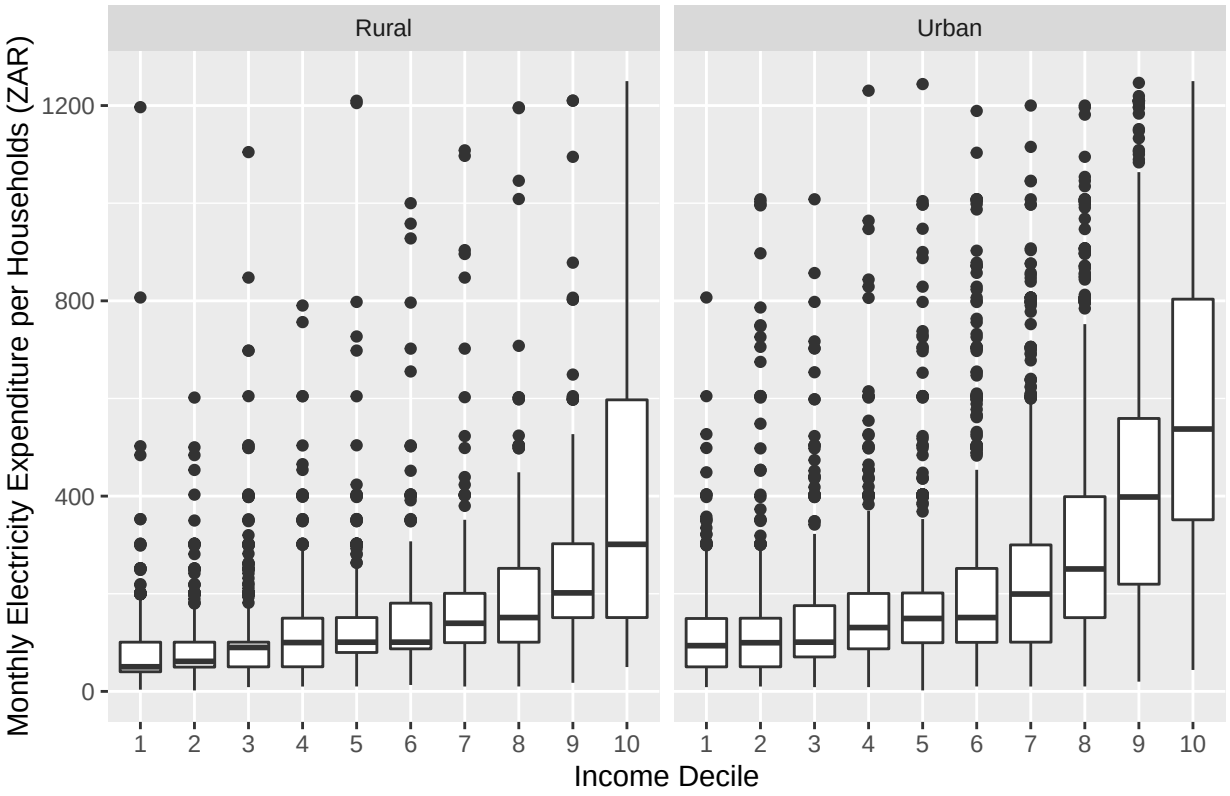


Figure 3: Income Inequality in South Africa

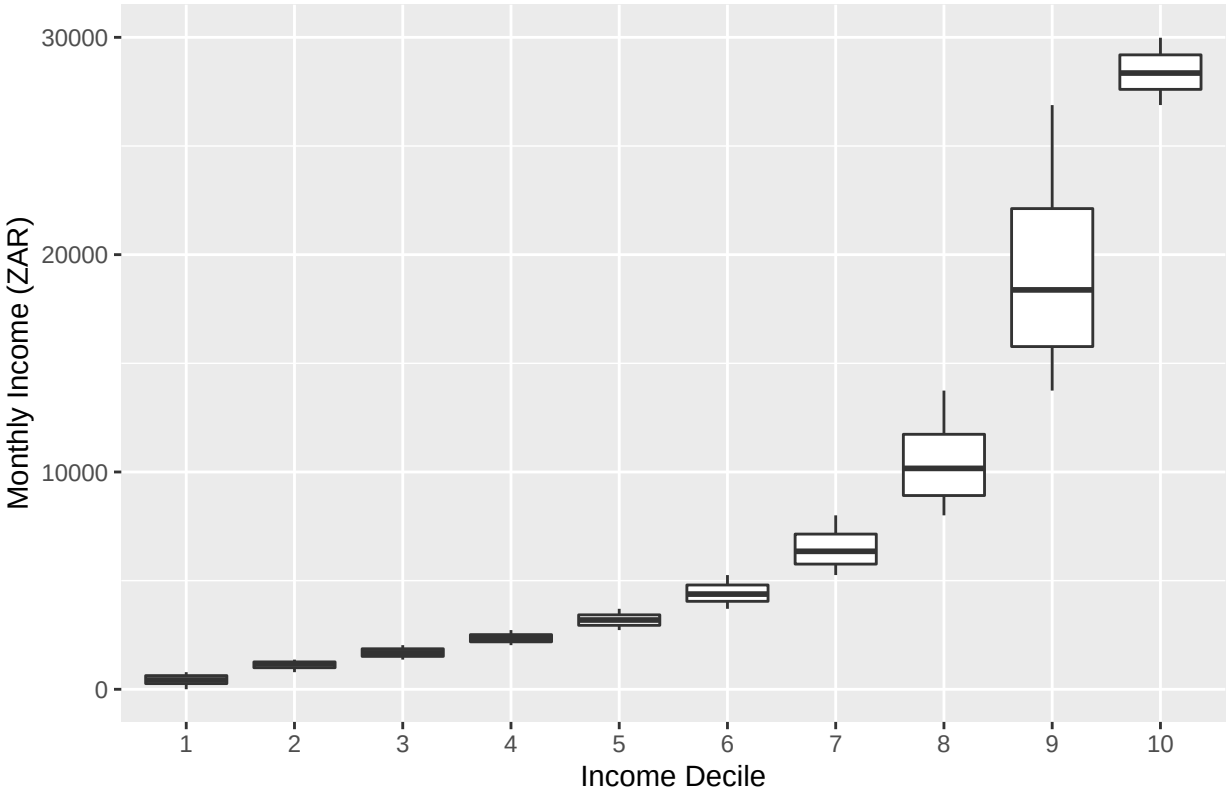


Figure 4: Portion of Income Spent on Electricity

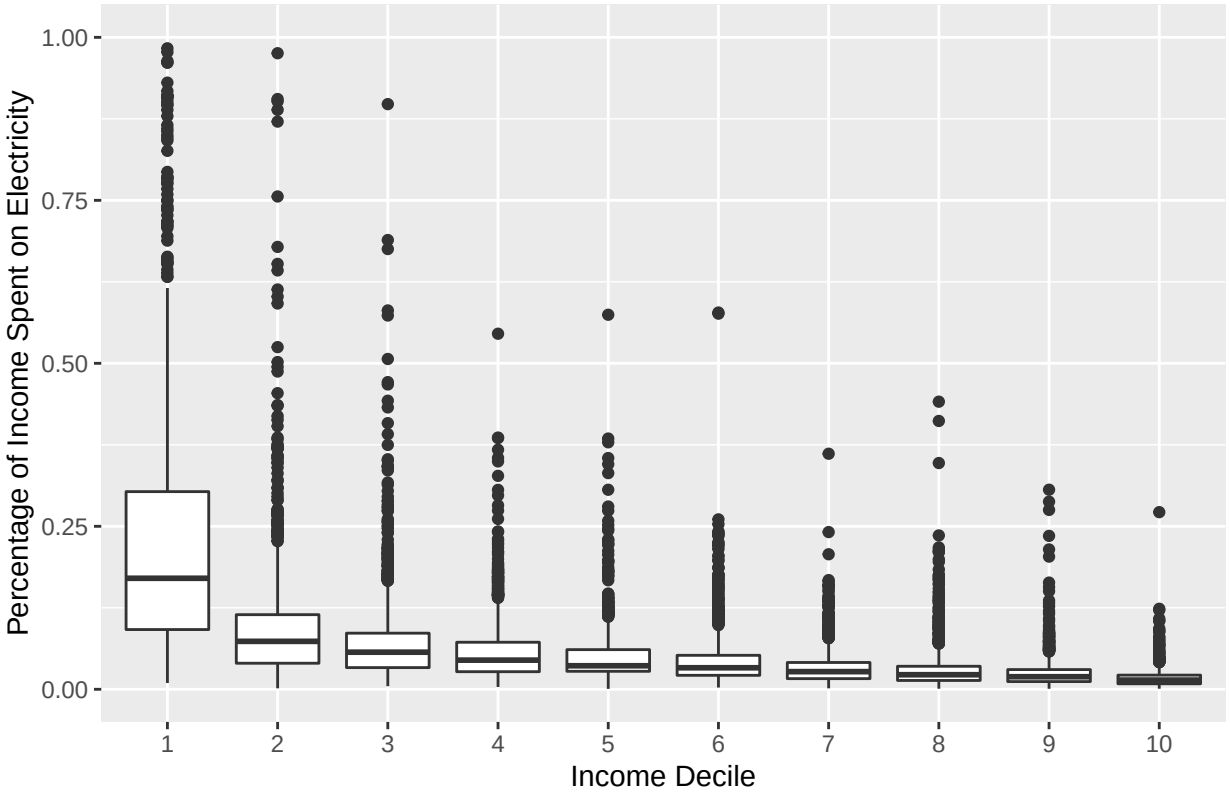
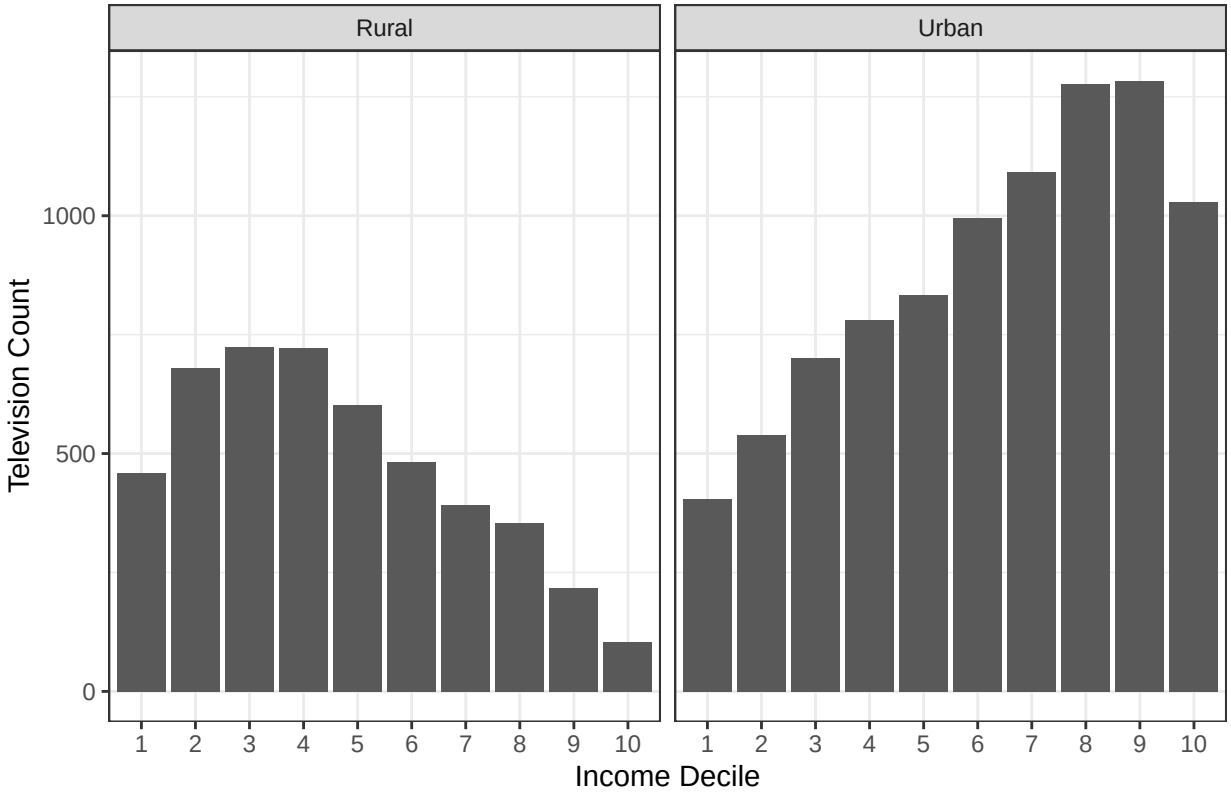


Figure 5: Television Ownership vs Income Decile by Urban/Rural



Data and Model

Data

The electricity spending and demographic data is a combination of two underlying datasets: (1) the 2010/2011 South African Income and Expenditure Survey (SA IES) conducted by South Africa's Department of Statistics and (2) electricity prices from The National Energy Regulator of South Africa (NERSA). The SA IES is a cross-sectional survey of 25,328 households in South Africa, containing demographic data on each household. When merged, the two datasets produce information on a common set of 16,851 households. The final dataset contains electricity spending, energy alternatives (e.g., kerosene or wood), demographics, household characteristics (including electrical appliances), income, and electricity price data for each household. Given that the dataset contains both urban and rural households, many of the energy alternative variables are zero for the urban households, resulting in a highly sparse covariate matrix. Overall, there are 230 covariates available for predictive modeling.

Model

To deal with sparsity and high variable count, I use a multivariate lasso regression model (from the *glmnet* package) for regularization. Regularization attaches a penalty to non-zero covariate coefficients. The lasso model provides information on the tradeoff between including move covariates and higher prediction error. This tradeoff information allows the modeler to exclude costly non-zero coefficients that would increase prediction variance. The lasso model is used to predict monthly spending on electricity ("elecsummonth") in 2015 ZAR (South African currency) using household, demographic, and market data. To prevent high-value variables from overwhelming low-value variables in the model, I log-transform the following variables: price per MWh of electricity ("price"), house size ("Hsize"), house age ("Q14AGE"), number of rooms ("totalroom"), and monthly income ("incomesummonth").

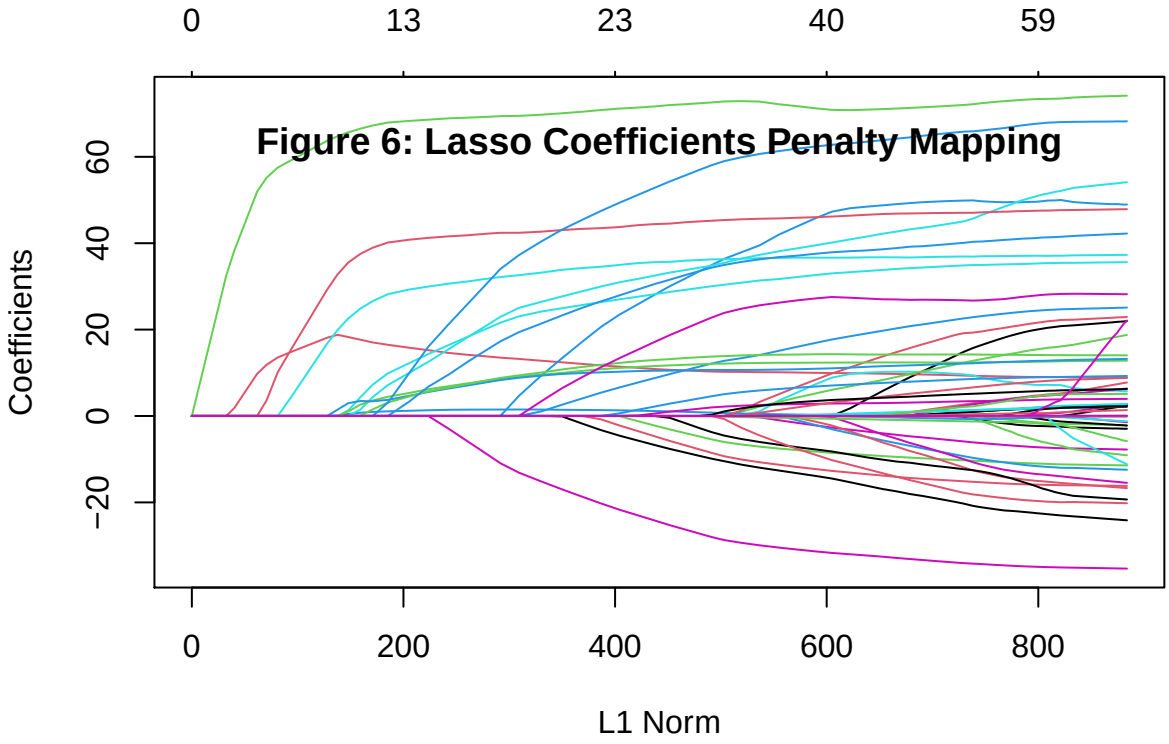
Out of the 230 available predictor variables (many of which are based on survey questions unrelated to home characteristics, demographics, or electricity usage), I selected 70 relevant predictor variables to input into the lasso model. Overall, I included log-transformed price and income variables, 14 demographic variables (e.g. race and education), 10 dwelling variables (e.g. dwelling type and size), 15 dummy variables for whether the households owns a particular electrical appliance (e.g. fridge and television), 9 geographic regions, 18 alternative energy source variables (e.g. firewood and diesel), and 2 seasonality variables (i.e. winter and summer).

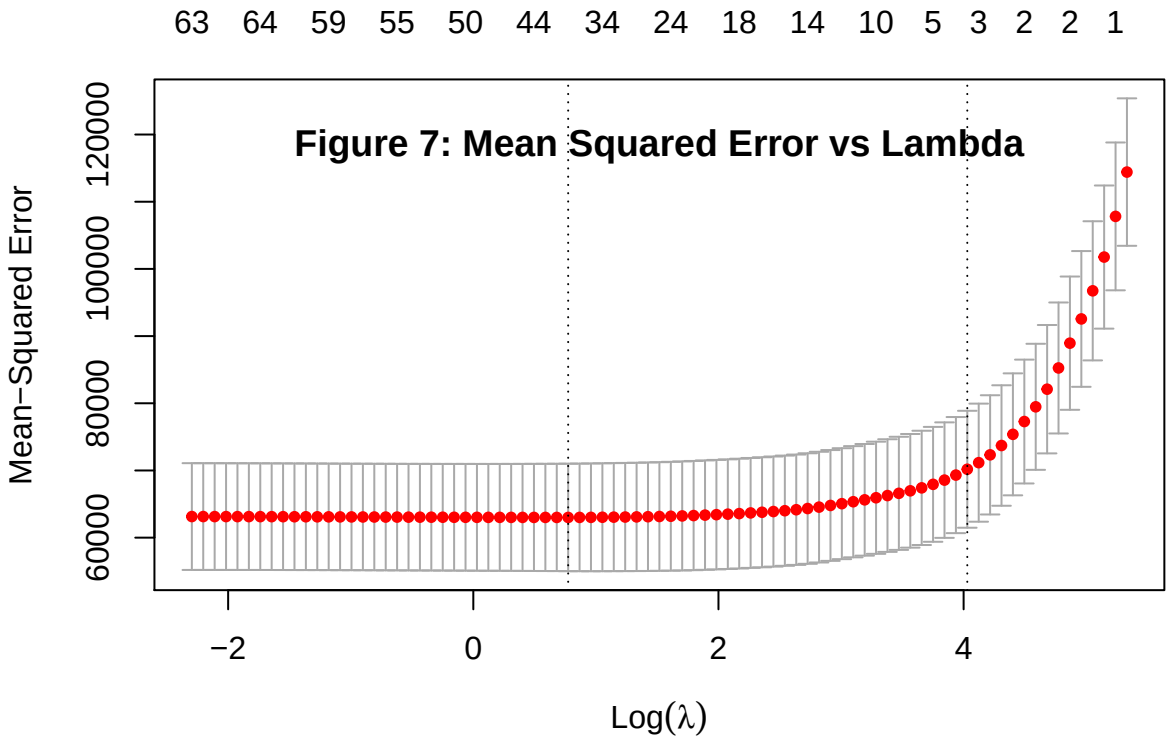
Finally, I account for the fact that some surveyed households consume zero electricity. Since I am only interested in predicting electricity spending that impacts load (electricity grid demand), I exclude households that do not consume any electricity. This step was performed in STATA prior to uploading the data in to R. Note that I do not exclude households receiving *free* electricity under the government subsidy program, "Free Basic Electricity" or "FBE". Spending for FBE households is included in the data, and these spending amounts are provided by the grid operator based on the prices concurrent with the timing of the free electricity provided.

The model is a five-fold cross-validated multivariate lasso regression applied to predict monthly household electricity spending. Prior to fitting the model to data, the dataset of 16,851 households is split into a training set (80%) and a test set (20%). The lasso model is fit to the training set data using a matrix of the 70 predictor variables. An optimal lambda of 2 is chosen to maximize non-costly predictive determinants in the model while minimizing mean squared error. This step is discussed further and visualized in Results section. The model is used to predict energy spending from covariate values in the test set in order to assess out-of-sample prediction accuracy.

Results

Figure 6 shows the coefficient penalty mapping for each of 70 possible coefficients in the model. Figure 7 shows the consequent relationship between the natural log of lambda and the mean squared error of the regression. Using Figure 7, I chose an optimal lambda of 2. At the point of $\log(2) = 0.693$ on the x-axis in Figure 7, the mean squared error in the model is roughly minimized while keeping 41 of the original 70 predictor variables in the regression. The tradeoff is thereby optimized by reducing prediction error while maintaining valuable covariate information in the model to use for prediction. The 41 coefficients used for regression after lasso regularization are shown in a matrix below, after Figure 7. These 41 covariates include log-transformed price and income variables, 9 demographic variables, 5 dwelling variables, 12 dummy variables for whether the households owns a particular electrical appliance, 4 geographic region indicators, 7 alternative energy source variables, and an indicator for the summer season.






```
## 71 x 1 sparse Matrix of class "dgCMatrix"
##                                1
## (Intercept)        -1.393871e+02
## ln_price            1.561634e+00
## ln_incomemonth      9.913170e+00
## dgender2           -8.754441e+00
## drace2              .
## drace3              4.800501e+01
## drace4              9.628731e+00
## dedu2               .
## dedu3               .
## dedu4               .
## dedu5               6.770231e+00
## IncomeDecile        1.662463e-01
## Consumptions        1.023510e-03
## under5             -3.112886e+00
## older60             .
## GenderOfHead       -3.334434e-02
## PopGrpOfHead        7.081229e+01
## ln_hsize            1.836956e+01
## ln_room             3.326429e+01
## ln_age              .
## dDtype1             .
## dDtype2             .
## dDtype3             1.107735e+01
## urban               .
```

```
## own                -3.889160e+00
## toilet              4.079492e+01
## FBE                -3.197368e+01
## winter              .
## summer             -1.305531e+01
## dradio1             .
## dstereo1            7.245556e+00
## dTV1                5.453809e-01
## dDVD1               2.924323e+00
## dfridge1            .
## dstove1             3.540098e+00
## dmicrowave1         1.237833e+01
## dwashing1           1.118663e+01
## dcomputer1          3.666792e+01
## dcamera1            .
## dcellphone1         3.870946e+00
## dlandlinephone1     4.635342e+01
## dDSTV1              1.427899e+01
## dinternet1          3.812794e+01
## dpowertool1         3.210883e-01
## dprov1             -1.459064e+00
## dprov2             -1.523512e+01
## dprov3             -3.386476e+00
## dprov4              .
## dprov5              6.308489e+01
## dprov6              .
## dprov7              2.739904e+01
## dprov8              .
## dprov9              .
## sqrtgas             .
## sqrtliquid         -6.999082e-01
## sqrtsolid          -8.803476e-02
## dgas                .
## dliquid             .
## dsolid             -8.833309e+00
## dpipewater1        -1.116612e+01
## paraffin            .
## petrol              .
## diesel              .
## firewoodBought      .
## firewoodFetched     .
## Charcoal           -8.741337e-02
## Candles             2.691029e-03
## Coal                .
## dungBought          .
## dungFetched         .
## otherhouseholdfuel  .
```

Using these 41 covariates, the out-of-sample root-mean-squared-error (RMSE) on the test set is 238.0866. This means that prediction error for household monthly electricity spending is roughly 238 ZAR. Figure 8 shows predicted spending versus actual spending for each household in the sample. If the predictions were perfect, all points would fall on the 1:1 diagonal line running through the plot. While the predictions are certainly positively correlated with actual values, there is very high variance in predictions around the line of perfect accuracy. Since actual monthly expenditures fall mainly below 500 ZAR, the RMSE of 238 ZAR is very high, indicating the model does a poor job of predicting household electricity expenditure each month. Figure 8 shows that at very low levels of expenditure (1-100 ZAR), the prediction appears to systematically *overestimate* expenditure. Conversely, at high levels of expenditure (1,000-1,500 ZAR), the prediction appears to systematically *underestimate* expenditure. Figure 9 adds nuance to prediction outcomes by showing predicted versus actual expenditures *for each income decile*. Figure 9 reveals that the model does a very poor job of estimating expenditure for low-income households, but appears to perform better for the top three income deciles. Notably, among the highest income decile, predictions are spread uniformly around the line of perfect prediction, indicating somewhat homoskedastic errors. At low income deciles, the predictions appear to be essentially useless.

Figure 8: Prediction vs Actual

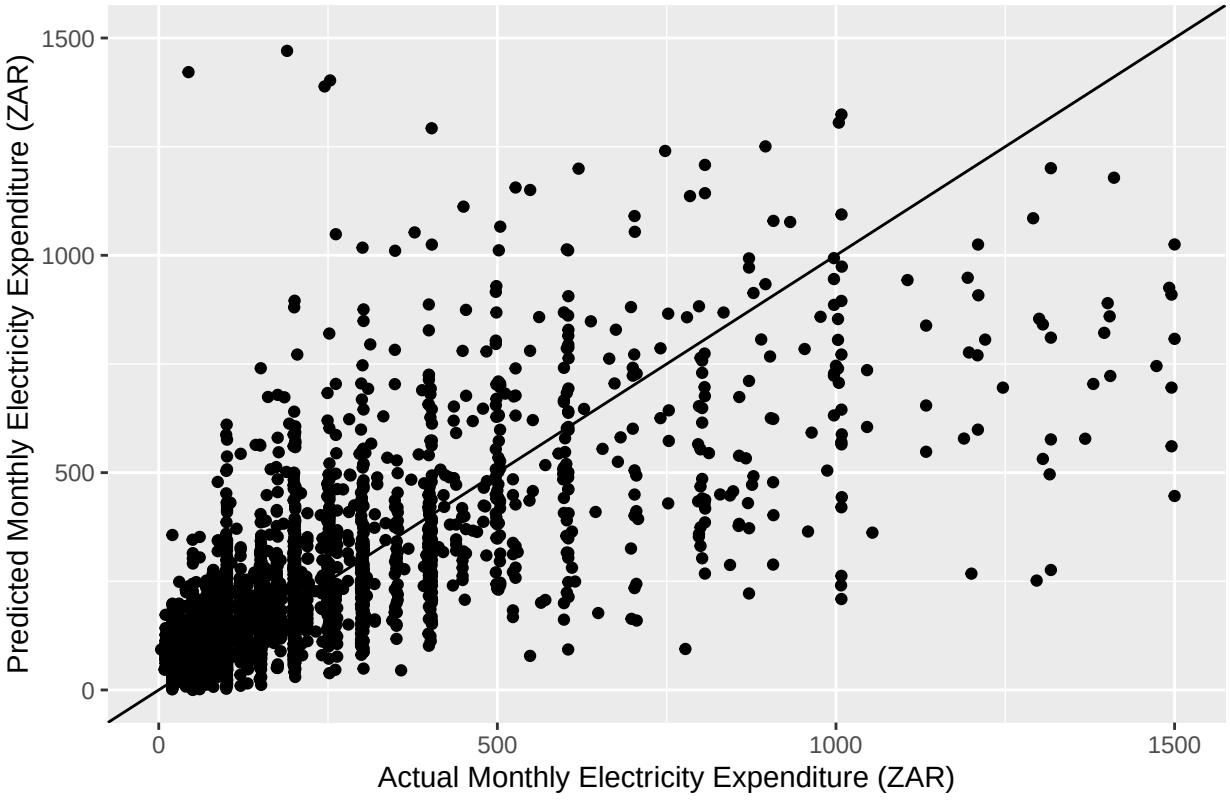
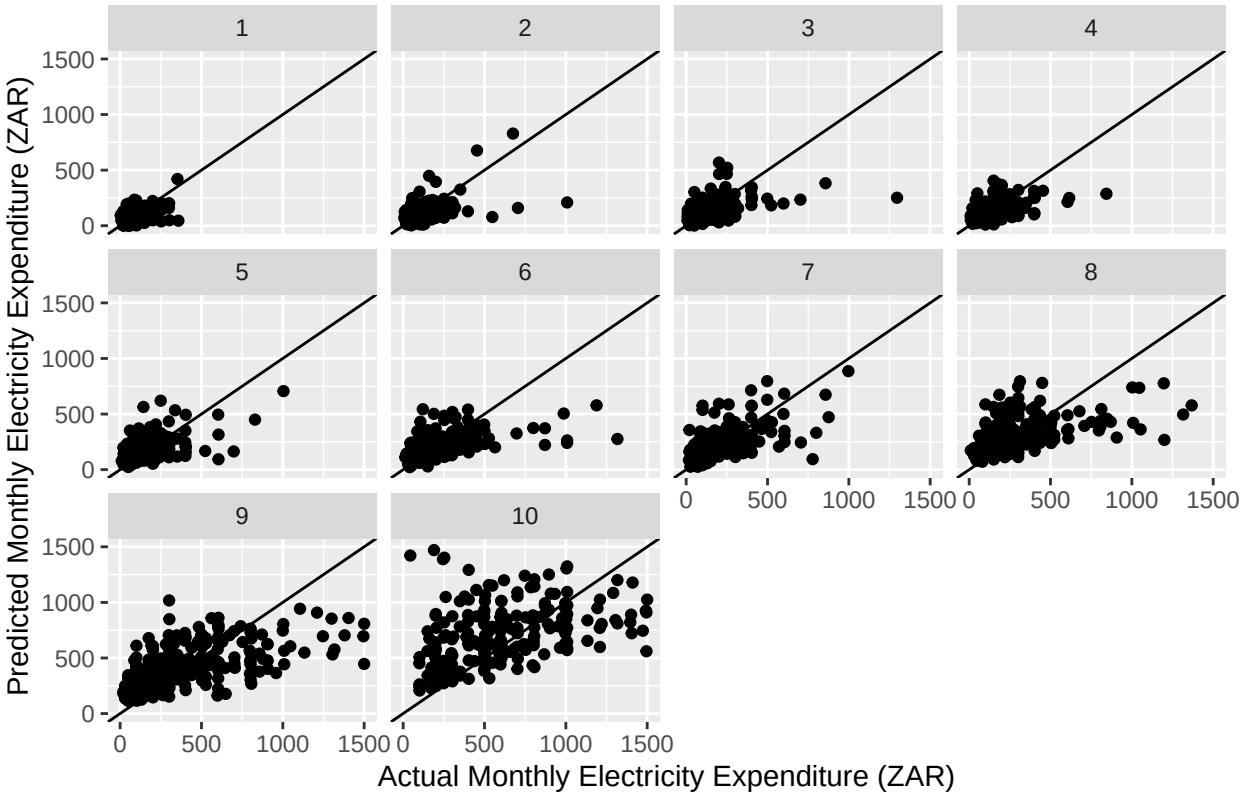


Figure 9: Prediction vs Actual (by Income Decile)



Conclusion

Overall, the prediction accuracy is poor since the RMSE (238 ZAR) is as large or larger than most of the low-expenditure households (those spending below 500 ZAR per month). One aspect of the data that may help explain this low prediction accuracy is the apparent *heaping* in the outcome variable. Figure 8 shows that many households suspiciously have exactly the same monthly expenditure, as seen in the vertical lines at nearly-regular intervals. For instance many houses report exactly 500 ZAR monthly spending. This may be a result of *rounding* during expenditure reporting in the data-gathering process. Heaping can make prediction difficult because rounding erodes the relationship between the outcome and the predictor variables. In other words, because heaping reduces variation in the outcome variable (by crowding observations together at certain intervals), there is less outcome variation available for predictor variables to exploit.

While heaping is a problem in this dataset, it does not seem to be the main reason for low prediction accuracy. The 41 variables used for regression appear to explain only some of the available variation in the outcome, regardless of heaping. As stated in results, the model overestimates expenditure for low-expenditure households and underestimates expenditure for high-expenditure households. This indicates that *quantile regression* should be used to better capture the unique relationships between expenditure and the covariates at different income levels. Households that can afford to spend more on electricity seem to have different relationships between spending behavior and their covariate characteristics than lower-income households. Quantile regression would estimate unique slopes and intercepts for each income decile. Therefore, while this analysis does a poor job of predicting expenditure, it still reveals important information for ESKOM–that ESKOM should separately estimate unique electricity consumption behavior at *different income levels* when making load forecasts.

Another useful quality of this model is its decent performance among households at the highest income decile. Facet ten in Figure 9 shows somewhat homoskedastic errors, indicating that the model may be an unbiased estimator among these high-earning households. It also indicates that the 41 predictor variables may be more relevant for high-income households than low-income households. Either low-income households need a different set of variables in order to generate good predictions, or expenditure at low-income households is too noisy to estimate at all. Therefore, ESKOM should develop separate electricity consumption (and expenditure) models for low-income and high-income household groups. Evidently, these groups have very different drivers of electricity demand. This is not surprising, considering the vast income inequality in South Africa. Low-income households and high-income households exist in very different economic contexts and rely on electricity for different needs and luxuries. The value of this analysis is therefore not in providing good predictions for overall household electricity expenditure but for revealing to ESKOM that it should model electricity usage separately by income-level in order to improve its forecasts and hopefully to thereby reduce load-shedding and its harmful impacts on low-income households.