Survival of the Fittest: Variable Selection on Agricultural Data from the Galápagos Islands

Michael Bostwick
Department of Statistics and Operations Research
University of North Carolina at Chapel Hill

Client: Francisco Laso
Department of Geography
University of North Carolina at Chapel Hill

March 28th, 2018

Abstract

Variable selection is an important first step when analyzing datasets with a large number of potential predictor variables. We apply two techniques, Forward Selection and Elastic Net, to find the most important variables in a dataset detailing over 200 socioeconomic measurements for 755 farms on the Galápagos Islands. Modeling five different outcome variables, we find the data available has the strongest linear relationships with the outcomes *productivity* and *land use choices*. For each of the outcome variables we present the top five predictor variables as well as a full set of coefficients for the optimally predictive model.

1 Introduction

1.1 Background

The Galápagos Islands make for a feasible and significant case study of complex systems. Due to their relative isolation and smaller size, the interaction of factors can more realistically be modeled for the Galápagos Islands than other systems. Yet, the Galápagos Islands also represents an important example of the competing forces of resource conservation and economic development in a rapidly changing environment. Prior work has created agent-based simulation models of the Galápagos, but with limited interaction parameters between agents, particularly in regards to farm success ([6], [7],[8]). In order to create a more detailed and perhaps more accurate simulation, the relationships between different factors on the island must be better understood. This work aims to search through a large number of possible relationships and identify the empirically most significant ones for future study and incorporation into simulation models.

1.2 Data

The data available to study the dynamics between agricultural measures and related factors primarily comes from the Censo de las Unidades de Producción Agropecuaria (UPA) de Galápagos (Census of Agricultural Production Units (UPA) of Galapagos) ([1]). This is a self-reported survey with data from 755 farms (UPAs) detailing many characteristics. The questions covered fall into the following categories:

- General characteristics (land area, age of landowner, etc.)
- Permanent crops (specific types and quantity)

- Temporary crops (specific types and quantity)
- Pastures (specific types and quantity)
- Tree crops (specific types and quantity)
- Livestock production (detailed by animal)
- Expenses (detailed by category)
- Workers (detailed by role)
- Land use (8 different categories)

In addition to this census, data is also available from satellite image classification including information on water, energy and road access. In total, under the direction of the client, 239 variables were selected for consideration in modeling relationships between potential predictors and five outcome variables of interest.

Some of the five outcome variables of interest came directly from survey responses, while others were derived from a combination of multiple variables. When analyzing a derived outcome variable all variables used in its original calculation were removed from consideration in the model. The client also denoted specific predictor variables to exclude from particular models when their inclusion would not be beneficial. For example, while the amount of crops sold in pounds was not directly used to calculate *net income*, the obvious relationship existent precluded it from inclusion. In addition, predictor variables that met one or more of the following criteria were removed prior to modeling: zero variance, extremely high (>0.99) or perfect correlation with other predictor variables, or linear dependence with other predictor variables (that is, two or more predictor variables could be linearly combined to create another predictor variable). The exact number of predictor variables included in each model varied slightly, but there were approximately 200 predictors variables examined for each model after the preceding steps were taken.

1.3 Organization of Report

The remainder of this report is divided into four sections. Section 2 provides a brief overview of the analysis so that the results can be understood. In Section 3 results for each of the 5 outcome variables are provided, where standard tables and graphs are repeated for each. A more in-depth explanation of the statistical methods used is contained in Section 4, but this section can be referenced as needed. Section 5 details important considerations when interpreting the results and suggests possible avenues for future work. Lastly, References and the Appendix, including additional tables and figures, can be found following Section 5.

2 Modeling

2.1 Challenges to address

The primary challenge in this analysis is the vast number of potential predictor variables. This challenge is twofold; 1.) when the number of predictors is large the determination of a reliable model is difficult and 2.) interpreting the coefficients of many predictors simultaneously is not an easy task for humans (and will make resulting simulations overly complicated). For this reason, the analysis focuses on the use of two variable selection techniques that aim to build a linear model with a subset of the available variables that still maintains a strong explanatory/predictive performance.

Secondarily, when performing standard linear regression the error is assumed to be normally distributed. While this does not mean the outcome variable necessarily needs to be normally distributed, large deviations from normality can cause issues. Several outcome variables in this study show strong non-normality, which can contribute to a poorly fitting model and unreliable estimates of the coefficients. In order to address this issue transformations to the data and modifications to the standard linear model were considered.

2.2 Overview of methods

A summary of the statistical methods used is presented here to allow for understanding of results. For further details see Section 4: Statistical Methods. For each of the outcome variables of interest

a set of linear models using the appropriate subset of predictors was built. Each relationship was modeled using Forward Selection and Elastic Net regression. Forward Selection fits a linear model by progressively adding variables to the model until a best fit is found. This results in only some of the variables being included, chosen in a discrete manner (computed using the R package 'leaps' [5]). Elastic Net regression fits a linear model by limiting the size of the coefficients so that they are smaller than in standard least squares, and for many variables actually shrunken to zero. Similar to Forward Selection this results in a smaller model, but variable selection can be carefully tuned as optimization is done in a more continuous way (computed using the R package 'glmnet' [3]).

In general, these techniques have slightly different aims. Forward Selection chooses a model that best explains the variance in the dataset at hand. Elastic Net chooses a model that can best make predictions on new data. Depending on the goals of analysis, one technique is not necessarily better so we do not compare the two quantitatively, but instead offer both results as varying perspectives on variable selection. While a variable being chosen by both methods provides stronger evidence that an important relationship exists, disagreement suggests exploring both possibilities instead of one method necessarily being incorrect.

3 Results

The Results section is broken into three subsections covering the outcome categories of interest: farm success, invasive species and land use choices. Within farm success we further divide into three specific measurements, resulting in a total of 5 different outcome variables. We explore each outcome variable in turn, presenting a visualization of the data, the top five predictor variables and analysis of optimal linear models.

3.1 Farm Success

The first three outcome variables of interest can be grouped together under the category of farm success:

- *Productivity*: total pounds of crops and livestock produced divided by the farm surface area.
- *Net income*: revenue from all products sold minus total expenses.
- *Number of workers supported*: total labor expenditures divided by a standard full-time worker's salary.

3.1.1 Productivity

The histograms of the *Productivity* variable (Figure 1) show a strong skewness, both when looking at all observations, and when zooming into observations between 0-10,000 lbs/hectare. To achieve a distribution closer to normal (bell-curved), which will benefit the linear model, we took the log_{10} transformation with resulting data shown in the last plot. Since the log transformation cannot be performed on zero values, we removed the 38 occasions of this from the dataset. Beyond the mathematical constraint, farms with zero production perhaps are not farms as typically defined.

We built linear models using both methods, Elastic Net and Forward Selection, on the log-transformed *Productivity* variable, recording an optimal model of any size and the best 5 variable model for each. The size of 5 variables is chosen to provide quick insight and not based on any specific statistical property. The results of the best 5 variable model are shown in Table 1, listed in order of entrance into the model. Next to variable names the direction of the relationship is indicated with a (+) or (-).

Elastic Net	**Forward Selection**
percpasture (-)	cantonSan Cristobal (+)
pc4None (+)	percinv (-)
v30a (-)	percperm2 (+)
percperm (+)	percbrush (-)
pc6 (-)	CPermanentesPAPAYA (+)

Table 1: *Modeling of Productivity, Top 5 features for both methods*

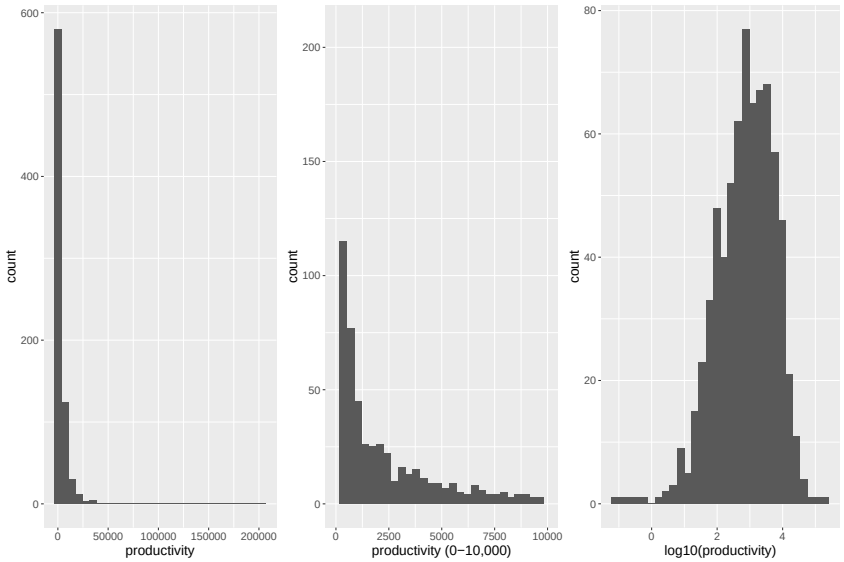


Figure 1: *Histogram of Productivity, from left to right showing the full range, zoomed into 0-10,000 lbs/hectare, and the log transformed nonzero values. The log transformed values exhibit the desired normal shape.*

Plots from the optimal models for Elastic Net and Forward Selection are shown in Figures 2 and 3. The cross- validation plot for Elastic Net can be understood as follows: the horizontal axis shows the number of variables included in the model (on top) and the corresponding lambda (λ) value (on the bottom), the vertical axis shows the Mean-Squared Error (MSE) represented as the red dots and surrounded by bars showing the standard deviation. The vertical dashed line to the left, λ_{min}, is found at the minimum MSE and the vertical dashed line to the right, λ_{1se} is at the largest lambda within one standard error of the minimum. The idea behind λ_{1se} is that similar error performance can be achieved with a smaller model, in this case a model with 40 fewer variables. Since our goal is to select a small amount of variables, we will generally use the model found at λ_{1se}.

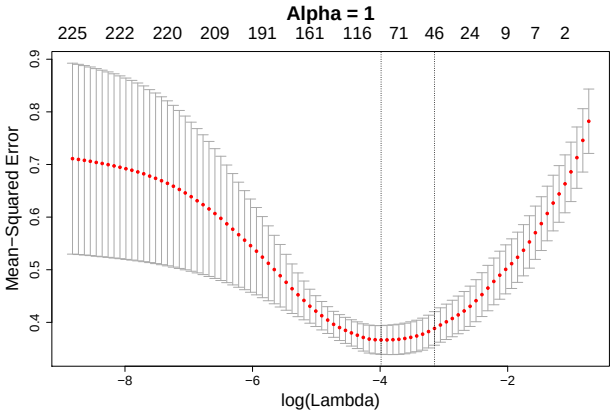


Figure 2: *Elastic Net Cross-validation plot for Productivity. The plot shows a desirable "U" shape where a moderate number of variables provides a significant improvement in prediction error.*

For Elastic Net we measure performance with Root Mean-Squared Error (RMSE), which is on average how far the predicted value is from the actual value across cross-validation runs. The 5 variable Elastic Net model had a RMSE of 0.78. Not counting the intercept term, the optimal model chosen for Elastic Net included 45 variables and had a RMSE of 0.61. To provide context for the RMSE, we can look at the point furthest to the right on the cross-validation plot in Figure 2 and see how the model would perform when no variables are included in the model, that is, just predicting the average outcome value.

For Forward Selection the plot, Figure 3, is much more straightforward. We plot the number of variables included versus the Bayesian Information Criterion (BIC) that we choose to minimize, and mark the optimal point in red. The convex shape of the plots highlights a common trend in variable selection; not including enough variables does not provide enough information, but beyond a certain point adding more variables may not be worth the added complexity.

We evaluate Forward Selection with R^2, which is the percentage of variability in the outcome variable that is explained by the model. The 5 variable Forward Selection model had an R^2 of 0.48 and the full model included 26 variables and had a R^2 of 0.63.

The coefficients estimated for both full models can be found in Table 9 in the Appendix. These numbers suggest that while the 5 variable model is helpful, there is a decent amount of information to be gained by adding more variables. Diagnostics of the linear fit of the optimal Elastic Net and Forward Selection models (plots shown in Figure 15 in the Appendix) do not raise any concerns.

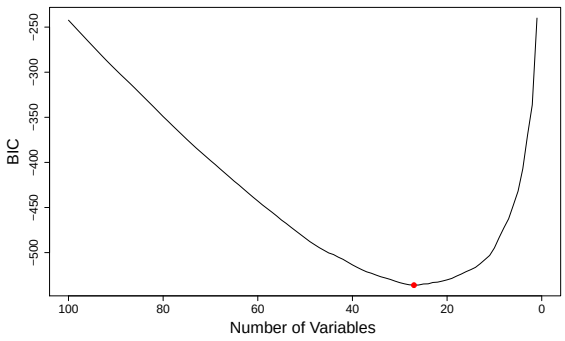


Figure 3: *Forward Selection plot for Productivity. The optimal model includes 26 variables.*

3.1.2 Net Income

The histograms for *Net Income* (Figure 4) show a symmetric shape, but a very spiky center and a few observations wide in the tails. We attempted to fit a model on the full dataset, but find the observations with large absolute values are obscuring other possible information in the model. After trying several cutoff thresholds and examining model fit, we removed the 28 observations that are outside the middle 95% of the data. The following models are fit on this reduced dataset of 727 observations.

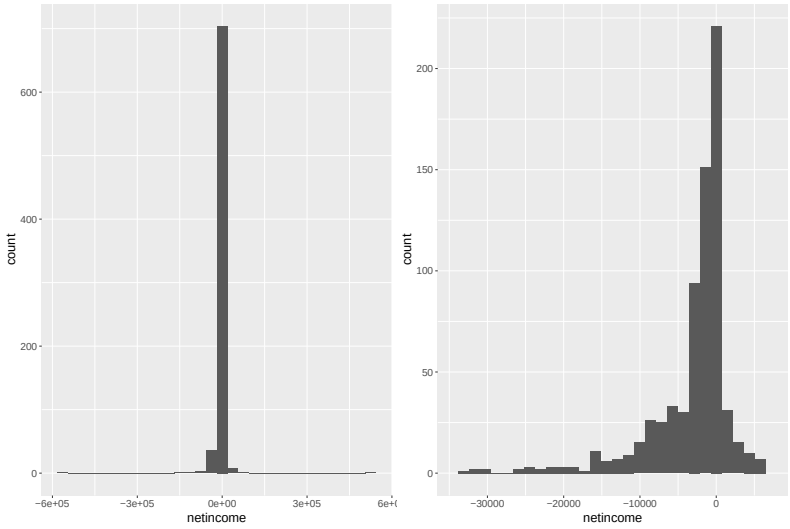


Figure 4: *Histograms of Net Income, with full data and the central 95% of data displayed. The values very far from the center are of concern when modeling.*

Elastic Net	**Forward Selection**
v3 (-)	cantonSan Cristobal (+)
v45 (-)	v30a (-)
produccionenlibrasproductocosechadoautoconsumo (+)	v3 (-)
CATTLETRUE (-)	AGUAAGUA POTABLE PUBLICA (+)
percperm2 (+)	v53a (+)

Table 2: *Modeling of Net Income, Top 5 features for both methods*

The results of the best 5 variable model are shown in Table 2, with a RMSE of 5790.35 for Elastic Net. For interpretation of the RMSE it is important to keep in mind the scale for *Net Income* is much

larger than that of log productivity. However, in this case adding variables has caused a minimal decrease in the error. The cross-validation plot for *Net Income* show wider error bars throughout the range of model sizes and just using the average *Net Income* would predict nearly as well as any other model. Since here λ_{1se} only includes the intercept, we chose the optimal Elastic Net model to be at λ_{min}, which included 9 variables and had a RMSE of 5768.30.

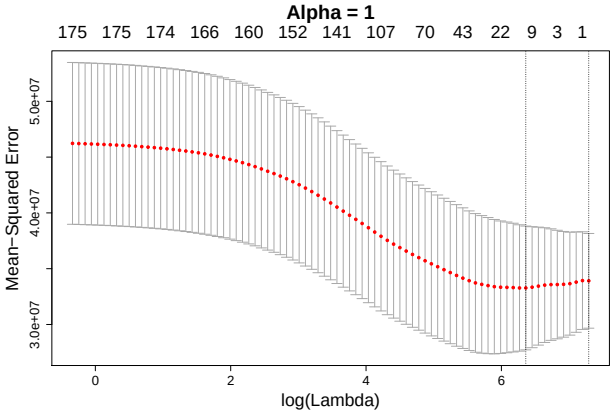


Figure 5: *Elastic Net Cross-validation plot for Net Income. The plot shows a much smaller improvement in prediction error.*

For Forward Selection the 5 variable model and full model had R^2 values of 0.15 and 0.19, respectively. The full Forward Selection model included 9 variables, with all but two overlapping with the Elastic Net choices. The coefficients estimated for both models can be found in Table 10 in the Appendix.

Since the full models are not much larger than the 5 variable models, the small improvements are not surprising. Diagnostics of the linear fit of the optimal Elastic Net and Forward Selection models (plots shown in Figure 16 in the Appendix) do not follow assumptions as closely as for the Production models, but are not so concerning as to disqualify the results. On the whole, the results from the various plots and diagnostics suggest that the relationships found for Net Income are worth investigating, but that a linear relationship provides little predictive power.

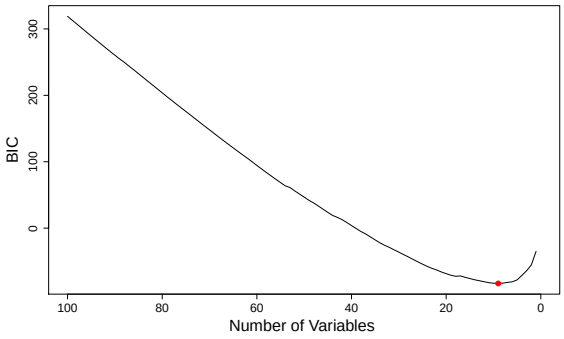


Figure 6: *Forward Selection plot for Net Income, which includes a small number of variables.*

3.1.3 Number of Workers Supported

The first plot in Figure 7 shows that a large percentage (65%) of the farms are not able to support any workers. For this reason, we divided up the modeling task for *Number of Workers Supported*. First, we used logistic regression to model the binary variable of whether a farm supports zero or more than zero workers. Secondly, we used linear regression to model the quantity of workers for just those 264 farms with a positive number of workers supported.

The top 5 variable models for logistic regression and linear regression can be found in Table 3 and Table 4, respectively. Sometimes Elastic Net will simultaneously choose to eliminate multiple variables, in this case there is not a 5 variable model so we show the 6 variable model for Elastic Net. For logistic regression we can measure performance with the misclassification rate, which on average was 0.31 on the test set for Elastic Net, compared to the 0.35 misclassification rate we would achieve if we simply predicted the majority class, zero workers, every time. The full Elastic Net logistic regression model had 6 variables and the same 0.31 misclassification rate. For Elastic Net

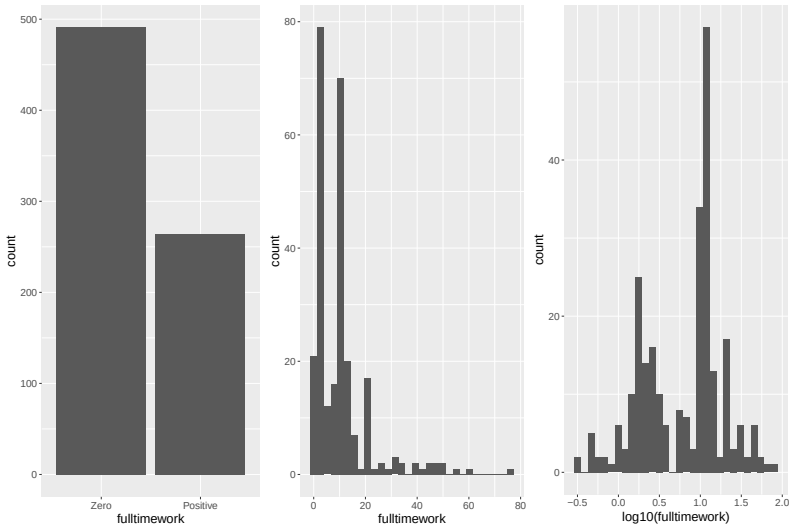


Figure 7: *Histograms of Workers, showing the proportion of zeroes, farms with nonzero workers, and the log transformed number of workers. The log transformation helps remove the skewness, but some bi-modality still remains.*

Elastic Net	**Forward Selection**
v3 (+)	v3 (+)
s4 (+)	GastosPecuarios (+)
v45 (+)	VentaLibras (+)
GastosPecuarios (+)	librasvendida (+)
CATTLETRUE (+)	perctemp2 (-)
CosechaLibras (+)	

Table 3: *Modeling of Binary Workers , Top features for both methods*

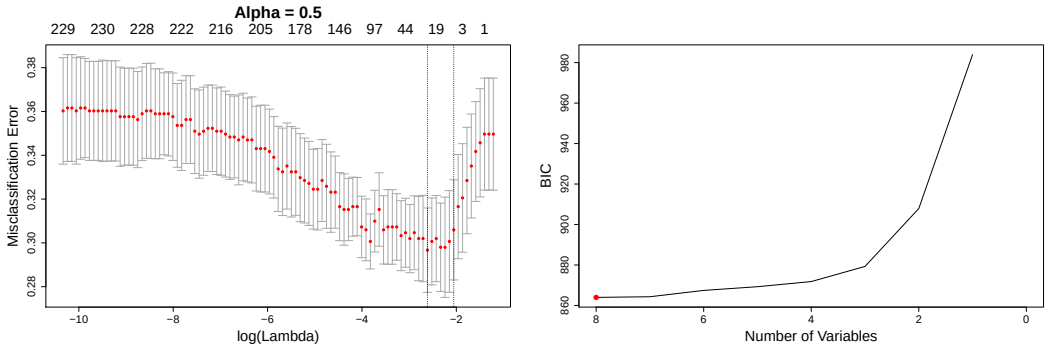
Elastic Net	**Forward Selection**
s4 (+)	cantonSan Cristobal (-)
v3 (+)	s9 (+)
biokSi (+)	ReclassCONSERVACION (-)
GastosAgricolas (+)	GastosPecuarios (+)
v44 (+)	ArbolesLIMON REAL (+)

Table 4: *Modeling of Nonzero Workers, Top 5 features for both methods*

linear regression, the 5 variable model resulted in a RMSE of 0.46 and the full model also included 5 variables, and therefore the same RMSE.

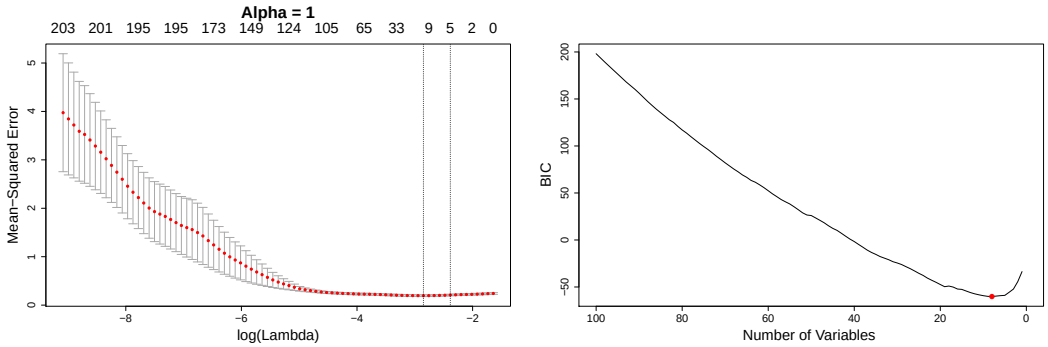
For logistic regression Forward Selection the 5 variable model had a misclassification rate of 0.26, which is measured on the original dataset, not a test set. The optimal Forward Selection logistic regression chose a model of size 7 also with a misclassification rate of 0.26. For linear regression, the 5 variable Forward Selection model had a R^2 of 0.34 and the optimal model had a R^2 of 0.34 using 8 variables.

In the Appendix, the coefficients estimated for logistic regression models can be found in Table 11 and the coefficients estimated for linear regression models can be found in Table 12. The fairly low misclassification rate coupled with low R^2 show that there is a fairly clear divide between farms that can or cannot support workers, but the specific number of workers is much harder to predict. Several variables are found to be most helpful for both the logistic regression and linear regression models, but there is still a fair bit of difference. One common theme of the variables chosen here is that many are related to cattle production.



(a) *Cross-validation plot for Elastic Net* (b) Forward Selection plot

Figure 8: *Workers Binary Variable Selection. The Elastic Net plot shows ideal predictive performance when including 5-25 variables. The computation for logistic regression Forward Selection stops at the optimal value so the rest of the plot is not included as before.*



(a) *Cross-validation plot for Elastic Net* (b) Forward Selection plot

Figure 9: *Nonzero Workers Variable Selection. Both plots show that adding variables does little to improve performance, highlighting a lack of linear relationship.*

3.2 Invasive Species

The analysis of *Invasive Species* follows much the same pattern as for *Number of Workers Supported*. Again, we have a large number of zeros in the outcome variable, farms with zero percent of their land covered by invasive species. We first modeled this binary variable using logistic regression and then performed linear regression on the log transformed values for the farms that have a percent coverage greater than zero. As can be seen in the far right plot of Figure 10 there are two very small values (0.005% and 0.08% prior to taking the log). They effect the fit of the model and are negligibly above zero so we removed them from the linear model, leaving 155 farms for modeling.

Elastic Net	**Forward Selection**
cantonSanta Cruz (-)	cantonSan Cristobal (+)
ABANDONEDTRUE (+)	cantonSanta Cruz (-)
cantonSan Cristobal (+)	cantonFloreana (-)
pc4None (-)	v30a (+)
v30a (+)	CTransitoriosMAIZ SUAVE CHOCLO (+)
CTransitoriosMAIZ SUAVE CHOCLO (+)	

Table 5: *Modeling of Binary Invasive , Top features for each method*

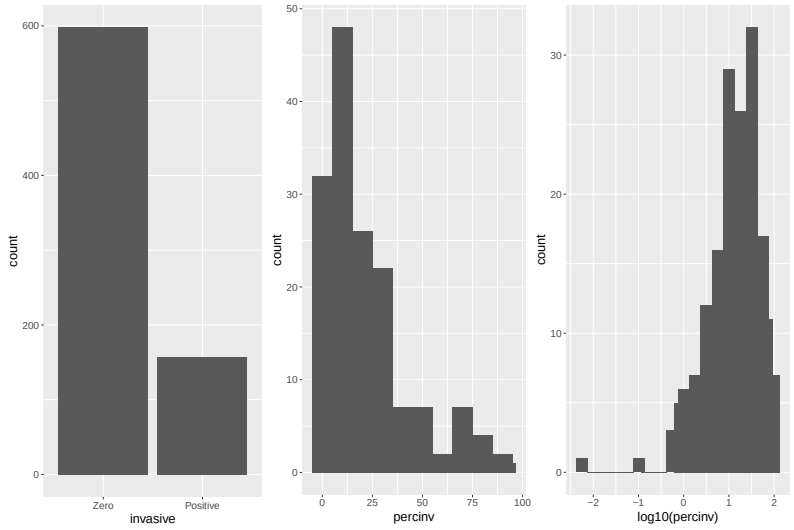
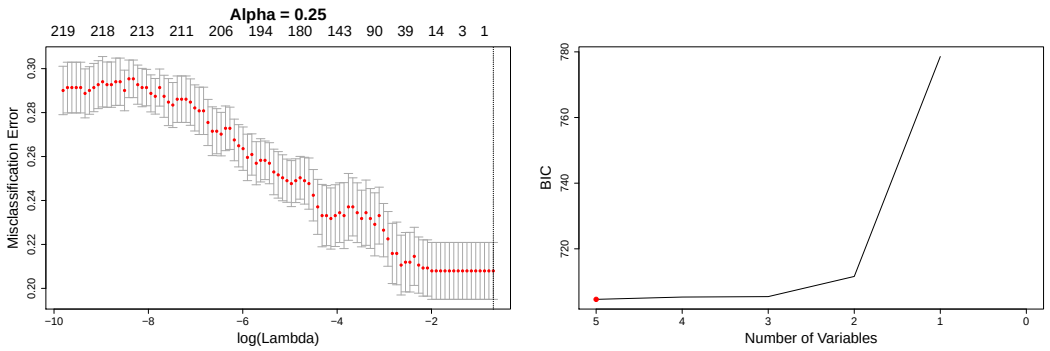


Figure 10: *Histogram of Invasive Species. To the left: the proportion of zeroes, in the middle: farms with greater than zero, and to the right: the log transformed percent of invasive species. The two very small values on the right plot are removed.*

Elastic Net	**Forward Selection**
PastosKING GRASS (-)	cantonSan Cristobal (+)
pc4None (+)	PastosMIEL O SETARIA (+)
cantonSan Cristobal (+)	CTransitoriosTOMATE RINON (-)
CTransitoriosSANDIA (-)	pc6 (-)
CTransitoriosPIMIENTO (-)	AGUAAGUA ENTUBADA PRIVADA (-)

Table 6: *Modeling of Nonzero Invasive , Top 5 features for each method*

The top 5 variable models for logistic regression and linear regression can be found in Table 5 and Table 6, respectively. For Elastic Net logistic regression the misclassification rate was 0.21 on the test set for the 5 variable model. For the Elastic Net linear regression 5 variable model the RMSE was 0.25 and the optimal model only included 4 variables with the same RMSE.

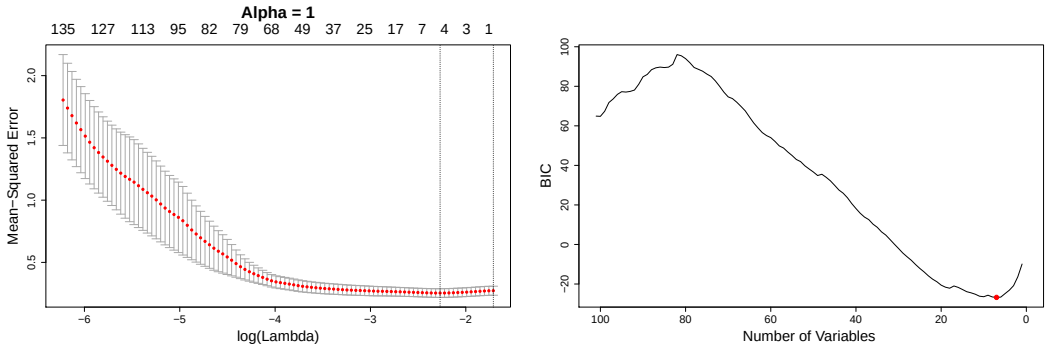


(a) *Cross-validation plot for Elastic Net*

(b) *Forward Selection plot*

Figure 11: *Invasive Binary Variable Selection. The Elastic Net plot shows adding variables does not improve just predicting the majority class. In Forward Selection the BIC very quickly levels off.*

For Forward Selection the 5 variable logistic regression model had a misclassification rate of 0.20 on the full dataset. The 5 variable linear regression model had an R^2 of 0.30 and the optimal model included 7 variables to increase the R^2 0.35.



(a) *Cross-validation plot for Elastic Net* (b) *Forward Selection plot*

Figure 12: *Nonzero Invasive Variable Selection. Similar to Figure 11 the lack of benefit from adding additional variables demonstrates a weak linear relationship.*

The cross-validation and forward selection plots in Figure 11 show that there was little gained from adding variables to the logistic regression model. The optimal Elastic Net model only included the intercept and the optimal Forward Selection model still only included 5 variables. For this reason the misclassification rates are not changed from above.

Since the percentage of farms with no invasive species is 20% the logistic regression was not able to do much more than just predict zero for all farms. This combined with the low performance of the linear model suggests that invasive species coverage cannot be well explained by the variables considered here. That said, more than seen in previous outcome variables, the canton variables seem to be particularly important for *invasive species*. Similar variables were chosen for Elastic Net and Forward Selection in logistic regression, but not for linear regression.

3.3 Land use choices

The analysis of land use choice requires a slightly different approach than used previously since the outcome variable is a categorical variable, with six different classifications of land use (each measured as a percentage of total land area):

- Permanent Crops
- Temporary Crops
- Fallow Land
- Tilled Land
- Pasture
- Brush

While not a perfect ordering, the classifications can generally be thought of progressing from land that has been most worked by the farmer to least worked. We visualized the land use data using a parallel coordinate plot (Figure 13) where each line represents a farm and the height represents the percent of land used for that category. Darker or thicker line areas indicate more farms falling along that path. We can see that there are farms primarily used for each of the different land use categories, but many more so for pasture, permanent crops and brush. If there were many horizontal, or near horizontal, lines on the graph that would indicate even distribution of land across multiple categories, but there is not much evidence of that in this plot.

There are multiple ways that this outcome variable can be formulated for modeling, but we decide to label each farm with its highest percentage land use category. There are only 9 such farms labeled as *fallow*, which is too small for modeling so we remove them from the consideration. The remaining categories and the number of farms are shown in Table 7.

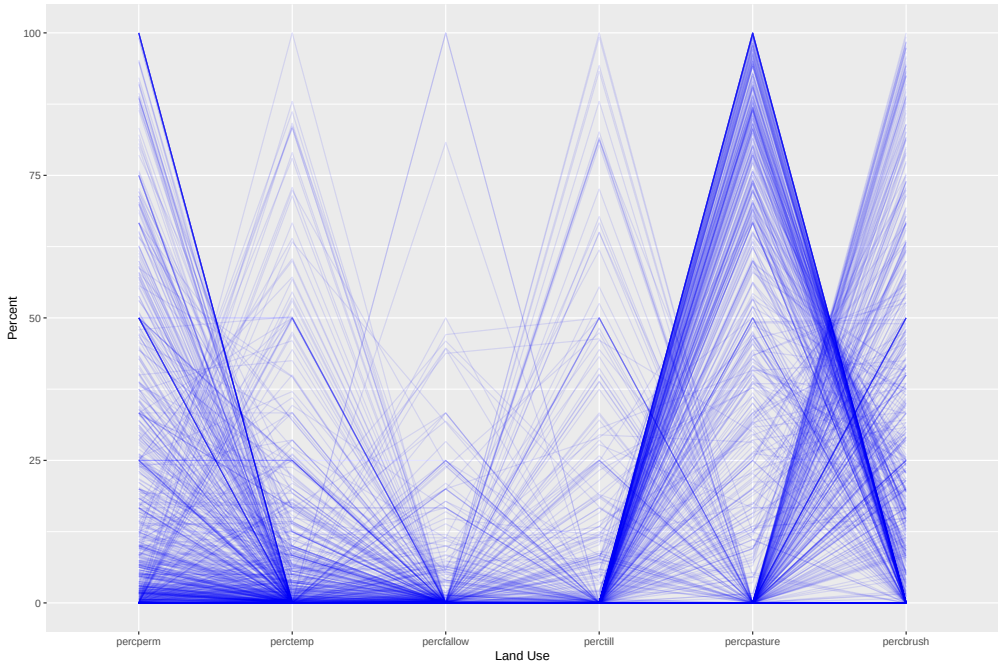


Figure 13: *Parallel Coordinate Plot of Land Use. There appear to be different types of farms, some primarily permanent crops, some primarily temporary crops, etc. The land use categories pasture, permanent crops and brush appear to be the most prevalent.*

Majority Category	Number of Farms
Perm	209
Temp	50
Till	36
Pasture	271
Brush	189

Table 7: *Landuse Category Count*

Using this derived categorical variable we used multinomial logistic regression with Elastic Net to build a model, which found an optimal set of coefficients for each category. The results from the 5 variable model, which had a misclassification rate of 0.39, are shown in Table 8. Different categories may have varying number of coefficients and *Perctill* had no variables included at this lambda level. As might be expected, presence of permanent crops predicted *Percperm*, presence of transitory crops predicted *Perctemp*, etc., but the specific crops found most predictive may be of interest.

Percperm	**Perctemp**	
CPermanentesCAFE (+)	CTransitoriosNABO (+)	
CPermanentesNARANJA (+)	CTransitoriosMAIZ DURO CHOCLO (+)	
pc4None (+)	librasvendida (+)	
VentaLibras (+)		
CPermanentesGUABA (+)		
Percpasture	**Percbrush**	**Perctill**
pc4None (-)	cantonSanta Cruz (-)	
v30a (+)	ArbolesAGUACATE (+)	
v3 (+)		
PastosBRACHIARIA (+)		

Table 8: *Modeling of Landuse, Top features for each category*

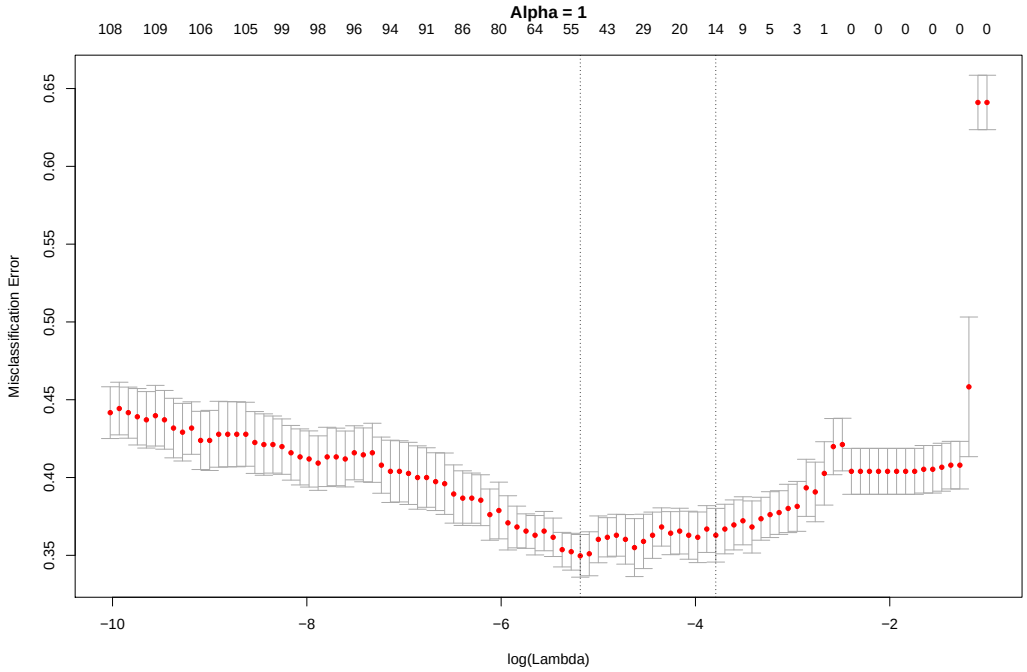


Figure 14: *Land Use Cross-validation plot for Elastic Net. The Elastic Net model shows a large improvement over just predicting the majority class. Since there are now 5 categories, such a strategy would result in a 0.65 misclassification rate.*

The plot of the Elastic Net cross-validation is shown in Figure 14. The optimal model achieved a 0.36 misclassification rate against the test set on average, which given 5 categories to choose from shows decent predictive strength. In the Appendix, the coefficients estimated for the optimal model can be found in Tables 15 and 16.

4 Statistical Methods

4.1 Generalized Linear Models

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon \tag{1}$$

Standard Linear Regression can be represented in matrix form as seen in equation 1 above, When there are n observations and p predictor variables, $\mathbf{Y}$ is a $n \times 1$ vector of the outcome variable, $\mathbf{X}$ is a $n \times p$ matrix of predictor variables, β is a $p \times 1$ vector of variable coefficients and ϵ is the error term. The standard linear model works best when the outcome variable $\mathbf{Y}$ has a normal distribution, and therefore takes continuous values. When the outcome variable is continuous, but not normal shaped (e.g., skewed like the productivity data) it can be possible to transform the data by taking the logarithm or something similar. However, when the outcome variable is discrete (such as binary labels of 1 and 0 denoting absence/presence of a feature) a further modification must be made. The outcome variable is clearly no longer normally distributed, as it is not even continuous. Without modification we could get predicted values below 0, above 1 or somewhere in between, none of which make sense in this context.

This calls for the use of logistic regression, a type of generalized linear model [2], in which we perform a logit transformation as seen in equation 2 below so that the $\mathbf{X}\beta$ can still be mapped to a continuous scale. In some respects this is a computational concern, but it also changes the way coefficients can be interpreted. For example, instead of a one unit change in X_1 predicting a β_1 change in the predicted Y, in this case it predicts a β_1 change in the log odds of Y.

$$log\frac{\Pr(Y=1)}{\Pr(Y=0)} = \mathbf{X}\beta \tag{2}$$

Equation 2 can be rewritten as below in equation 3, which gives the predicted probability of an observation being class 1.

$$\Pr(Y=1) = \frac{e^{\mathbf{X}\beta}}{1+e^{\mathbf{X}\beta}} \tag{3}$$

This can be extended beyond binary variables to categorical variables when there are K classes, referred to as Multinomial Logistic Regression. Now to calculate the probability of class k, we use equation 4 below.

$$\Pr(Y=k) = \frac{e^{\mathbf{X}\beta_k}}{\sum_{l=1}^{K-1} e^{\mathbf{X}\beta_l}} \tag{4}$$

4.2 Performance Measures

There are many measures of fit for linear models. When there are many possible predictor variables, care must be taken to use appropriate measures, as some measures will favor just adding all of the variables to the model. For example, if we aim to minimize the mean square error adding more predictor variables to the model will always be encouraged. Since that is not desired, measurements like Bayesian Information Criterion (BIC) can be used. BIC is a combination of how well the model fits the data and a penalty term for the number of predictor variables included in the model. The goal is to minimize BIC, that is the model with the best balance of small size and goodness of fit. BIC is chosen over other potential measurements because it puts a large penalty on the inclusion of additional variables.

Another approach is to use cross-validation. In this technique the dataset is first split into k equally sized sets. Then a model is fit using $k-1$ of the sets (training sets) and evaluated on the remaining 1 (test) set. This is repeated k times, each time reserving a different 1 test set, and then results across the k runs are averaged. The benefit of this is that model building and model evaluating are happening on different portions of the data, so we can distinguish if the model is picking up on generalizable patterns or just random noise. Using cross-validation the average test set mean square error is an appropriate measure of model fit. We can also capture the standard deviation across the k runs to measure variability, which is shown in the error bars of the Elastic Net cross-validation plots.

4.3 Best Subset and Forward Selection

The essential goal of variable selection is to find the best combination of predictor variables to explain the outcome variable. As discussed above, when we have many possible predictors we often want to put a constraint on the problem so that all variables are not included. Such a constraint might be limiting the number of variables included or that the model found can generalize to other data. Best subset selection, the most natural, but computationally difficult way is to try all possible combinations of variables and select the best fitting combination. However, when the number of variables, p, is large this quickly becomes infeasible, as there are 2^p possible combinations.

One approach to tackle the computational complexity discussed above is to restrict the search for the optimal number of predictor variables, which is what Forward Selection does. In this algorithm, we start with an empty model and iteratively add a new variable at each stage that most increases the fit. This procedure can work well, but may not find the optimal solution. As an example, consider a case where X_1 is the single most predictive variable, but the combination of X_2 and X_3 is the best two variable combination. The algorithm will first add X_1, but then regardless whether it adds X_2 or X_3 next, it will have found a suboptimal solution. In general, we can decide to stop adding variables once we have reached an optimal performance measure like BIC or cross-validation test error.

4.4 Regularized Regression

$$\min_{\beta} \|\mathbf{Y} - \mathbf{X}\beta\|_2^2 \qquad \text{(linear model)}$$

$$\min_{\beta} \|\mathbf{Y} - \mathbf{X}\beta\|_2^2 + \lambda||\beta||_2^2 \qquad \text{(ridge regression)}$$

$$\min_{\beta} \|\mathbf{Y} - \mathbf{X}\beta\|_2^2 + \lambda||\beta|| \qquad \text{(LASSO)}$$

The above notation of $||\cdot||_2^2$ and $||\cdot||$ are defined in general as: $||\mathbf{X}||_2^2 = x_1^2 + x_2^2 + ... + x_n^2$ and $||\mathbf{X}|| = |x_1| + |x_2| + ... + |x_n|$. As shown in the top equation of the standard linear model, we try to find the β, that is a vector of coefficients, that minimizes the squared difference between the true $\mathbf{Y}$ and the predicted $\hat{\mathbf{Y}}$ (which is $\mathbf{X}\beta$). In regularized regression we do the same thing, but also add a second term that we look to simultaneously minimize. This second term adds a penalty scaled by λ for increasing values of β, so the two terms must be balanced. The optimal model will find a balance between fitting the outcome variable closely, but not having too large of coefficient values. The difference between Ridge regression and LASSO is how we add up the coefficients. In Ridge Regression the coefficients are squared and then summed, in LASSO we take the absolute value of the coefficients and then sum them. LASSO will encourage most of the coefficients to go to zero, thus only including a small number of terms in the model. Ridge regression will encourage the coefficient values to be spread out among predictor variables, leaving all of the variables in the model, but helping to offset negative effects of correlated predictor variables.

$$\min_{\beta} \|\mathbf{Y} - \mathbf{X}\beta\|_2^2 + \lambda[(1-\alpha)||\beta||_2^2 + \alpha||\beta||] \qquad \text{(Elastic Net)}$$

The technique that is used in this analysis is a combination of the Ridge and LASSO penalties, called Elastic Net. As can be seen in the equation above both the square of the coefficients and the absolute value of the coefficients is included, with the contribution of each controlled by the size of alpha (α) which takes a value between 0 and 1. Elastic Net, thus combines the favorable properties of Ridge and LASSO, in that it can achieve both sparse models and can handle correlated predictor variables. Both the λ and the α can be set using cross-validation (as discussed above) to appropriate values for the particular dataset. In each Elastic Net cross-validation plot found in the Results section, the specific α used is labeled at the top of the graph.

5 Limitations and Future Work

There are a few key considerations that should be kept in mind when interpreting this analysis. The first is that relationships discovered in this analysis are correlational in nature and cannot be assumed to be causal. Just because farms with a higher coverage of invasive species have lower productivity does not necessarily mean the invasive species causes lower productivity. It could be that lower productivity causes higher invasive species coverage. Or there could other factors not captured in the model that influence both productivity and invasive species. In order to determine causality, relationships of interest should be tested in a designed experiment.

Secondly, p-values and confidence intervals for coefficients were intentionally not included in the analysis. In standard regression analysis we pre-specify the model and then test which variables are found to be significant. However, when using Elastic Net and Forward Selection like we have done here, we do not specify the model ahead of time, but instead let the data decide the model. This violates the significance test assumption and can lead to misleadingly small p-values (see Chapter 6 of [4]). While there is some recent work ([9]) suggesting this may be acceptable under certain assumptions, as well as some advanced techniques to try to adjust for this, it is recommended to view the results in this report as an exploratory analysis rather than definitive evidence.

Future analysis might look to explore better fitting relationships, particularly for the outcome variables that had poor RMSE and R^2 values. The relationships modeled in this report only considered linear combinations of predictor variables to predict/ explain the outcome variables. Modifications could include adding interaction terms (i.e., x_1x_2) or nonlinear terms (i.e., x_1^2 or binary transformations $x_1 > 10$). Exploring all possible modifications of this type is not computationally feasible, but with domain knowledge a subset of theorized relationships could be tested.

For some of the outcome variables the island was found to be a significant varriable, but it may be more helpful to model each of the islands separately. A potential limitation to such an approach would be the small sample sizes for some of the islands, so this would be best carried out with a smaller set of potential predictor variables. Lastly, the variables used here primarily covered socioeconomic dimensions. The addition of physical and biotic variables may help better predict/explain the outcome variables or may change the importance of previously highlighted socioeconomic variables.

References

[1] Censo De Las Unidades De Producción Agropecuaria De Galápagos, sinagap.agricultura.gob.ec/pdf/censo_galapagos/cuestionario_censo_upa_galapagos.pdf.

[2] Dobson, A. J., & Barnett, A. (2008). An introduction to generalized linear models. CRC press.

[3] Friedman, J., Hastie, T., & Tibshirani, R. (2009). glmnet: Lasso and elastic-net regularized generalized linear models. R package version, 1(4).

[4] Hastie, T., Tibshirani, R., & Wainwright, M. (2015). Statistical learning with sparsity: the lasso and generalizations. CRC press.

[5] Lumley, T., & Miller, A. (2009). Leaps: regression subset selection. R package version 2.9. Online at http://CRAN. R-project.org/package= leaps.

[6] Miller, B. W., Breckheimer, I., McCleary, A. L., Guzmán-Ramirez, L., Caplow, S. C., Jones-Smith, J. C., & Walsh, S. J. (2010). Using stylized agent-based models for population environment research: a case study from the Galápagos Islands. Population and environment, 31(6), 401-426.

[7] Valdivia, G., Wolford, W., & Lu, F. (2014). Border crossings: New geographies of protection and production in the Galápagos Islands. Annals of the Association of American Geographers, 104(3), 686-701.

[8] Walsh S.J., Mena C.F. (2013) Perspectives for the Study of the Galápagos Islands: Complex Systems and Human-Environment Interactions. In: Walsh S., Mena C. (eds) Science and Conservation in the Galápagos Islands. Social and Ecological Interactions in the Galápagos Islands, vol 1. Springer, New York, NY

[9] Zhao, S., Shojaie, A., & Witten, D. (2017). In Defense of the Indefensible: A Very Naive Approach to High-Dimensional Inference. arXiv preprint arXiv:1705.05543.

[10] Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67(2), 301-320.

6 Appendix

Variable	Elastic Net	Forward Selection
Intercept	3.133 4	3.334 9
CPermanentesOTROS BANANOS	0.186 6	0.276 5
CPermanentesPLATANO	0.136 6	0.257 5
CTransitoriosTOMATE RINON	0.116 7	0.205 7
CTransitoriosYUCA	0.011 9	NA
numcultivo	$5.940\,0 \cdot 10^{-06}$	$8.770\,0 \cdot 10^{-06}$
PastosBRACHIARIA	−0.159 5	−0.262 6
PastosELEFANTE	−0.091 6	−0.208 1
PastosKING GRASS	−0.064 8	NA
pc4None	0.050 7	NA
pc6	−0.001 7	−0.002 0
ArbolesGUABA	0.080 4	0.141 5
ArbolesGUANABANA	0.000 1	NA
ArbolesGUINEO	0.090 4	0.333 2
ArbolesLIMON MANDARINA	0.063 1	0.204 2
ArbolesNARANJA	0.028 1	NA
ArbolesPLATANO	0.017 6	NA
ad11	$1.840\,0 \cdot 10^{-05}$	NA
produccionenlibrasproductovendido	$7.740\,0 \cdot 10^{-06}$	$2.000\,0 \cdot 10^{-05}$
v3	$-9.630\,0 \cdot 10^{-05}$	NA
v30a	−0.001 3	NA
c12	$3.010\,0 \cdot 10^{-06}$	$1.190\,0 \cdot 10^{-05}$
a7a	$2.750\,0 \cdot 10^{-06}$	NA
ga9Si	0.023 5	NA
ga9a	$7.470\,0 \cdot 10^{-06}$	0.000 1
ga15cualADECUACION UPA	−0.291 5	−1.524 0
ga15cualMANTENIMIENTO DE CAFdb	−0.276 9	−1.938 6
ga15cualPACHETE	0.644 4	1.803 7
e30	0.017 6	NA
percperm	0.001 8	NA
perctemp	0.004 3	NA
perctill	−0.000 9	−0.008 1
percpasture	−0.007 1	−0.010 9
percinv	−0.002 7	−0.010 5
percbrush	−0.005 2	−0.009 6
percinv2	−0.002 3	NA
d3Si	−0.026 1	NA
ReclassCONSERVACION	−0.155 0	−0.337 5
ReclassPECUARIO	−0.067 5	−0.113 7
ABANDONEDTRUE	−0.064 9	NA
CONSERVATIONTRUE	−0.050 3	NA
FORESTRYTRUE	−0.084 9	−0.145 5
LODGINGTRUE	−0.013 5	−0.129 8
ENERGIAELENERGIA SOLAR PRIVADA	−0.654 9	−1.400 8
VIASDEACASFALTADA	−0.063 7	−0.133 5
RELIEVEABRUPTO	−0.063 8	NA
RELIEVEPLANO	NA	0.141 3

Table 9: *Full coefficient list for Production model*

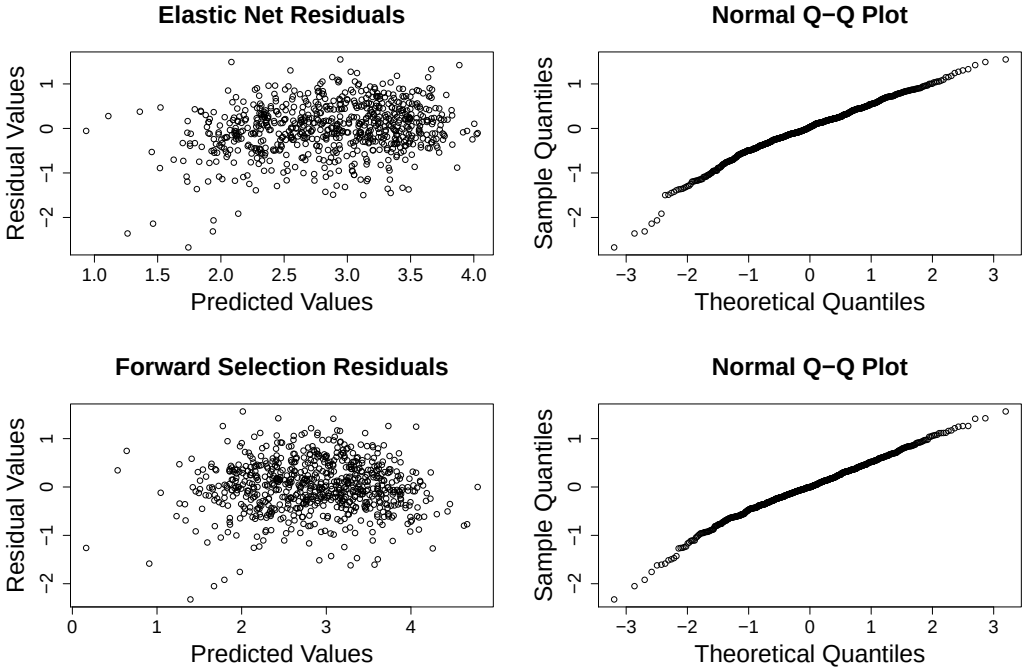


Figure 15: *Diagnostic Residual plots for Production. The plots on the left show the predicted values vs. the residuals and both demonstrate a desirable lack of pattern. The plots of the right examine the normality of the residuals, both staying close to the desired straight diagonal pattern.*

Variable	**Elastic Net**	**Forward Selection**
Intercept	$-2\,042.710\,1$	$-1\,963.536\,5$
CTransitoriosFREJOL TIERNO	$-420.731\,6$	$-2\,665.119\,5$
PastosBRACHIARIA	$-219.199\,8$	NA
ad4	$0.156\,8$	$1.696\,3$
produccionenlibrasproductocosechadoautoconsumo	$0.049\,6$	$0.098\,0$
v3	$-24.101\,1$	$-63.533\,9$
v45	$-51.680\,1$	$-131.248\,3$
percperm2	$4.593\,3$	$19.940\,5$
CATTLETRUE	$-226.104\,5$	NA
AGUAAGUA POTABLE PRIVADA	$-5\,209.352\,4$	$-21\,382.103\,6$
v44	NA	$5.024\,9$
ALCANTARILPOZO SEPTICO O CIEGO PUBLICO	NA	$-3\,896.765\,5$

Table 10: *Full coefficient list for Net Income model*

Variable	**Elastic Net**	**Forward Selection**
Intercept	$-0.800\,1$	$-1.377\,1$
s4	$0.002\,0$	NA
CosechaLibras	$1.330\,0 \cdot 10^{-07}$	NA
v3	$0.006\,4$	$0.026\,3$
v45	$0.008\,2$	NA
GastosPecuarios	$8.020\,0 \cdot 10^{-08}$	$5.640\,0 \cdot 10^{-05}$
CATTLETRUE	$0.020\,6$	NA
VentaLibras	NA	$2.150\,0 \cdot 10^{-05}$
librasvendida	NA	$7.530\,0 \cdot 10^{-05}$
perctemp2	NA	$-0.027\,0$
CTransitoriosRABANO	NA	$1.206\,3$
VIASDEACASFALTADA	NA	$0.486\,5$

Table 11: *Full coefficient list for Binary Workers model*

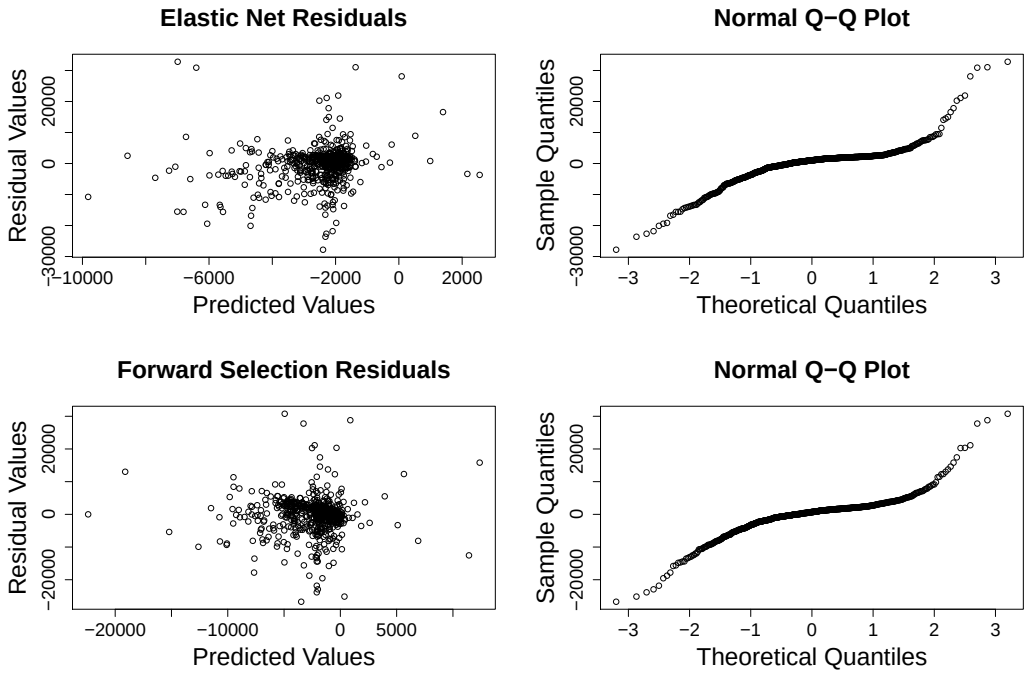


Figure 16: *Diagnostic Residual plots for Net Income. The plots of the predicted values vs. the residuals both demonstrate a desirable lack of pattern. The plots examining the normality of the residuals do not follow a diagonal line as closely as would be hoped, but are not an extreme departure.*

Variable	**Elastic Net**	**Forward Selection**
Intercept	$0.702\,9$	$0.577\,5$
s4	$0.001\,2$	$0.002\,9$
v3	$0.000\,6$	NA
v44	$1.860\,0 \cdot 10^{-05}$	NA
GastosAgricolas	$3.490\,0 \cdot 10^{-06}$	$3.960\,0 \cdot 10^{-05}$
biokSi	$0.035\,5$	$0.363\,4$
CTransitoriosCILANTRO	NA	$0.280\,9$
ArbolesLIMON MANDARINA	NA	$0.249\,7$
a24f	NA	$5.330\,0 \cdot 10^{-05}$
ga15cualLIMPIEZA DE TERRENO	NA	$0.531\,5$
FARMINGTRUE	NA	$-0.127\,2$

Table 12: *Full coefficient list for Nonzero Workers model*

Variable	**Elastic Net**	**Forward Selection**
Intercept	$-1.337\,3$	$-1.055\,0$
cantonSan Cristobal	NA	$0.193\,5$
cantonSanta Cruz	NA	$-1.771\,0$
cantonFloreana	NA	$-16.599\,3$
v30a	NA	$0.007\,6$
CTransitoriosMAIZ SUAVE CHOCLO	NA	$0.902\,8$
CPermanentesNARANJA	NA	$-0.691\,2$

Table 13: *Full coefficient list for Binary Invasive model*

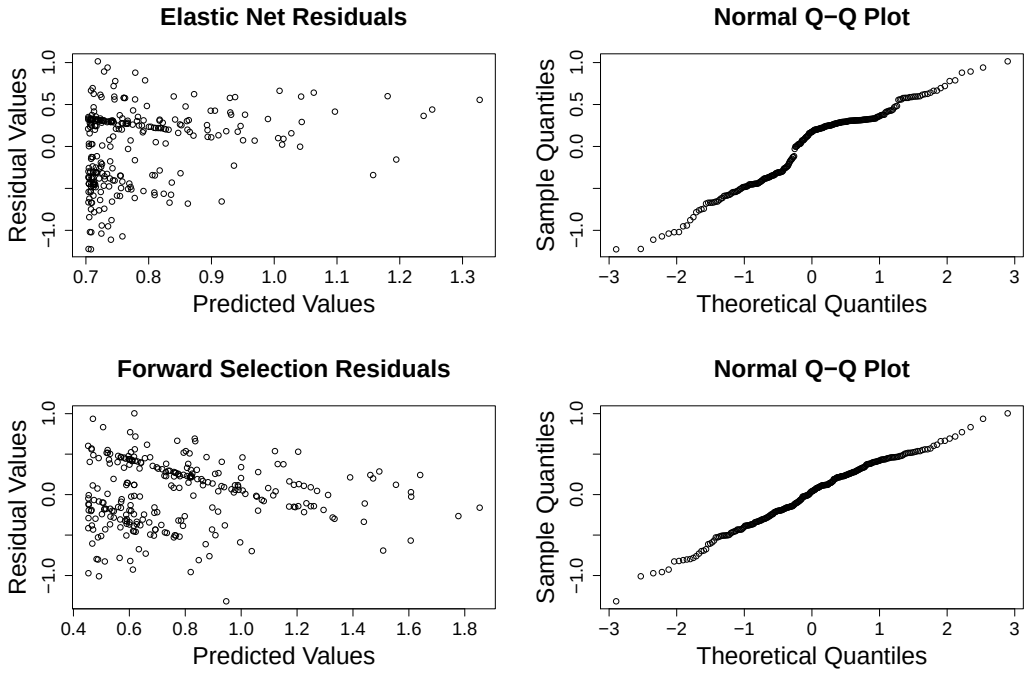


Figure 17: *Diagnostic Residual plots for Nonzero workers. The plots of predictions versus residuals shows points bunch at the left end, but the QQ plot closely follows a diagonal line.*

Variable	**Elastic Net**	**Forward Selection**
Intercept	1.088 0	1.363 4
cantonSan Cristobal	0.036 2	NA
CTransitoriosSANDIA	−0.106 8	−0.556 9
PastosKING GRASS	−0.236 5	−0.614 0
pc4None	0.072 9	0.202 0
CPermanentesPAPAYA	NA	−0.360 2
a7c	NA	−0.003 5
AGUAAGUA DE POZO PUBLICA	NA	−1.243 8
TELEFONOFNO TIENE	NA	−0.236 9

Table 14: *Full coefficient list for Nonzero Invasive model*

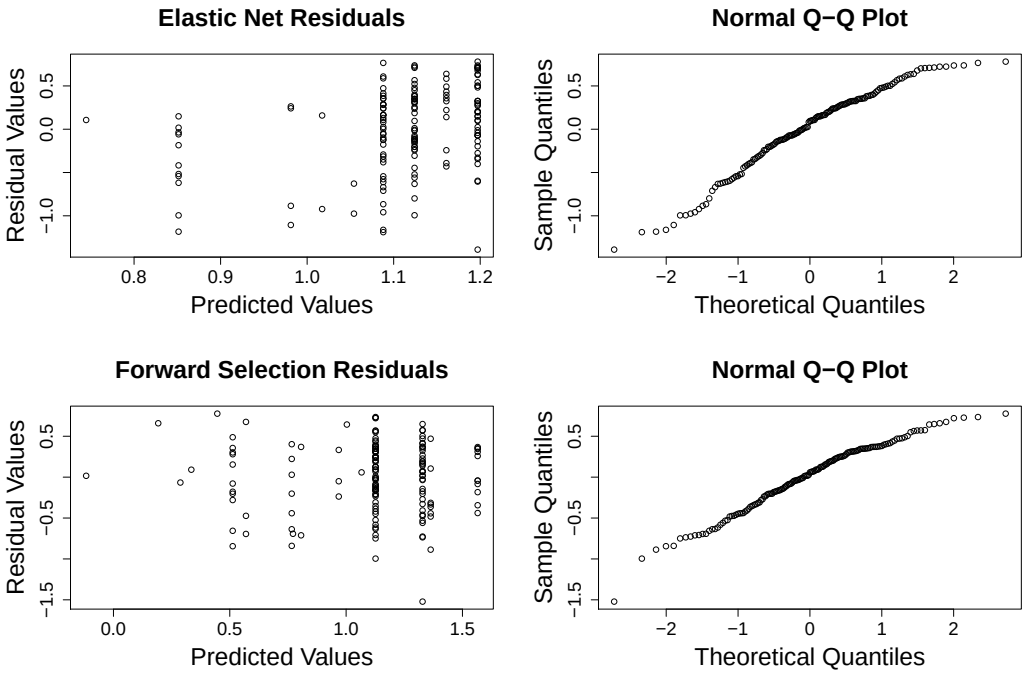


Figure 18: *Diagnostic Residual plots for Nonzero Invasive. Due to only a few variables being included in these models and some of them being discrete, the predictions occur at a limited number of distinct points.*

Variable	Percperm	Perctemp	Perctill
Intercept	$0.070\,1$	$-1.031\,0$	$-0.987\,9$
s4	$-0.000\,6$	NA	NA
CPermanentesAGUACATE	$0.027\,2$	NA	NA
CPermanentesCAFE	$0.986\,5$	NA	NA
CPermanentesGUABA	$0.253\,1$	NA	NA
CPermanentesMANDARINA	$0.010\,7$	NA	NA
CPermanentesNARANJA	$0.837\,1$	NA	NA
CPermanentesOTROS BANANOS	$0.106\,6$	$-0.140\,8$	NA
p9	$0.000\,7$	NA	NA
VentaLibras	$1.080\,0 \cdot 10^{-05}$	NA	NA
CTransitoriosAPIO	$-0.127\,3$	NA	NA
CTransitoriosBROCOLI	$-0.391\,2$	NA	NA
CTransitoriosFREJOL TIERNO	$-0.235\,5$	NA	NA
CTransitoriosMAIZ DURO CHOCLO	$-0.038\,9$	$0.980\,0$	NA
CTransitoriosTOMATE RINON	$-0.013\,8$	NA	NA
CTransitoriosVAINITA	$-0.036\,5$	NA	NA
pc4None	$0.278\,0$	NA	NA
ArbolesMANDARINA	$0.089\,6$	NA	NA
ga15cualADECUACION UPA	$0.102\,7$	NA	NA
to57e	$0.000\,9$	NA	NA
biokSi	$0.236\,4$	NA	NA
ENERGIAELNO TIENE	$-0.061\,0$	NA	NA
INTERNETINTERNET PRIVADO	$-0.340\,5$	NA	NA
ALCANTARILALCANTARILLADO PUBLICO	$0.176\,0$	NA	NA
ALCANTARILSIN INFORMACION	$-0.324\,4$	NA	NA
RELIEVEONDULADO	$-0.050\,5$	NA	NA
CTransitoriosCOL	NA	$0.114\,2$	NA
CTransitoriosHIERVITA	NA	$0.042\,2$	NA
CTransitoriosMELON	NA	$0.625\,1$	NA
CTransitoriosNABO	NA	$1.767\,8$	NA
CTransitoriosPIMIENTO	NA	$0.211\,5$	NA
librasvendida	NA	$1.480\,0 \cdot 10^{-05}$	NA
ga15cualPERSONAL PARA SEMBRAR	NA	$2.348\,9$	NA
e30f	NA	$0.914\,6$	NA
o4f	NA	NA	$0.001\,2$
ga15cualGASTOS HERRAMIENTAS	NA	NA	$1.554\,3$
ga15cualHERRAMIENTA DE TRANAJO	NA	NA	$1.412\,0$
ga15cualPARA PREVERCION MEDICINA	NA	NA	$1.687\,8$
GastosPecuarios	NA	NA	$1.250\,0 \cdot 10^{-06}$

Table 15: *Full coefficient list for Landuse model*

Variable	Percpasture	Percbrush
Intercept	1.458 4	0.490 4
cantonSan Cristobal	−0.068 2	NA
CTransitoriosCILANTRO	0.015 6	NA
PastosBRACHIARIA	0.153 7	NA
PastosKING GRASS	0.163 9	NA
pc6	0.003 3	NA
ArbolesPAPAYA	−0.286 3	NA
v3	0.007 6	NA
v30a	0.013 0	NA
v45	0.000 3	NA
o4b	0.000 2	NA
e29f	0.000 6	NA
d3Si	0.125 1	NA
ENERGIAELENERGIA SOLAR PRIVADA	0.025 8	NA
cantonSanta Cruz	NA	−0.476 4
c10Si	NA	−0.020 9
c14	NA	0.005 1
r3	NA	0.073 1
CPermanentesNARANJILLA	NA	0.227 1
CPermanentesPINA	NA	0.102 1
t28	NA	$1.490\,0 \cdot 10^{-05}$
ArbolesAGUACATE	NA	0.321 6
ArbolesGUAYABA	NA	0.716 7
ArbolesNARANJA	NA	0.035 8
ArbolesNARANJA AGRIA	NA	−0.332 8
oe	NA	0.000 4
ForestalAGUACATE	NA	0.039 6
ForestalCEDRELA	NA	0.431 6
ga15cualHERRAMIENTAS DE TRABAJO	NA	0.517 4
ga15cualLIMPIEZA DE LA FINCA	NA	0.221 0
ga15cualTRABAJADORES	NA	0.418 3
GastosAgricolas	NA	$-2.950\,0 \cdot 10^{-05}$
AGUAAGUA ENTUBADA PUBLICA	NA	−0.065 6
AGUANO TIENE	NA	0.091 3
ALCANTARILPOZO SEPTICO O CIEGO PRIVADO	NA	0.233 5
VIASDEACASFALTADA	NA	0.118 2
VIASDEACSENDERO	NA	0.215 1

Table 16: *Full coefficient list for Landuse model (cont.)*