Predictive Modeling of Logistics Performance Index using Sparse Regression Models

Odok Eric Oyenga

114865

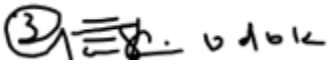
Thesis presented in partial fulfillment of the academic requirement for the Masters of Science in Statistical Science of Strathmore University

December, 2021

DECLARATION

I declare that this work has not been previously submitted and approved for the award of a degree by this or any other University. To the best of my knowledge and belief, the thesis contains no material previously published or written by another person except where due reference is made in the thesis itself.

Odok Eric Oyenga



10/11/2021

APPROVAL

The thesis of Eric Oyenga Odok was reviewed and approved by the following:

Prof. Samuel Mwalili,
Senior Lecturer, Institute of Mathematical Sciences,
Strathmore University

Dr. Elphas Okango,
Lecturer, Institute of Mathematical Sciences,
Strathmore University

Dr. Godfrey Achono Madigu,
Dean, Institute of Mathematical Sciences,
Strathmore University

Dr. Bernard Shibwabo,
Director of Graduate Studies,
Strathmore University

DEDICATION

This thesis is dedicated to my beautiful wife, Dr. Awuor and my favourite niece Hera.

ACKNOWLEDGEMENT

I would like to thank my supervisors Prof. Samuel Mwalili and Dr. Elphas Okango, for their guidance and immense contribution to ensure a flawless document. My special gratitude also goes to all my lecturers in the department of mathematical sciences at Strathmore University. To Dr. Mary Ochieng', this project would not be complete without her support and consistent follow-up. Lastly, to all my classmates, I cherish our interaction.

Table of Content

List of Tables

List of Figures

ABBREVIATIONS

LPI –Logistics Performance Index

LASSO – Least Absolute Shrinkage and Selection Operator

RRM – Ridge Regression Model

MLRM – Multiple Linear Regression Model

E_Net – Elastic Net Regression Model

PCA – Principal Component Analysis

ABSTRACT

The Logistics Performance Index (LPI), developed by The World Bank, is the only interactive benchmarking tool countries use to identify challenges and opportunities in trade logistics. It was developed using Principal Component Analysis and is a mean average of severely correlated variable scores; this poses two major problems: the susceptibility to outliers of mean computed measures and multicollinearity in prediction leading to overfitting. It is therefore critical to choose prediction techniques carefully. Regression is one of the many techniques, which can reliably predict the correct LPI. This paper accessed four regression models through median computed LPI, which is less vulnerable to outliers; the multiple linear regression model (MLRM), ridge regression model, elastic net model and LASSO model. The first observation was that mean and median computed LPI's were not different; in prediction, they both overfitted in the test data. Mean computed LPI, however, overfitted more than median. MLRM used all six variables to produce the best fit to the training set ($RMSE$ = 0.0497, AIC = − 952), however, tested on unseen data, it was the least precise ($RMSE$ = 0.0438). On the other hand, LASSO did not fit the training set well ($RMSE$ = 0.3627, AIC = −318) but was the most precise predictive model ($RMSE$ = 0.0436). LASSO, through variable shrinkage and selection, eliminated one irrelevant variable, timeliness. The two models were not significantly different (P = 0.2951, at 95% CI); the value addition through LASSO was parsimony. While MLRM used all six variables, LASSO used five to generate similar models. Policymakers could reliably use the top three variables that explained 80% of the variability in the model: logistics quality, infrastructure and tracking. Improving the physical infrastructure, increasing logistics management skills, and implementing intelligent technologies could improve trade competitiveness.

CHAPTER ONE

INTRODUCTION

1.1 Background to the study

The phrase "The world is a global village" has been used to describe situations where countries, regions and continents are brought together through interactions and activities such as sports, tourism and most importantly, trade. The term "Globalization" describes this phenomenon. Today, almost all countries in the world trade directly or indirectly with each other.

Supply chain and logistics are specifically emerging as key enablers to globalization since the mobility of products and people in an efficient manner are critical (Hausman (2005). The OECD (2005) estimates logistics costs at 15% of the total global turnover, contributing to international trade. Because logistics is at the core of stimulating economic development, several countries have developed comprehensive national logistics strategies (World Bank, 2016).

A fact-based metric provides a reliable benchmark to assess policy impact and compare global advancements. The Logistics Performance Index (LPI) developed recently by the World Bank is a crucial part of global efforts to understand logistics performance in the context of increasingly complex supply chains (World Bank, 2016) and assist countries in identifying opportunities in their trade logistics. LPI is the first and the only index to compare trade logistics around the world. The first was published in 2007, the second in 2010, and every two years after that (The World Bank, 2010), the most recent version is LPI 2020.

1.2 Statement of the problem

LPI is computed as a weighted average of six logistics performance variables (tracking and tracing, customs, infrastructure, logistics services, international shipments and timeliness). The Principal Component loadings computed from an eigenvector are used as weights and are relatively constant through the years (World Bank, 2014), as shown below.

$$(LPI)_i = w_1 c_{i1} + w_2 c_{i2} + w_3 c_{i3} + w_4 c_{i4} + w_5 c_{i5} + w_6 c_{i6},$$

Where, $w_1 = 0.40, w_2 = 0.42, w_3 = 0.40, w_4 = 0.42, w_5 = 0.41, w_6 = 0.40$ are the Principal Component loadings, and $c_{ij,(i=(1,2,3,\ldots,160),j=(1,2,3,\ldots,6))}$ are the standardized variable scores for

individual countries. The weights are equal making LPI a simple average of standardized variable scores (World Bank, 2016).

The first problem is that simple means are highly susceptible to outliers; they skew and produce non-representative measures (S.Manikandan, 2011). Secondly, loadings are components of eigenvectors; they do not provide any information or interpretation on or about the original data (LPI variable scores) (Holland, 2019). Thirdly, used as weights, loadings do not represent the expected "Importance" of variables; instead, they assume equal effect across the years, which is not consistent with reality (Ghojogh & Crowley, 2019). Lastly, in modelling, they introduce strong multicollinearity that leads to inaccurate estimates of coefficients, standard errors and P-Values, degrading predictability of scores (Mundfrom & Smith, 2018).

The result is that policymakers cannot focus and predict the performance of specific areas to improve trade logistics.

1.3 Main objective

The main objective of this study was to develop predictive models for Logistics Performance Index using Sparse Regression models.

1.4 Specific objectives

i. To formulate multiple linear regression model (MLRM), ridge regression model (RRM), the elastic net (E_Net) and the least absolute shrinkage and selection operator (LASSO) for logistics performance index (LPI)
ii. To determine the best predictive model for logistics performance index (LPI)
iii. To select important logistics performance variables using the best model

1.5 Significance of the study

The important variables selected through the predictive model would inform policymakers on the reforms needed in trade logistics to enhance global competitiveness.

CHAPTER TWO

LITERATURE REVIEW

2.1 Logistics Performance Index (LPI)

LPI is a derived measure from average scores of responses from countries based on 10-15 questions on six components, namely: customs, infrastructure, quality of service, timeliness, ease of arranging shipments and tracking & tracing. A statistical technique called Principal Component Analysis (PCA) generates the weights (loadings) of scores. The sum of the product of the weights (loadings) and standardized scores produce LPI (World Bank, 2016).

PCA is a dimensionality reduction technique developed in 1901 to reduce the dimension of datasets $\boldsymbol{X}$ without losing a lot of information. It transforms data from a d -dimensional space $\boldsymbol{X}$ into a new coordinate system $\mathbf{Y}$ of dimension p.

$$\boldsymbol{X} \epsilon \boldsymbol{R}^{d} \rightarrow \boldsymbol{Y} \epsilon \boldsymbol{R}^{p}, \text{where } p \ll d.$$

Higher-dimensional spaces are complex and difficult to analyze. The goal of PCA is to reduce this complexity but preserve as much variance in the original data as possible. The new variables that form the new coordinate system are called Principal Components (PC's), denoted by $\mathbf{U_1}, \mathbf{U_2}, \ldots . \mathbf{U_d}$. These PC's are orthogonal linear transformations of the original variables, and there are at most d of them. The first PC ($\mathbf{U_1}$) is called the first principal component and has the maximum variance, thus accounting for the most variability in the data. The second PC ($\mathbf{U_2}$) is called the second principal component and has the second-highest variance and so on, until $\boldsymbol{d^{th}}$ PC ($\mathbf{U_d}$), which has the minimum variance (Holland, 2019).

Suppose that $\mathbf{X} = (\mathbf{X_1}, \mathbf{X_2}, \ldots \mathbf{X_n})$ in a d- dimensional space ($\mathbf{X_i} \epsilon \boldsymbol{R}^{d}$) is projected on $\mathbf{U_1}$, that is, $\mathbf{U_1^T X}$ where $\mathbf{U_1}$ is $1xd$, and $\mathbf{X}$ is dxn. $\mathbf{U_1^T X}$ is in a lower dimension, $1xn$. It is also hoped that the variance of $\mathbf{U_1^T X}$ is as large as possible. Maximizing the variance with unknown $\mathbf{U_1}$ creates an optimization problem,

$$\text{Max } Var\ (\mathbf{U_1^T X}).$$

$Var\ (\mathbf{U_1^T X}) = \mathbf{U_1^T S U_1}$, where $\mathbf{S}$ is a dxd sample covariance matrix and,

$$\text{Max } Var\ \left(\mathbf{U_1^T X}\right) = \text{Max } (\ \mathbf{U_1^T S U_1}).$$

This function is quadratic and has no upper bound. The length of the vector $\mathbf{U_1^T U_1}$ is fixed to define the problem. In this case, it is fixed to one; however, one can use any length. The new optimization problem is given by;

$$\text{Max } (\mathbf{U_1^T S U_1})$$
$$\text{s.t } \mathbf{U_1^T U_1} = 1.$$

Using the Lagrangian approach,

$$L(\mathbf{U_1}, \lambda_1) = \mathbf{U_1^T S U_1} - \lambda_1(\mathbf{U_1^T U_1} - 1),$$
$$\frac{\delta L}{\delta \mathbf{U_1}} = 2\mathbf{S U_1} - \mathbf{2}\lambda_1 \mathbf{U_1} = 0,$$
$$\mathbf{S U_1} = \lambda_1 \mathbf{U_1}.$$

λ_1 is the eigenvalue of sample covariance matrix **S** corresponding to the maximum eigenvector of **S.**

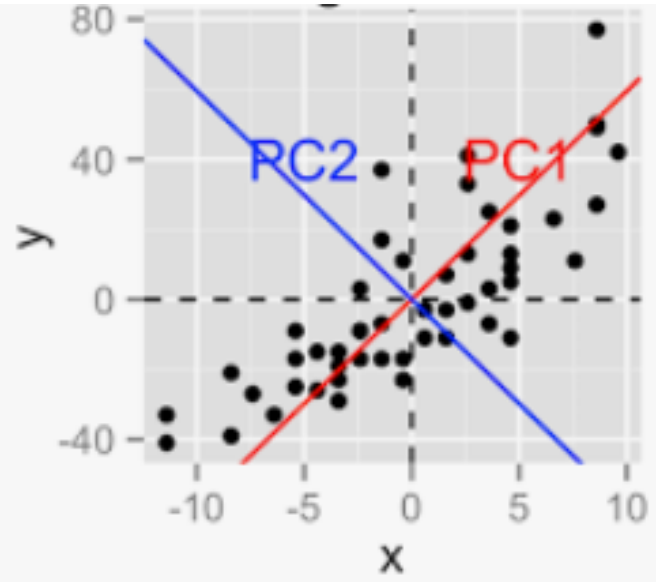


Figure 2.1 Principal Component Analysis (Source: Analyticsvidhya)

PCA is applicable for square matrices; however, for a dataset such as LPI that is rectangular a more subtle approach of singular value decomposition is used. Suppose,

$$\mathbf{X = U \Sigma V^T},$$

Where **U** is the eigenvector of $\mathbf{XX^T}$ and **V** the eigenvector of $\mathbf{X^T X}$.

If $\mathbf{X}$ is centralized, that is, $\mathbf{X}' = (\mathbf{X} - \mathbf{M})$, where $\mathbf{M}$ is the overall mean of the matrix,

$$\mathbf{X}' = \mathbf{U}\,\mathbf{\Sigma}\,\mathbf{V}^{\mathrm{T}},$$

$\mathbf{U}$ and $\mathbf{V}$ are the sample covariance matrices $\mathbf{X}'\mathbf{X}'^{\mathbf{T}}$ and $\mathbf{X}'^{\mathbf{T}}\mathbf{X}'$ **eigenvectors,** respectively. The components of $\mathbf{U}$ and $\mathbf{V}$ are called loadings. Principal components examine the magnitude and direction of the coefficients for the original variables. The larger the coefficient (loading) value, the more important the corresponding variable is in calculating the component (Ghojogh & Crowley, 2019).

Principal component analysis was not used in the construction of LPI as a dimensionality reduction technique but for generating the loadings. Loadings are weights used to generate the orthogonal and linear combinations of central values of the original variables; however, they are closely associated with the sample variance-covariance matrices through the eigenvectors and can only be interpreted with respect to the direction and magnitude of PC's. Secondly, the interpretation of PCA is not as direct and simple; one of the limitations of this old age method is that interpreting results may be as subtle as it gets. Applying them directly to the original variables as weights grossly violates the basic principle of the concept described thus far (Holland, 2019).

Robust techniques for assigning weights and selecting variables exist; for instance, Jafar, Wilco & Lorant (2020) developed a model for measuring the relative importance of LPI indicators using the Best Worst Method (BWM). This method was able to help direct policy priorities compared to the current LPI; however, because of the correlation between the variables, the effect was only mild. Other weighting methods like, Raking, Greg, and Logit regression (Kalton & Flores-Cervantes, 2003) are effective, but in prediction, they are affected by multicollinearity. Recent predictive modelling techniques can weight, select, or eliminate variables and cancel the effect of multicollinearity to develop robust models (Zou & Hastie, 2005). The ensuing literature discusses them.

2.2 Predictive Modelling

Predictive models are a set of techniques used to make inferences about uncertain future events. The key concepts underlying predictive modelling were developed around the beginning of the nineteenth century, with Legendre and Gauss. Through their work on linear regression, in a book titled “méthode des moindres carrés", a "least-squares method" (Paris, 2012). Techniques

developed thus far based on the linear regression approach could only model simple linear relationships. The other computationally expensive methods lagged. Towards the end of 1980, with increased computational power and access to data (Bromley A.G., 1998), advanced statistical research emerged. Classification and regression tree algorithms were some of the first to tap into the new possibilities. Hastie and Tibshirani simulated "generalized additive models" for a class of generalized linear models, demonstrating the new possibilities (G. James and others, 2013).

However, the emerging statistical techniques and methods fell short in addressing the widening gap between data generation and understanding. The data analysis process was very time consuming; additionally, database managers and system engineers had little or no mathematical background to work with information at their disposal. Machine learning was born in that instance to automate the discovery process. Statistical learning algorithms augmented Machine Learning by embedding core mathematical techniques. The convergence of the two technologies opened possibilities to many applications (Cunningham and others, 2014).

In high dimensional data sets (Key feature of the new datasets) with more explanatory variables than the sample size, $p \gg n$ the traditional analysis techniques like Ordinary least squares (OLS) developed in the 1900's could not adequately analyze them. In this context, techniques of subsetting (choosing a subset from available set of variables) to reduce the dimensionality p and shrinkage (reducing the size of the coefficients) were developed by Charles Stein in the 1960s and expanded upon later by Stanley Sclove and Tibshirani recently. These techniques reduced complex, "wide" or "fat" datasets to a more familiar "tall" structure that would specifically improve the prediction accuracy of simple linear regression (Tibshirani, 1996).

In OLS, the smaller the error, the better the model fits the data. The two main subcomponents, error due to "bias" and "variance", are critical to understanding the error. They help in describing the behaviour of a prediction model and thus improve the data fitting process. The error due to bias is the difference between the predicted value of the model and the observed value. The error due to variance, on the other hand, is the variability of a model prediction. The model is iterated, and the differences between the predictions and realizations are calculated. The ideal model is one with low bias and low variance, which rarely happens. There is, therefore, a need for a trade-off between bias and variance to avoid over or underfitting. These two are related to the complexity of the model. As the model complexity increases, for instance, by increasing the number of

explanatory variables, bias reduces, and variance increases, the primary concern becomes variance. The reverse is true. The task is to find a "sweet" spot for the model, where the increase in bias reduces variance (Frank & Friedman, 1993).

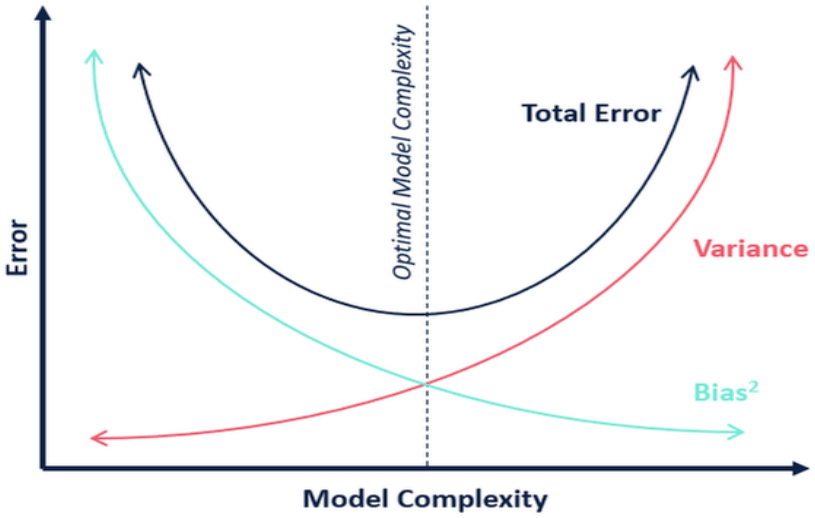


Figure 2.2 Bias-Variance trade-off (Source: Analyticsvidhya)

Gauss Markov Theorem states that OLS has the least variance amongst all linear unbiased estimators (Halliwell, 2015); the question not answered in this theorem is whether there could be a biased estimator with an even smaller variance than OLS. The approach targeted shrinkage parameters to find a biased estimator with a smaller variance than OLS.

Suppose that the OLS estimator, $\beta_{ols,}$ is replaced by something smaller,

$$\beta_k = \frac{1}{1+\lambda}\,\beta_{ols},$$

If $\lambda = 0$, then $\beta_k = \beta_{ols}$. If $\lambda > 0$, then β_K approaches 0. λ, is referred to as shrinkage estimator. If λ is chosen well, an estimator β_K more precise than β_{ols} can be estimated. Whereas β_{ols} is unbiased, β_K is biased. Finding the right λ means balancing the two errors; the following expression estimate λ relatively well (Sclove, 1963).

$$\lambda = \frac{p\sigma^2}{\sum\beta_k^2} \qquad \text{(i)}$$

If the coefficients are large relative to their variances, set λ to be small; on the other hand, if the coefficients are small, set λ to be large to approach zero. The information in (i) is usually not available to assist formulate an optimal λ. Supposed σ^2 in (i) is known, Stein and Willard proposed a specification:

$$\beta_k = \left(1 - \frac{(p-2)\sigma^2}{\sum\beta_{ols}^2}\right)\beta_{ols}. \qquad \text{(ii)}$$

Sclove further proposed to shrink the estimates to 0 if a negative β_k is obtained. This he proposed using the following formulae,

$$\beta_k = \left(1 - \frac{(p-2)\sigma^2}{\sum\beta_{ols}^2}\right)^+ \beta_{ols} \qquad \text{(iii)}$$

Where, $\left(1 - \frac{(p-2)\sigma^2}{\sum\beta_{ols}^2}\right)^+ = \max\left(\left(1 - \frac{(p-2)\sigma^2}{\sum\beta_{ols}^2}\right)^+, 0\right)$

Therefore, the test statistics removes the coefficient by reducing it to zero or shrinks them to some negligible value (Trevor Hastie, 2013).

These shrinkage and subsetting techniques solved the complexities of high dimensional datasets, $p \gg n$ and multicollinearity. The need to shrink and subset was additionally very instrumental in developing weights and assigning "importance" by way of ranking.

Much research has been done on the high dimensional (big) and sparse data sets and little on low dimensional and non-sparse. The intent was to explore them on the LPI dataset with the later characteristics.

The ensuing chapter utilized the shrinkage parameter to derive shrinkage techniques applicable to LPI datasets.

CHAPTER THREE

RESEARCH METHODOLOGY

3.1 Research Design

This was a retrospective cross-sectional study of the World Bank's International Logistics Performance Index, LPI 2014 and LPI 2016, published in 2014 and 2016 respectively by the World Bank's Global Trade and Regional Integration Team (World Bank, 2016).

3.2 Dataset and Preprocessing

LPI 2014 and LPI 2016 was generated using data from the first 10-15 questions of an online survey. Each survey respondent rated eight overseas countries on the following six variables.

i. Infrastructure; The quality of transport–related development
ii. Customs; The efficiency of customs clearance
iii. International Shipments; The arrangement of competitively priced shipments
iv. Timeliness; The timely delivery of goods with desired quality
v. Tracking and Tracing; The ability to track and trace goods on transit
vi. Logistics Quality; The overall competence and quality of logistics services.

The web engine for LPI 2014 was designed to use a Uniform Sampling Randomized (USR) method to gain adequate responses from unrepresented countries. Country i was chosen with probability $\frac{N-n_i}{2N}$, where n_i was the sample size of country i and N, the total sample size. Selected countries were also based on important exports and imports of the respondent's country. Landlocked countries used the neighbouring nations connecting them with international markets. Without randomization, the engine would select countries with high trade volume alone; USR increased the chances of selecting lower trade volume countries using non-uniform probability (World Bank, 2014). LPI ranked one hundred and sixty countries based on responses from 1,000 logistics professionals in 130 countries. LPI 2016 research design was similar to LPI 2014; however, the survey was done in two phases, between October and December 2015 and March to April 2016 in phases one and two, respectively; this helped increase the response rate. The index also ranked one hundred and sixty countries based on responses from 1,051 logistics professionals (World Bank, 2016).

Respondents evaluated countries using a five-point Likert scale (1-5), with the lowest score indicating "Poor" and the highest "Excellent" in both datasets. The country scores for the variables were then averaged across all respondents (World Bank, 2016). The scores were rounded off to two decimal places, and simple averages imputed missing values.

The World Bank LPI was compared with the mean and median computed LPIs. The mean and median computed LPIs were generated through simple mean and median averages of the country scores for the six variables, respectively.

3.3 Simple descriptive statistics and multicollinearity assessment

Summary statistics was used to find patterns and access the magnitude of the relationship between the variables. The analysis delved further into multicollinearity.

Multicollinearity assessment was done to check if the predictors were correlated with one another, this would imply that one predictor would be estimated using another predictor in the same model. The problem would be how to distinguish between the individual effects of the predictors on the target variable. Say a linear predictor,

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2.$$

Where, β_1 would be the increase in y for a unit increase in x_1 while adjusting for x_2 , and β_2 the increase in y for a unit increase in x_2 while adjusting for x_1. Since x_1 and x_2 would be highly correlated, a change in x_1 would also cause a change in x_2, making it difficult to see the individual effect on y (Shrestha, 2020).

To detect multicollinearity in the dataset, Variance Inflation Factors (VIF) approach was used. $VIF = \frac{1}{1-R^2}$, where, R^2 was the coefficient of determination. The closer the R^2 was to one, the higher the VIF, and the higher the multicollinearity (Shrestha, 2020).

3.4 Multiple Linear Regression Model (MLRM)

The aim was to predict LPI scores of 160 countries ($\mathbf{Y}$) using scores of six predictors ($\mathbf{X} = \boldsymbol{x_1, x_2, \ldots x_6}$). $\boldsymbol{x_1}$ = Logistics quality, $\boldsymbol{x_2}$ = Infrastructure, $\boldsymbol{x_3}$ = International shipment, $\boldsymbol{x_4}$ = Customs, $\boldsymbol{x_5}$ = Tracking and tracing, and $\boldsymbol{x_6}$ = Timeliness. Each of these variables had 160 scores from different countries, that is, $\boldsymbol{x_i}$ = (160 x 1) dimension. The model was given by, $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$,where, $\boldsymbol{Y}$ =(160 x 1), $\boldsymbol{X}$ = (160 x 6), $\boldsymbol{\beta}$ = (6 x 1), dimensions and $\boldsymbol{\varepsilon} \sim N(0, \sum)$. $\mathbf{X}$ was known and

fixed, $\mathbf{Y}$ was known and fixed but $\boldsymbol{\beta}$ which quantified the effect of the variable was unknown and randomly distributed. $\boldsymbol{\beta}$ was estimated from the sample using the Ordinary Least Squares (OLS) approach. It was expected that the estimated $\widehat{\boldsymbol{\beta}}$ had the sum of square residuals as small as possible and minimized the loss function,

$$\sum(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}})^2 = \left|\left|\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right|\right|^2.$$

$\boldsymbol{\beta}$ was estimated by OLS method, and was expected to give rise to the best linear unbiased estimator (BLUE),

$$\widehat{\boldsymbol{\beta}}_{\boldsymbol{OLS}} = (\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-1}(\mathbf{X}^{\mathbf{T}}\mathbf{Y}),$$

satisfying Gauss-Markov properties**.**

$$E(\widehat{\boldsymbol{\beta}}_{\boldsymbol{OLS}}) = \boldsymbol{\beta}, \text{ and}$$

$$Var(\widehat{\boldsymbol{\beta}}_{\boldsymbol{OLS}}) = \sigma^2(\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-1} < Var(\widehat{\boldsymbol{\beta}}_{\boldsymbol{i}}),$$

where $\widehat{\boldsymbol{\beta}}_{\boldsymbol{i}} \neq \widehat{\boldsymbol{\beta}}_{\boldsymbol{OLS}}$ would be other estimators.

That is, the true parameter and the estimated statistic would be similar and uncertainty minimal. The unknown variance would be, $\hat{\sigma}^2 = \frac{\epsilon^T\epsilon}{n-m}$, where, $\epsilon = \mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}_{\boldsymbol{OLS}}$. The residual would be decomposed into, error resulting from bias, error resulting from variance and unexplained error. That is;

$$E(\epsilon) = \left(E(\mathbf{X}\widehat{\boldsymbol{\beta}}) - \mathbf{X}\boldsymbol{\beta}\right)^2 + E\left(\mathbf{X}\widehat{\boldsymbol{\beta}} - E(\mathbf{X}\widehat{\boldsymbol{\beta}})\right)^2 + \sigma^2 = Bias^2 + Var + \sigma^2.$$

Therefore, the first objective was that the estimated model $\widehat{\mathbf{Y}} = \mathbf{X}\widehat{\boldsymbol{\beta}}_{\boldsymbol{OLS}}$, would fit the data well and be a good fit for a new dataset, LPI 2016. The new predictor scores ($\mathbf{X}'$) would predict the new LPI ($\mathbf{Y}' = \mathbf{X}'\widehat{\boldsymbol{\beta}}_{\boldsymbol{OLS}}$) in such a way that the predicted $\mathbf{Y}'$ would be as close as possible to the actual $\mathbf{Y}$. $\mathbf{Y} - \mathbf{Y}'$ being the prediction error (PE) would be as minimal as possible. The second objective was to reach an economical model; this used a backward elimination method. The method started with a full model comprising of the six variables and applied p-value elimination criteria to develop a model with the highest predictive value using only significant variables. The two scenarios (mean and medium calculated LPI) applied the backward model selection method to arrive at the final

model. The two models were then evaluated based on how well they fitted the training dataset (LPI 2014) using AIC and RMSE.

3.5 Ridge Regression Model (L_2 Penalty)

The OLS estimator is always unbiased; however, the variance would increase exponentially when the predictors are correlated (Trevor Hastie, 2013). To control variance, removing some predictors in the model would be a viable solution; the problem would be that the predictors eliminated (setting coefficient β to zero) from the model would be unknown. Instead of elimination, the coefficients that are too far from zero would be penalized by a factor, in a way that the complexity of the model would decrease while retaining all variables, a process called regularization (Tibshirani, 1996).

The loss function would be similar to MLRM but with a constraint that prevents coefficients $\boldsymbol{\beta}$ from growing large and exploding. The optimization problem below would give rise to a ridge solution,

$$\text{minimize } \sum_i^n \left(y_i - x_i^T \hat{\beta}\right)^2 \;\; s.t \; \sum_j^p \beta_j^2 \; \leq t, \text{ where}$$

$\mathbf{X} = (\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots \boldsymbol{x_n})$ standardized**,** and $\mathbf{Y} = (\boldsymbol{y_1}, \boldsymbol{y_2}, \ldots \boldsymbol{y_n})$ centred, would result **in the** penalized residual sum of squares (PRSS) also called ridge regression, L_2 and the loss function is given by,

$$\sum \left(y_i - x_i^T \hat{\beta}\right)^2 + \lambda\left(\sum_{i=1}^m \hat{\beta}_i^2\right) = \left|\left|\mathbf{Y} - \mathbf{X\beta}\right|\right|^2 + \boldsymbol{\lambda}\left|\left|\boldsymbol{\beta}\right|\right|^2 = \left|\left|\mathbf{Y} - \mathbf{X\beta}\right|\right|^2 + \boldsymbol{\lambda}\left|\left|\mathbf{0} - \boldsymbol{\beta}\right|\right|^2 .$$

In this paper, n = 160 and m = 6.

The computation of $\boldsymbol{\beta}$ was similar to MLRM for the first part plus the term $\boldsymbol{\lambda}\left(\sum_{i=1}^m \hat{\beta}_i^2\right)$. With $\boldsymbol{\lambda}$ being the regularization penalty.

The estimated $\boldsymbol{\beta}$ from ridge regression had a smaller average prediction error than $\widehat{\boldsymbol{\beta}}_{OLS}$. Using Maximum Likelihood Estimation, a unique solution was as follows,

$$\frac{\delta PRSS}{\delta \hat{\beta}_{Ridge}} = -2\boldsymbol{X^T}(\boldsymbol{Y} - \boldsymbol{X^T\beta}) + 2\boldsymbol{\lambda\beta}$$

$$\hat{\beta}_{Ridge} = (\mathbf{X^T X} + \boldsymbol{\lambda}\mathbf{I})^{-1}(\mathbf{X^T Y}),$$

I being the identity matrix. As $\boldsymbol{\lambda}$ tended to zero, the Ridge estimator approached OLS estimator. As λ tended to infinity, the Ridge estimator approached zero.

3.5.1 Proving that $\boldsymbol{\beta}_{\boldsymbol{Ridge}}^{\boldsymbol{\lambda'}}$ was biased

$$\boldsymbol{\beta}_{\boldsymbol{Ridge}}^{\boldsymbol{\lambda'}} = (\mathbf{X^T X} + \boldsymbol{\lambda}\mathbf{I})^{-1}(\mathbf{X^T Y}),$$

Let $\mathbf{W} = \mathbf{X^T X}$,

$$= (\mathbf{W} + \boldsymbol{\lambda}\mathbf{I})^{-1}\mathbf{W}(\mathbf{W^{-1}X^T Y}),$$

$$= (\mathbf{W}(\mathbf{I} + \boldsymbol{\lambda}\mathbf{W^{-1}})^{-1}\mathbf{W}((\mathbf{X^T X})^{-1}\mathbf{X^T Y}),$$

$$= (\mathbf{I} + \boldsymbol{\lambda}\mathbf{W^{-1}})^{-1}\mathbf{W^{-1}W}\widehat{\boldsymbol{\beta}}_{OLS},$$

$$= (\mathbf{I} + \boldsymbol{\lambda}\mathbf{W^{-1}})\widehat{\boldsymbol{\beta}}_{OLS}.$$

Taking expectations on both sides,

$$\mathbf{E}\,(\boldsymbol{\beta}_{\boldsymbol{Ridge}}^{\boldsymbol{\lambda'}}) = \mathbf{E}\,((\mathbf{I} + \boldsymbol{\lambda}\mathbf{W^{-1}})\widehat{\boldsymbol{\beta}}_{OLS}),$$

$$= (\mathbf{I} + \boldsymbol{\lambda}\mathbf{W^{-1}})\boldsymbol{\beta}, \text{ for } \boldsymbol{\lambda} \neq \mathbf{0}.$$

$$\mathbf{E}\,(\boldsymbol{\beta}_{\boldsymbol{Ridge}}^{\boldsymbol{\lambda'}}) \neq \boldsymbol{\beta}.$$

As $\boldsymbol{\lambda}$ increased, bias increased while variance decreased (Kennard & Hoerl, 1970). A Careful balance was achieved by choosing an optimal $\boldsymbol{\lambda}$.

3.5.2 Choosing the optimal $\boldsymbol{\lambda}$

A standard practice using cross validation was adopted to tune $\boldsymbol{\lambda}$ (Zou & Hastie, 2005). The approach chose $\boldsymbol{\lambda}$ that minimized the mean squared error of the model. A machine learning method, which evaluated the accuracy of a statistical model using a training set, and tested the performance of the model in a new dataset called test set, was used. Training set $N = 160$ was partitioned into $\mathrm{K} = 10$ sets of equal sizes. Each set had 16 data points. A finite but large sequence of possible $\boldsymbol{\lambda}$ were generated randomly**.** Nine folds were used to fit the model $\widehat{\mathbf{Y}}$. The remaining fold was used to compute cross-validation error (Mean Square Error (MSE)) as follows;

$$(\text{CV Error})_{\mathrm{k}}^{\lambda} = |\mathrm{N_K}|^{-1}\sum_{x,y\epsilon N_K}\left(y - \hat{Y}_k^{\lambda}\right)^2,$$

and the overall cross-validation error given by;

$$(\text{CV Error})^{\lambda} = K^{-1}\sum_{k=1}^{k}(\text{CV Error})_k^{\lambda},$$

The optimal $\boldsymbol{\lambda}'$ was the one with the minimum $(\text{CV Error})^{\lambda}$ (Zou & Hastie, 2005) and was used to fit RRM. Ridge regression shrunk the variables that were not important in the model. Two ridge regression models (Mean and median computed LPI) were generated using the training dataset, the models were then compared and assessed using Akaike Information Criteria and Root Mean Squared Error.

3.6 The LASSO ($\mathbf{L_1}$ Penalty)

The Least Absolute Shrinkage and Selection Operator (LASSO) was similar to the Ridge regression only that the penalty term was the sum of the absolute $L1$ values of coefficients.

Tibshirani (1996) developed LASSO, in the attempt to address the sparsity and variable selection through the optimization problem subject to the sum of absolute values of coefficients. The LASSO coefficients are the solutions to the optimization problem,

$$\text{Minimize } \sum_i^n \left(\mathbf{Y_i} - \mathbf{X_i^T}\widehat{\boldsymbol{\beta}}_\mathbf{j}\right)^2 \;\; s.t \; \sum_j^p |\widehat{\boldsymbol{\beta}}_\mathbf{j}| \; \leq t,$$

where, $\mathbf{X_i} = (\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots \boldsymbol{x_n})$ and $\mathbf{Y_i} = (\boldsymbol{y_1}, \boldsymbol{y_2}, \ldots \boldsymbol{y_n})$ are standardized and centered respectively.

The equivalent loss function is given by,

$$\sum_{i=1}^n \left(\mathbf{Y_i} - \mathbf{X_i^T}\widehat{\boldsymbol{\beta}}_\mathbf{j}\right)^2 + \boldsymbol{\lambda} \sum_{j=1}^m |\widehat{\boldsymbol{\beta}}_\mathbf{j}|.$$

Similar to Ridge regression, the tuning parameter $\boldsymbol{\lambda}$ controlled the amount of regularization. For high values of $\boldsymbol{\lambda}$ many coefficients were reduced to zero, therefore eliminating the variables from the model, and deriving a set of sparse solutions. A few features were assigned the entire effect, and the others reduced to zero. This way LASSO performed model selection. Unlike the Ridge regression, $\widehat{\boldsymbol{\beta}}_{\boldsymbol{LASSO}}$ had no numerical closed form solution. The applicable computational solutions involved quadratic programming techniques from convex optimization (Tibshirani, 1996). The computation was made simpler by using packages in R. Two models were generated, mean and median computed LPI models were then assessed using AIC and RMSE.

3.7 Elastic Net

Elastic net combined the best of ridge and LASSO to minimize the loss function given by,

$$\sum_{i=1}^{n} \frac{\left(\mathbf{Y_i} - \mathbf{X_i^T}\widehat{\boldsymbol{\beta}}_\mathbf{j}\right)^2}{2n} + \boldsymbol{\lambda}\left(\frac{1-\alpha}{2}\right) \sum_{j=1}^{m} \widehat{\boldsymbol{\beta}}_\mathbf{j}^2 \; + \alpha \sum_{j=1}^{m} |\widehat{\boldsymbol{\beta}}_\mathbf{j}|$$

where, $\mathbf{X_i} = (\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots \boldsymbol{x_n})$ and $\mathbf{Y_i} = (\boldsymbol{y_1}, \boldsymbol{y_2}, \ldots \boldsymbol{y_n})$ were standardized and centered respectively. $0 < \alpha < 1$ was a mixing parameter between ridge and LASSO. λ and α were the two parameters tuned to give the optimal $\widehat{\boldsymbol{\beta}}_{\boldsymbol{E\_net}}$. The mean and median computed LPI models were generated using E_Net and compared using RMSE and AIC.

3.8 Selecting relevant components

Using the principle of economy (Occam's razor) or the law of parsimony, the most preferred model had the smallest number of explanatory variables (Sober, 1981). Two simple graphical methods were used to perform variable selection. The first was a plot of the tuning parameter $\boldsymbol{\lambda}$ against coefficients, where the interest was to determine the optimal set of variables in the model. The second plot of the "fraction of deviance explained" and $\boldsymbol{\lambda}$ quantified the variance produced by selected variables. Graphical inspection was adopted to conclude on the mentioned plots.

3.9 Determining the best predictive model

RMSE was used to assess the best predictive model.

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{N}}$$

where, N = Number of observations, y_i= Actual observations, $\hat{y}_i$= Estimated observation, N = 160, i ≤ 6, y_i= Actual LPI scores, $\hat{y}_i$= Estimated LPI scores.

To generate interval estimates, bootstrapped sampling method was used. 160 samples were drawn with replacement from the training set and 1000 replications used to calculate the statistic.

CHAPTER FOUR

DATA ANALYSIS, PRESENTATION AND INTERPRETATION

4.1 Introduction

This chapter summarized groups and presented results according to evaluation criteria. A comparison was done and notable differences were denoted by asterisks (*). Values under comparison were denoted by **boldface**, and values that emerged different after empirical comparison were denoted by boldface and asterisk (*).

4.2 Summary descriptive statistics

Table 4.1 show that the scores range between 1.50 and 4.71. Customs score the lowest averagely and timeliness the highest. The overall LPI for 2014 is a simple mean of the variable scores.

Table 4.1 Summary statistics for Overall LPI Scores, Simple mean & Median score

overall LPI score	Simple Average	Median Score	Overall LPI score(Ex. Timeliness)
2.89	**2.89**	2.83	2.81*

Summary statistics in Table 4.2 observed timeliness as a convenient outlier. The overall LPI was higher (2.89) with timeliness in the model; the median (2.83) was closer to the mean (2.81) with timeliness excluded. In the absence of timeliness, the simple mean average was closer to the median score and was less skewed. The overall LPI score (Ex. Timeliness) was also reduced and more representative. It was also worth noting that the Overall LPI Score (World Bank LPI) was similar to the simple average.

Table 4.2 Summary statistics for determining the outlier

	Customs	Infrastructure	International shipments	Logistics quality	Tracking & tracing	Timeliness
Min.	1.50*	1.50*	1.71	1.70	1.75	1.88*
Mean	2.73	2.77	2.86	2.80	2.90	3.25*
Max.	4.21	4.32	3.82*	4.10	4.17	4.71*

* denotes the outlier

Kenya's LPI_2014 excluding timeliness was approximately similar to its median score as shown in Table 4.3. The outlier, rather than the typical value dominated the measure. The median as a measure of central tendency offered a good solution to correcting this problem.

Table 4.3 Summary statistics LPI_2014 for Kenya

Overall LPI score	Median score	Median Score (Ex. Timeliness)	Overall LPI score(Ex. Timeliness)
2.81	2.84	**2.65**	**2.64**

4.3 Assessing multicollinearity

Table 4.4 show high VIF scores (>5) for all variables. This signalled severe multicollinearity and a major problem in modelling and prediction. Most notable was the similar trend of correlation between 2014 and 2016. Logistics quality and infrastructure had the highest VIF and international shipment and timelessness the least.

Table 4.4 Variable Inflation Factor (VIF) using Overall LPI 2014 and 2016

Label	Customs	Infrastructure	International shipments	Logistics quality	Tracking & tracing	Timeliness
Overall LPI score_2014	**12.2 ***	**22.9 ***	8.2	**29.0 ***	**13.8 ***	7.3
Overall LPI score_2016	**17.3 ***	**28.5 ***	**17.7 ***	**34.9 ***	**28.5 ***	**10.0 ***

*&**boldface** denotes statistically significant (95% C.I) VIF (High multicollinearity)

This relationship was relevant in fitting and interpreting regression models.

4.4 λ selection

Choice of λ was not relevant for MLRM; however, for RRM, the precise λ that minimized RMSE for mean and median computed LPI were 0.0912 and 0.0523, respectively. E_Net (0.0424, 0.0080) and LASSO (0.0150, 0.0033) reported lower values for mean and median respectively. As λ got smaller, variance grew larger. The choice of λ played a key role in variable selection.

4.5 Variable selection using λ

Figure 4.1(a) is a plot of coefficients vs. log λ for RRM (Mean). The coefficients were closer together, meaning that mean computation assigned equal weights to the coefficients, assuming that all LPI components were equally important. As λ increased, the coefficients decreased steadily, approaching zero, and all variables retained in the model. Figure 4.1 (b) is a plot of coefficients vs. log λ for RRM (Median). Unlike RRM (Mean), the coefficients had significant weights and drew closer with increase in λ.

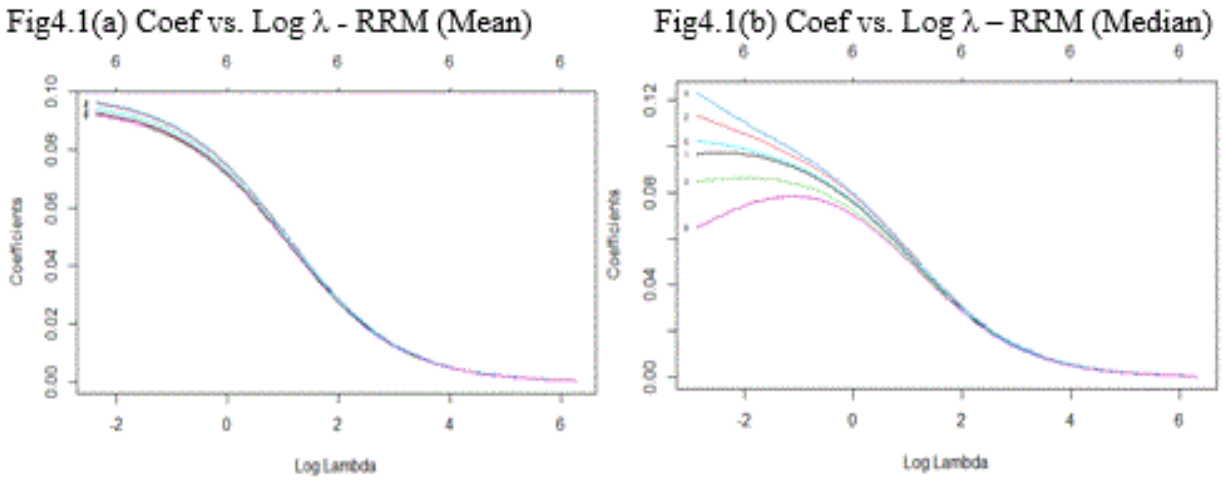


Figure 4.1 Plots for Coefficient vs. Log λ for Mean and Median computed RRM

Figure 4.2 (a) is a plot of Coefficients and Fraction Deviance Explained for RRM (Mean); the six variables explained approximately 100% of variation. Figure 4.2 (b) for RRM (Median), on the other hand, had approximately 95% of the variation explained by the six variables. The coefficients became highly inflated after that. The model started overfitting; this meant that RRM (Median) could fit about 95% of training data efficiently using all variables without overfitting. RRM (Mean), on the other hand, fitted the data almost perfectly. This model most certainly fitted random error in the data rather than the actual relationship. It reduced the generalizability outside the training set.

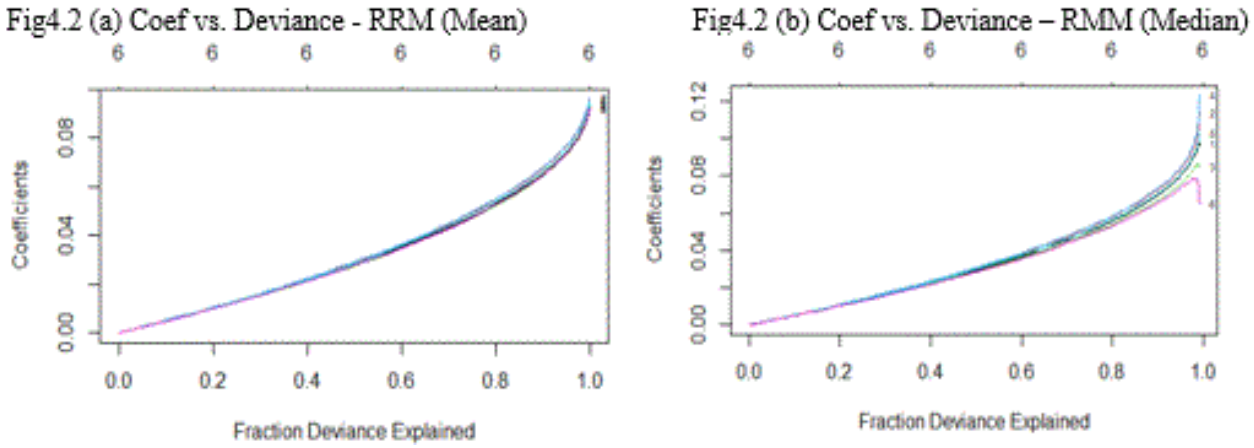


Figure 4.2 Plots for Coefficient vs Deviance for Mean and Median computed RRM

Figure 4.3 (a) is a plot of coefficients vs log λ for LASSO model computed using mean. The variables were close together, but as λ increased, the infrastructure and logistics quality coefficients increased, and the other variables decreased steadily. Beyond the optimal lambda (0.015), coefficients for timeliness, customs, international shipments and tracking were reduced to zero. This phenomenon, referred to as variable selection, forced the previous two to explain most variation. The trend was similar to Figure 4.3 (b) (the median approach), albeit with the bigger coefficient weights.

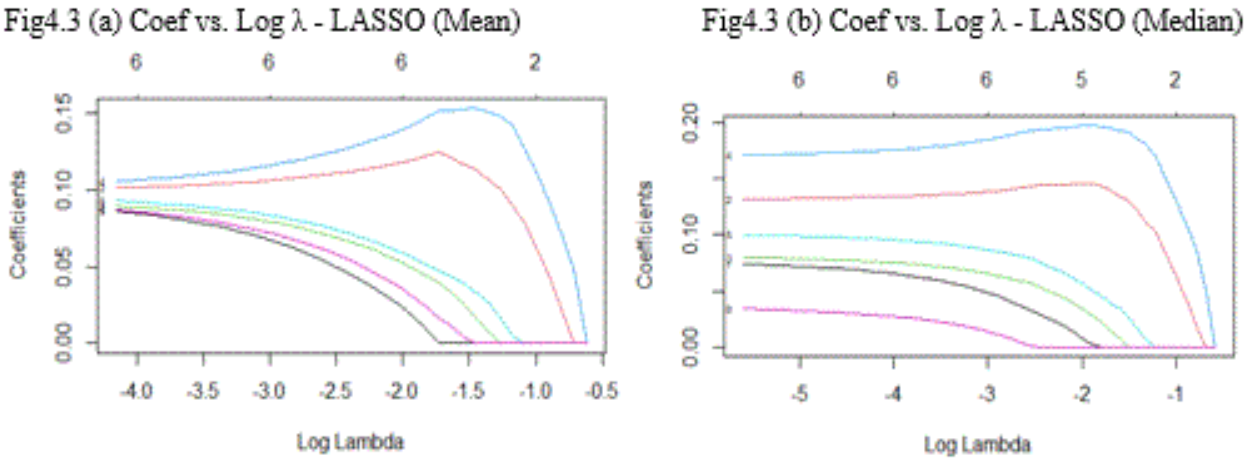


Figure 4.3 Plots for Coefficient vs Log λ for Mean and Median computed LASSO

In Figure 4.4 (a) and Fig 4.4 (b), the two dominant variables, logistics quality and infrastructure, in mean and median models, respectively, contributed significantly to above 60% variation in the models. Whereas in the mean model, five variables explained 80% of the variation in data, in the median model, three variables (logistics quality, infrastructure and tracking) explained 80% of the variation in the data. Comparing LASSO to RRM, the aspect of variable selection was evident; the final LASSO model had three variables that contributed significantly to the model based on the LPI_2014 training dataset, resulting in a parsimonious model.

Fig4.4 (a) Coef vs. Deviance - LASSO (Mean)

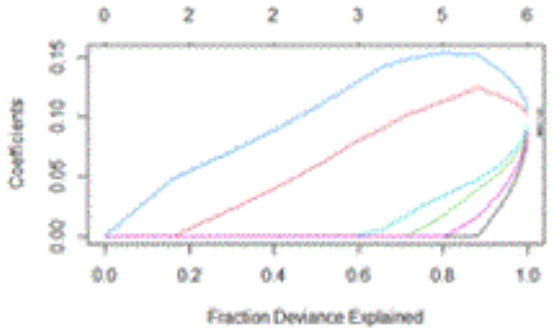


Fig4.4 (b) Coef vs. Deviance - LASSO (Median)

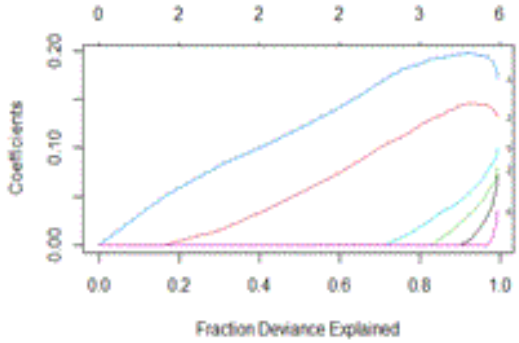


Figure 4.4 Plots for Coefficient vs Deviance for Mean and Median computed LASSO

Figure 4.5(a) and Figure 4.5 (b) are plots of Elastic Net (Mean) and Elastic Net (Median) models. There were notable differences in the weights of the coefficients. The median approach weighted the coefficients based on their importance; the mean approach assumed equal importance. Both models retained the six variables, meaning that Elastic net did not perform variable selection, instead of reducing the variable coefficients towards zero as in the case of RRM.

Fig4.5(a) Coef vs. Log λ – E_Net (Mean)

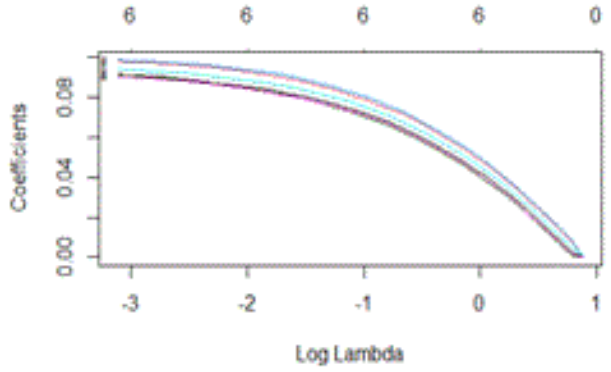


Fig4.5(b) Coef vs. Log λ – E_Net (Median)

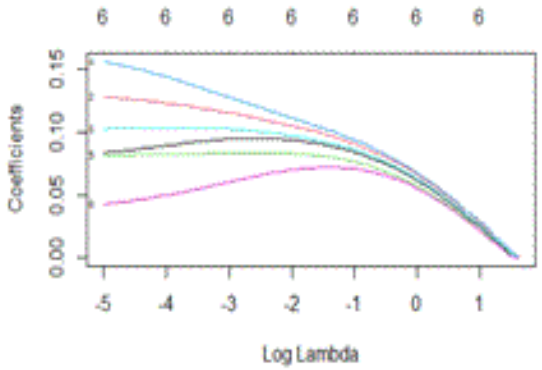


Figure 4.5 Plots for Coefficient vs Log λ for Mean and Median computed E_Net

Figure 4.6 (a) and Figure 4.6 (b) are plots of coefficients and explained deviance for Elastic Net (Mean) and Elastic Net (Median), respectively. The trends for both models were very similar to RRM. The amount of variation explained by the six variables was about 95% for the median case. After the optimal λ, the inflation and deflation indicated the possibility of over or under fitting. This model was closer to RRM than LASSO and did not perform variable selection.

Fig4.6 (a) Coef vs. Deviance – E_Net (Mean)

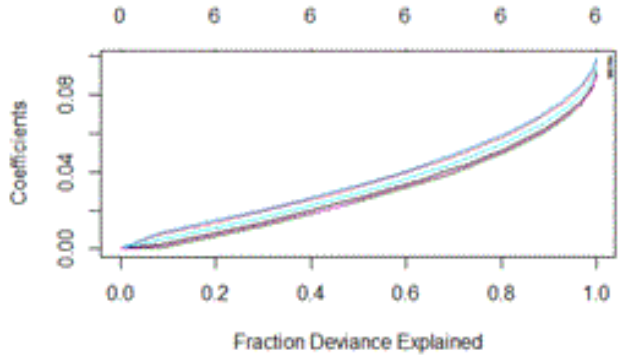


Fig4.6 (b): Coef vs. Deviance – E_Net (Median)

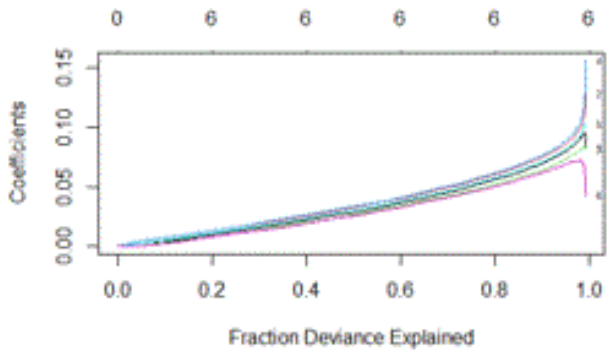


Figure 4.6 Plots for Coefficient vs Deviance for Mean and Median computed E_Net

The mean approach of computing LPI weighted variables approximately equally. It further overfitted generally in the training dataset for both Ridge regression and Elastic Net models. The median approach introduced some randomness and explained about 95% of the variation using the six variables; this was still overfitting; however, it was better than the mean approach. The LASSO

model using the median approach provided an interesting balance alluded to earlier. By generating a parsimonious model with three variables explaining 80% of the variation in data, it started to underfit beyond that limit. The next sub-section will drill down to the best training model that explained the variation in the training data better.

4.6 The best training model

Regression coefficients for MLRM, RRM, LASSO, and E_Net are provided in Table 4. 5, along with their root mean squared errors (RMSE) and AIC (in brackets). The first part compares the models based on the World Bank's mean computation approach, and the second is the proposed median approach. At 95% Confidence interval, the multiple linear regression model (MLRM) of the World Bank's LPI was near perfect (RMSE (Mean) =0.000) with all the six variables statistically significant (P-Value < 0.05). The model fitted too close to the training set, and the regression coefficients represented noise rather than the actual relationship. MLRM was still the best using the median approach, with considerable variability (RMSE (median) =0.0497). With the high multicollinearity in the data, the result pointed to the median as a safer technique to address overfitting compared to the mean. The rest of the models showed generally higher variability in the median case as compared to the mean.

Table 4.5 Comparison of regression models based on the training dataset (LPI_2014)

Mean	MLRM Coef (95% CI)	RRM Coef (95% CI)	LASSO Coef (95% CI)	E_Net Coef (95% CI)
Intercept	2.894(2.894, 2.894)	2.894(2.894, 2.894)	2.894(2.894, 2.894)	2.894(2.894, 2.894)
Customs	0.096(0.096, 0.096)	0.093(0.093, 0.093)	0.086(0.086, 0.086)	0.091(0.091, 0.091)
Infrastructure	0.099(0.099, 0.099)	0.096(0.096, 0.096)	0.101(0.101, 0.101)	0.098(0.098, 0.098)
International shipments	0.094(0.094, 0.094)	0.092(0.092, 0.092)	0.089(0.089, 0.089)	0.091(0.091, 0.091)
Logistics quality	0.099(0.099, 0.099)	0.096(0.096, 0.096)	0.106(0.106, 0.106)	0.098(0.098, 0.098)
Tracking & tracing	0.097(0.097, 0.097)	0.094(0.094, 0.094)	0.093(0.093, 0.093)	0.094(0.094, 0.094)
Timeliness	0.095(0.095, 0.095)	0.092(0.092, 0.092)	0.087(0.087, 0.087)	0.090(0.090, 0.090)
Median				
Intercept	2.869(2.812, 2.926)	2.869(2.812, 2.927)	2.869(2.813, 2.926)	2.869(2.811, 2.927)
Customs	0.076(0.030, 0.123)	0.097(0.049, 0.145)	0.074(0.028, 0.120)	0.084(0.036, 0.132)
Infrastructure	0.131(0.080, 0.182)	0.113(0.059, 0.167)	0.132(0.083, 0.181)	0.127(0.075, 0.178)
International shipments	0.081(0.037, 0.125)	0.085(0.041, 0.129)	0.080(0.036, 0.124)	0.081(0.036, 0.127)
Logistics quality	0.169(0.099, 0.238)	0.123(0.053, 0.194)	0.171(0.101, 0.240)	0.154(0.085, 0.223)
Tracking & tracing	0.101(0.042, 0.159)	0.102(0.046, 0.159)	0.100(0.041, 0.158)	0.102(0.045, 0.160)
Timeliness	0.036(-0.001, 0.074)	0.065(0.027, 0.103)	0.034(-0.002, 0.070)	0.044(0.006, 0.081)
RMSE –Mean (AIC)	**0.0000*(-4790)**	0.2338(-457)	0.3395(-337)	0.3094(-370)
RMSE –Median (AIC)	**0.0497*(-952)**	0.2570(-426)	0.3627(-318)	0.3309(-345)

*&**boldface** denotes the best model

The major differences with the models were the variance-bias trade-offs of parameter estimates. In contrast, MLRM reduced bias at the expense of variance, the other three introduced bias to reduce variance and find the optimal level of model complexity. However, as seen in Table 4.5, the penalized models could not reduce the variance better than MLRM; this meant that MLRM reduced the bias and increased the precision of the model better than any other in the training set. MLRM using the median approach emerged as the best training model.

4.7 The best predictive model

With MLRM emerging superior to the penalized regression techniques, it was subjected to a new dataset (test set) to assess its predictive power. The model, alongside the other three, were used to predict the LPI_2016.

Table 4.6 compare the prediction models based on the test set (LPI 2016). RRM, LASSO, and E_Net reduced the prediction error, but MLRM (Mean) increased it by 0.004. Mean computed LPI was generally not a robust approach as it increased the prediction error of the label.

Median computed LPI reduced the prediction error in all models. LASSO (Median) was the most precise predictive model (RMSE = 0.0436) for LPI 2016. A one-sample t-test comparing the two models showed no significant difference (t = -1.1117 (-0.0001, 0.0003); P = 0.2951, at 95% CI). The results implied that MLRM (Median) and LASSO (Median) had a similar predictive performance.

Table 4.6 Comparison of the prediction models based test dataset (LPI_2016)

Overall LPI 2016	MLRM	RRM	LASSO	E_Net
RMSE(Mean)	**0.0040***	0.0187	0.0192	0.0186
RMSE(Median)	0.0438	0.0469	**0.0436***	0.0437

Due to the proximity of the measures, all the models were likely overfitting the data. The overfitting could be attributed to the fact that LPI was a derived measure of the predictor variables. Despite this, using occam's razor principle, LASSO was still the better model. LASSO model had three predictors that accounted for 80% of the variation in the model and five that accounted for the full model; this is contrary to MLRM's six, but with similar predictive power.

Table 4.7 The best predictive model

Median	LASSO Coef (95% CI)
Intercept	2.869 (2.813, 2.926)
Customs	0.074 (0.028, 0.120)
Infrastructure	0.132 (0.083, 0.181)
International shipments	0.080 (0.036, 0.124)
Logistics quality	0.171 (0.101, 0.240)
Tracking & tracing	0.100 (0.041, 0.158)

4.8 Variable Selection

Figure 4.7 show the ranking of variables based on their importance. LASSO ranked logistics quality and competence the highest, followed by infrastructure and tracking & tracing. International shipments and customs ranked closely, and timeliness was not significant in the model.

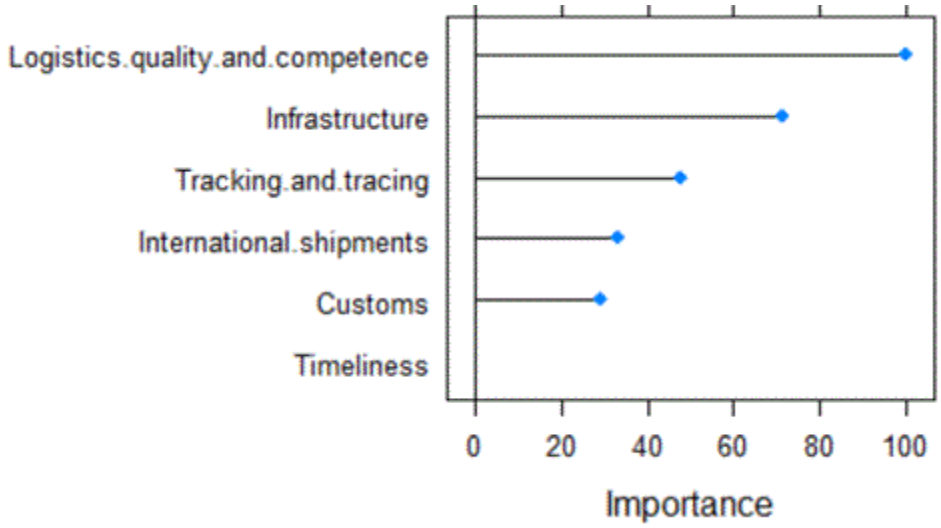


Figure 4.7 Variable Importance using LASSO (Median) model

CHAPTER FIVE

SUMMARY OF FINDINGS, DISCUSSIONS AND CONCLUSIONS

5.1 Summary of findings

In an attempt to develop a robust predictive model for Logistics Performance Index (LPI), using the LPI 2014 dataset, several results emerged consistent with the literature.

The LPI 2014 dataset was a well-organized small dataset with 160 subjects and six variables, with no missing values and scores ranging between one and five. This thesis referred to it as a "low-dimensional and non-sparse dataset". The LPI developed by The World Bank using PCA for weighting the variables turned out to be simple mean averages of the variable scores. Outliers affected LPI heavily; for instance, removing timeliness, a consistent outlier, changed the measure significantly. The best replacement for the mean was the median. Like the mean, it is a measure of central tendency but not affected by outliers. The weights assigned to the variables were also approximately similar under the mean computation, meaning that all the variables affected LPI equally; the median, on the other hand, had significantly different weights, signifying variation in the component effects.

Descriptive analysis of the variable scores in LPI 2014 and LPI 2016 datasets produced severe correlation. From the VIF outputs, any random variable sufficiently predicted the LPI. This multicollinearity was consistent in LPI 2014 and LPI 2016, resulting in a similar correlation structure between the two years. This similarity affected LPI Predictability irrespective of the computation approach (mean or median). The exact effect was not directly on the overall LPI predictability but the individual components.

The best predictive model was one that produced the least prediction error for LPI. Used as the base model, the multiple linear regression model was the most precise; however, it was the least in prediction. The least absolute shrinkage and selection operator was the best in prediction; however, a one-sample paired test produced a statistically insignificant difference between the two models.

Whereas the multiple linear regression model used all the six variables in modelling, the least absolute shrinkage and selection operator used five to produce a slightly better predictive model.

It further ranked logistics quality, infrastructure and tracking as the most relevant variables accounting for 80% variability in the model.

5.2 Discussions

LPI is a tool that is critical in benchmarking the performance of trade logistics; the key advantages are that the inputs (the six variables) are directly observable and measurable, giving significant insights into the opportunities for development. Cemberci et al. (2015) concluded that Improving timeliness, tracking and tracing, and international shipments increased the global competitive index (GCI). Marti et al. (2014) studied the significance of the LPI components in emerging economies using the gravity model. All the components scores had positive correlations with the international trade performance. Civelek et al. (2015) regressed LPI with GCI and GDP. He first investigated the pairs, and the result was a statistically significant relationship between them. The multiple regression between LPI, GCI and GDP was also significant; this shows that LPI and the components are relevant for measuring global performances.

Instinctively, all six components affect the overall rating of LPI differently; this means that the components ought to have different weights based on their significance. The World Bank computed LPI give approximately similar component weights: Customs (0.40); Infrastructure (0.42); International shipments (0.40); Logistics quality and competence (0.42); Tracking and tracing (0.41) and Timeliness (0.40). The resulting LPI is a simple average of the variable scores. It is affected by outliers, and in this context, an improvement in one variable generally lead to better performance and vice-versa. Secondly, logistics is a system of interrelated and interconnected components, one part affects another, and a few critical parts hold the system together. World Bank's LPI assumes that the parts are mutually exclusive and disjoint, which is a very simplistic view. In LPI 2014 dataset, all the variables are correlated to some extent, as shown by the VIF scores. The assumption of independence is wrong.

A simpler yet logical assumption is to weigh variables according to some criteria. Jafar, Wilco & Lorant (2020) developed a model for measuring the relative importance of LPI indicators using the Best Worst Method (BWM). However, because of the correlation between the variables, the effect was only mild. Multicollinearity further affected prediction. Median LPI computation for controlling outliers and penalized regression approach for weighting and prediction are possible alternatives. Penalized approaches assign weights to variables in a regression model, considering

the level of variable correlation and their importance in the model. The more critical a variable is, the bigger the weight.

In this paper, LASSO predicted the future precisely better than any model; it was most importantly, able to rank variables based on their importance. Policymakers can skew the scarce resources towards important inputs. The aspect of maximizing the value of a dollar is also very important. LASSO rides on this premise; it predicts 80% of the variation in LPI using three variables only. Policymakers can focus resources on the three variables efficiently.

Penalized models have been used extensively in "high dimensional and sparse" datasets with a considerable positive outcome in data analysis. LPI was a "low dimensional and non-sparse" traditional and conservative form of the dataset; however, LASSO performed very similarly to the former case. It is, therefore, safe to say that based on the LPI dataset analyzed in this paper, sparse (penalized) regression models would perform consistently irrespective of the structure (Wide or tall) of the dataset.

5.3 Conclusion

While weighting the variables for a robust metric is critical, PCA does not provide any form of weighting through its loadings. Policymakers cannot decide on which areas to prioritize and skew resources. On the other hand, the LASSO model selects variables, and a clear direction can be achieved for policymakers. In this paper, improving the skills of employees in the logistics sector and increasing general visibility of products on transit through autonomous and intelligent technologies in a quality physical infrastructure is critical to making the country globally competitive.

5.3.1 Limitations

The validity of data collection tool or reliability of various sources of information could not be verified. The median computed LPI did not weigh the variables, it only controlled for outliers.

5.3.2 Further Research

A replication study on the performance of penalized models in “low-dimensional non-sparse” datasets may validate this research further. Using a differently weighted LPI rather than median as a label should be explored further. In this case, the median was used as an alternative to the mean (The World Bank LPI) to control outliers. Further research should explore alternative prediction

techniques like neural network with Tensorflow, random forest, support vector machine and gradient boosts.

REFFERENCES

Alexandre Belloni, V. C. (2014). Inference on Treatment Effects after Selection amongh High-Dimensional Controls. *The Review of Economic Studies*, 42.

Bank, W. (2010). *Trade and Transport Facilitation Assessment: A practical Toolkit for Country Implementation.* World Bank Publications.

Chai, T., & Draxler, R. (2014). Root mean square error (RMSE) or mean absolute error (MEA). *Geoscientific Model Development Discussions* (p. 1). ResearchGate.

Development, O. f.-o. (2005). *OECD Economic Studies.* Berkeley: Organisation for Economic Co-operation and Development.

Frank, I., & Friedman, J. (1993). A Statistical view of some chemometrics regression tools. *Technometrics*, 35.

Ghojogh, B., & Crowley, M. (2019). Unsupervised and Supervised Principal Component Analysis: Tutorial. *Arxiv*, 25.

Hallin, M. (2014). Gauss-Markov Theorem in Statistics. *ResearchGate*, 4.

Holland, S. M. (2019). *Principal Components Analysis (PCA).* Athens: Department of Geology, Univerisity of Georgia.

Joshi, S. (2015). *Designing and Implementing Global Supply Chain Management.* California: IGI Global.

Kalton, G., & Flores-Cervantes, I. (2003). Weighting Methods. *Journal of Official Statistics*, 81-97.

Kennard, & Hoerl. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12.

Mundfrom, D., & Smith, M. L. (2018). The Effect of Multicollinearity on Prediction in Regression Models. *ResearchGate*, 6.

Pham, H. (2019). A New Criterion for Model Selection. *Rutgers*, 12.

Sclove, S. (1963). Improved estimators for coefficients in linear regression. *Journal of the American Statistical Association*, 63.

Shrestha, N. (2020). Detecting Multicollinearity in Regression Analysis. *American Journal of Applied Mathematics and Statistis*, 42.

Sober, E. (1981). The principle of Parsimony. *The British Journal for the Philosophy of Science*, 32.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Royal Statistical Society*, 22.

Trevor Hastie, R. T. (2013). *The Elements of Statistial Learning: Data Mining, Inference, and Prediction.* Springer Science.

World Bank. (2014). *Logistics Performance Index (LPI) Report: The Gap Persists.* Washington: The World Bank. Retrieved from Worldbank.org.

World Bank. (2016). *Connecting to Compete: Trade Logistics in the Global Economy.* Washington: World Bank.

World Bank. (March 20, 2014). *Logistics Performance Index (LPI) Report: The Gap Persists.* Washington: The World Bank.

X.N. Iraki, B. L. (2018). *24-hour Economi: A new Frontier in Kenya's Economic Development.* National Economic and Social Council.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, series B*, 67.

APPENDICES

APPENDIX A: Similarity (Originality) Report

Document Information

Analyzed document	Thesis-Draft.docx (D103814613)
Submitted	5/4/2021 10:36:00 PM
Submitted by	
Submitter email	Eric.Odok@strathmore.edu
Similarity	4%
Analysis address	library.strath@analysis.urkund.com

APPENDIX B: Ethical Clearance Confirmation

Strathmore UNIVERSITY

22nd November 2021

Mr Odok Eric,
eric.odok@strathmore.edu

Dear Mr Odok,

RE: Predictive Modeling of Kenya's Logistics Performance Index using Least Absolute Shrinkage and Selection Operator

This is to inform you that SU-IERC has reviewed and **approved** your above SU-master's research proposal. Your application reference number is **SU-IERC1045/21.** The approval period is **22nd November 2021 to 21st November 2022.**

This approval is subject to compliance with the following requirements:

i. Only approved documents including (informed consents, study instruments, MTA) will be used
ii. All changes including (amendments, deviations, and violations) are submitted for review and approval by SU-IERC.
iii. Death and life-threatening problems and serious adverse events or unexpected adverse events whether related or unrelated to the study must be reported to SU-IERC within 48 hours of notification
iv. Any changes, anticipated or otherwise that may increase the risks or affected safety or welfare of study participants and others or affect the integrity of the research must be reported to SU-IERC within 48 hours
v. Clearance for export of biological specimens must be obtained from relevant institutions.
vi. Submission of a request for renewal of approval at least 60 days prior to expiry of the approval period. Attach a comprehensive progress report to support the renewal.
vii. Submission of an executive summary report within 90 days upon completion of the study to SU-IERC.

Prior to commencing your study, you will be expected to obtain a research license from National Commission for Science, Technology, and Innovation (NACOSTI) https://research-portal.nacosti.go.ke/ and obtain other clearances needed.

Yours sincerely,

or: Prof Fred Were,
Chairperson; SU-IERC

STRATHMORE UNIVERSITY INSTITUTIONAL ETHICS REVIEW COMMITTEE (SU-IERC)
22 Nov 2021
TEL: +254 (0)70303400
P.O BOX 59857-00200
NAIROBI-KENYA

APPENDIX C: Thesis Correction Form

Strathmore University

Office of Graduate Studies

THESIS CORRECTION FORM

Name of Candidate:	Student Number:
Eric Oyenga Odok	114865
Faculty/School/Institute: Mathematical Sciences	Degree: MSc (Statistical Science)
Title of Thesis: Predictive Modelling of Logistics Performance Index using Sparse Regression Models	

Summarise the types of corrections done in your thesis (*Attach a detailed report*).

(1) Amendment of thesis title

(2) Re-writing the abstract to make it concise and direct

(3) Enriching the discussion section with more details and comparing it with previous results from other literatures. Aligning the language in the entire paper to appear more scientific.

(4) Putting the abbreviations page in the right place, just before the abstract. Placing full stops and commas appropriately

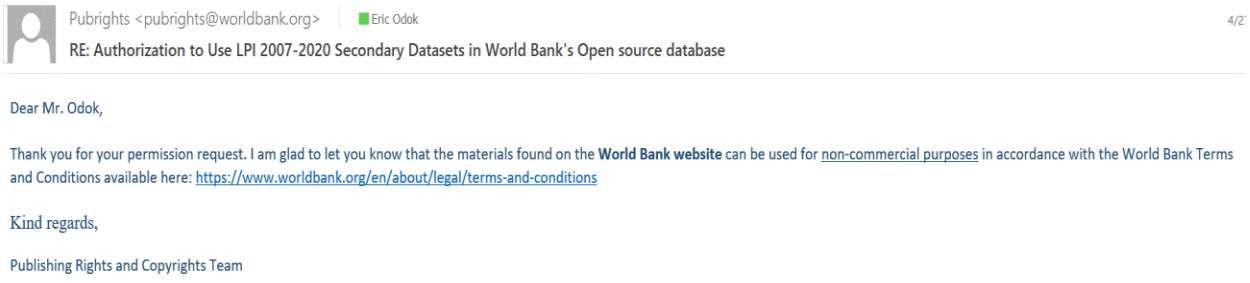
(5) Avoiding repeating Logistics performance index (LPI) and instead using LPI alone throughout the paper. Also aligned on sentence construction and grammar, then included list of tables and list of figures exhaustively

Committee Members:

Principal Supervisor Prof. Samuel Mwalili	Signature:	Date: 10/11/2021
Co-Supervisor	Signature:	Date: 16/10/2021
Internal Examiner	Signature:	Date:
Director of Graduate Studies	Signature:	Date:

SGS-04-13-04/11

APPENDIX D: Authorization to Use LPI 2007-2020 Secondary Datasets in World Bank's Open source database



Pubrights <pubrights@worldbank.org> Eric Odok 4/2

RE: Authorization to Use LPI 2007-2020 Secondary Datasets in World Bank's Open source database

Dear Mr. Odok,

Thank you for your permission request. I am glad to let you know that the materials found on the **World Bank website** can be used for non-commercial purposes in accordance with the World Bank Terms and Conditions available here: https://www.worldbank.org/en/about/legal/terms-and-conditions

Kind regards,

Publishing Rights and Copyrights Team