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 8th, 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 Elasticnet, to find the most important factors 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 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 its 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 models of the Galápagos, but with limited interaction parameters between agents, particularly in regards to farm success ([5], [6],[7]). 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: Modeling, Section 3: Results, Section 4: Statistical Methods, and Section 5: Limitations and Future Work. 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 at the end.

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 will be 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 ElasticNet 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' [4]). ElasticNet 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' [2]).

In general, these techniques have slightly different aims. Forward Selection chooses a model that best explains the variance in the dataset at hand. Elasticnet 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

3.1 Farm Success

The first three outcome variables of interest can be grouped together under the category of farm success; labeled as productivity, net income and number of workers supported. Productivity is calculated as the total pounds of crops and livestock produced divided by the farm surface area. Net income is calculated as the the revenue from all products sold minus total expenses. Number of workers supported is calculated as the total labor expenditures divided by a standard full-time worker's salary.

3.1.1 Productivity

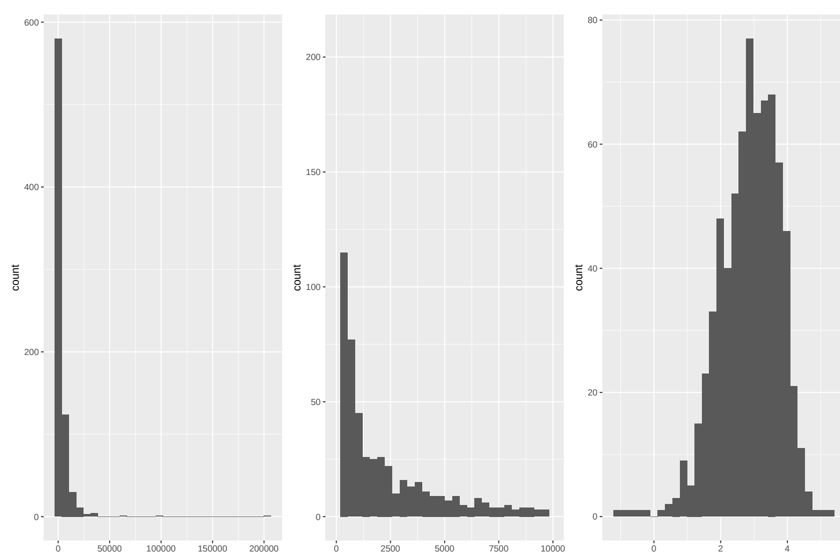


Figure 1: Histogram of Productivity, from left to right showing the full range, between 0-10,000 lbs/hectare, and the log transformed nonzero values. The log transformed values exhibit the desired normal shape.

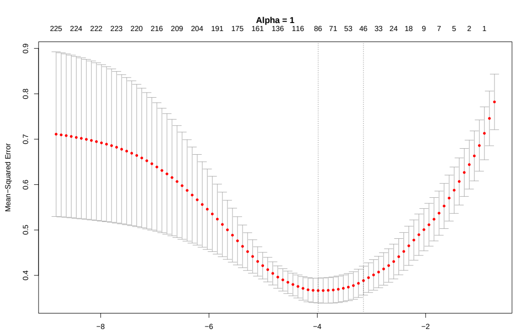
The histograms of the Productivity variable (Figure 1) show a strong skewness, both when looking at all observations, and when zooming into observations between 0-10,000 lbs/hectare. To achieve a distribution closer to normal (bell-curved), which will benefit the linear model, we took the log_{10} transformation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

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

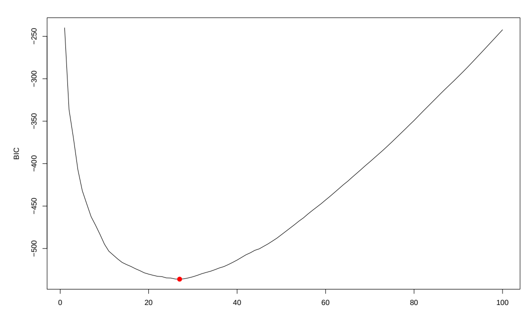
Table 1: Modeling of Productivity, Top 5 features for both methods

We built linear models using both methods, Elasticnet 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 for any specific statistical property. The results of the best 5 variable model are shown in Table 1, listed in alphabetical order. Next to variable names the direction of the relationship is indicated with a (+) or (-). There is some overlap between the variables selected by each of the methods, but also unique choices made by each method. The Root Mean-Squared Error (RMSE) for the 5 variable Elasticnet model is 0.78 and the R^2 for Forward Selection is 0.48. The RMSE is on average how far the predicted value is from the actual value and R^2 is the percentage of variability in the outcome variable that is explained by the model. To provide context for the RMSE, we can look at the point furthest to the right on the cross-validation plot and see how the model would perform when no variables are included in the model, that is just predicting the average outcome value.

Plots from the optimal models for Elasticnet and Forward Selection are shown in Figure 2. The cross- validation plot for Elasticnet 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}. For Forward Selection the plot is much more straightforward, we plot the number of variables included versus the information criterion that we choose to minimize and mark the optimal point in red.



(a) Cross-validation plot for Elasticnet



(b) Forward Selection plot

Figure 2: Productivity Variable Selection. The Elasticnet plot shows a desirable "U" shape where a moderate number of variables provides a significant improvement in prediction error. The optimal Forward Selection model includes 26 variables.

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. The optimal model chosen for Elasticnet includes 45 variables (not counting the intercept term) and using Forward Selection we choose a model of 26 variables, almost all of which are also included in the Elasticnet model. The coefficients estimated for both models can be found in Table 10 in the Appendix. For the full models, the RMSE and R^2 are 0.61 and 0.63, respectively. 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 Elasticnet and Forward Selection models (plots shown in Figure 13 in the Appendix) do not raise any concerns.

3.1.2 Net Income

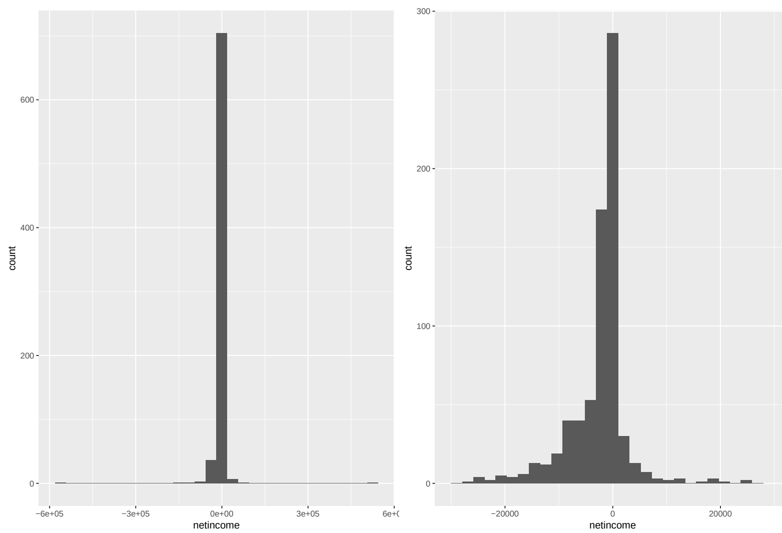


Figure 3: Histograms of Net Income, with full data and data within one standard deviation displayed. The values very far from the center are of concern when modeling.

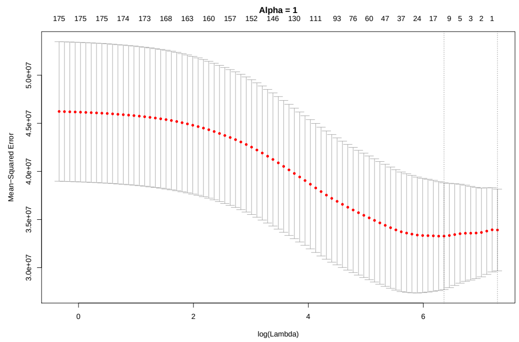
The histograms for Net Income (Figure 3) show a symmetric shape, but a very spiky center and a few observations wide in the tails. We attempted to fit a model on the full dataset, but find the observations with large absolute values are obscuring other possible information in the model. After trying several cutoff thresholds and examining model fit, we removed the 28 observations that are beyond one standard deviation (30,478) from the mean in either direction. The following models are fit on this reduced dataset of 727 observations.

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

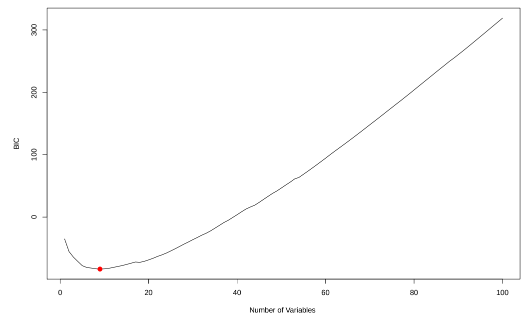
Table 2: Modeling of Net Income, Top 5 features for both methods

The results of the best 5 variable model are shown in Table 2, with a RMSE of 5790.35 for Elasticnet and a R^2 of 0.15 for Forward Selection. 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 Elasticnet model to be at λ_{min}, which includes 9 variables. For Forward Selection, we also included 9 variables, with all but two overlapping with



(a) Cross-validation plot for Elasticnet...



(b) Forward Selection plot

Figure 4: Net Income Variable Selection. The Elasticnet plot shows a much smaller improvement in prediction error and Forward Selection includes a small number of variables.

the Elasticnet choices. The coefficents estimated for both models can be found in Table 11 in the Appendix. The full models had an RMSE and R^2 of 5768.30 and 0.19, respectively. 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 Elasticnet and Forward Selection models (plots shown in Figure 14 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 is not very strong.

3.1.3 Number of Workers Supported

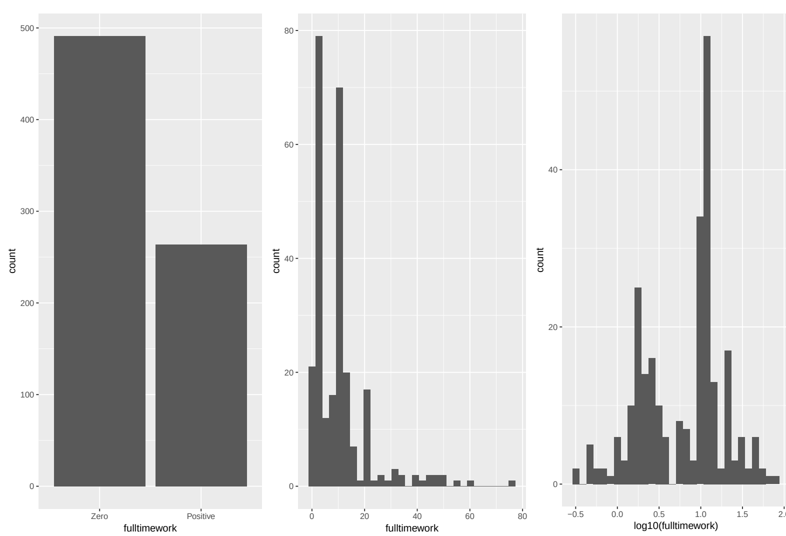


Figure 5: Histograms of Workers, showing the proportion of zeroes, farms with nonzero workers, and the log transformed number of workers. The log transformation helps remove the skewness, but some bi-modality still remains.

The first plot in Figure 5 shows that a large percentage of the farms are not able to support any workers. For this reason, we divided up the modeling task for this outcome variable. 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.

Sometimes Elasticnet will simultaneously choose to eliminate multiple variables at the same time, in this case there is not a 5 variable model so we show the 6 variable model for Elasticnet. For logistic

Elasticnet	Forward Selection
CATTLETRUE (+)	GastosPecuarios (+)
CosechaLibras (+)	librasvendida (+)
GastosPecuarios (+)	perctemp2 (-)
s4 (+)	v3 (+)
v3 (+)	VentaLibras (+)
v45 (+)	

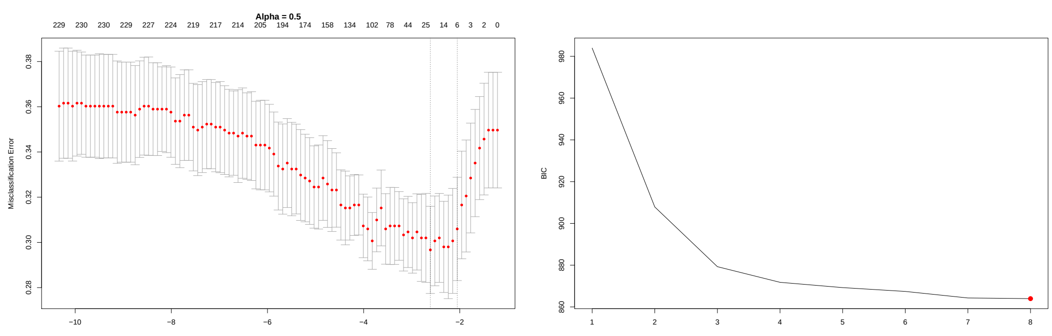
Table 3: Modeling of Binary Workers , Top 5 features for each method

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

Table 4: Nonzero Workers

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

regression we can measure performance with the misclassification rate, which on average is 0.31 on the test set for Elasticnet and 0.26 on the full dataset for Forward Selection. These can be compared to the 0.35 misclassification rate we would achieve if we simply predicted the majority class, zero workers, every time. For the linear model 5 variable model the RMSE is 0.46 and the R^2 0.30. The 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.

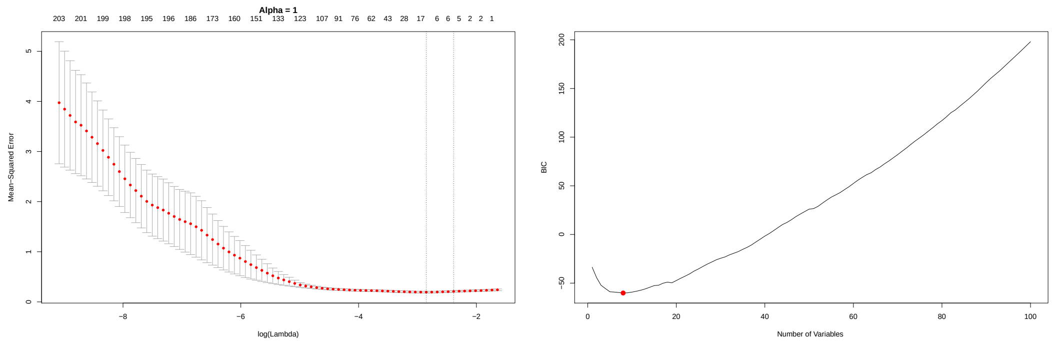


(a) Cross-validation plot for Elasticnet

(b) Forward Selection plot

Figure 6: Workers Binary Variable Selection. The Elasticnet 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.

For logistic regression Elasticnet chose a model of size 6 and Forward Selection chose a model of size 7. Since these models are not much larger the misclassification rates remained at 0.31 and 0.26, respectively. For linear regression, the full models for Elasticnet and Forward Selection had a RMSE and R^2 of 0.46 and 0.34, respectively. The Elasticnet full model had 5 variables and the Forward Selection full model had 8 variables. In the Appendix, the coefficients estimated for logistic regression models can be found in Table 12 and the coefficients estimated for linear regression models can be found in Table 13.



(a) Cross-validation plot for Elasticnet

(b) Forward Selection plot

Figure 7: 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 8 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.

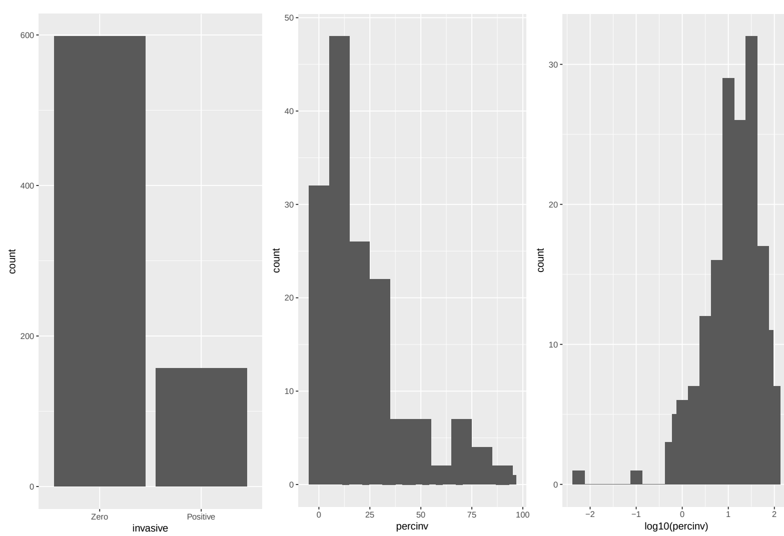


Figure 8: Histogram of Invasive Species. To the left: the proportion of zeroes, in the middle: farms with greater than zero, and to the right: the log transformed percent of invasive species. The two very small values on the right plot are removed.

The top 5 variable models for logistic regression and linear regression can be found in Table 6 and Table 7, respectively. For logistic regression the misclassification rate was 0.21 on the test set for Elasticnet and 0.20 on the full dataset for Forward Selection. For the linear model 5 variable model the RMSE was 0.25 and the R^2 was 0.30.

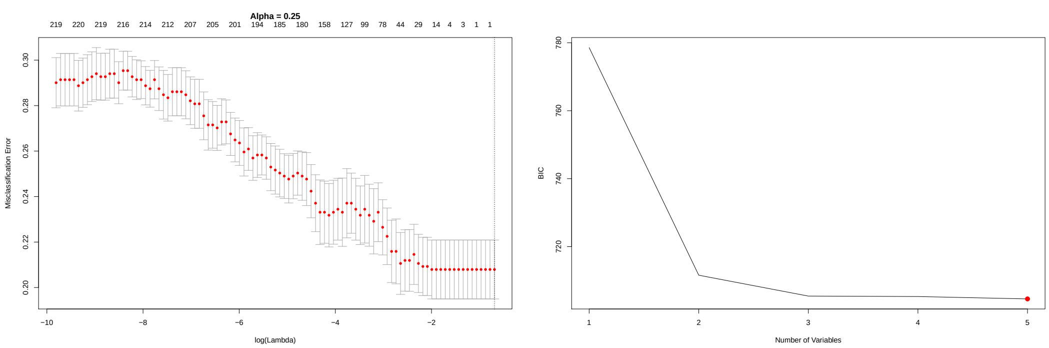
The cross-validation and forward selection plots in Figure 9 show that there was little gained from adding variables to the logistic regression model. The optimal Elasticnet model only included the intercept and the optimal Forward Selection model still only included 5 variables. For this reason the misclassification rates are not changed from above.

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

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

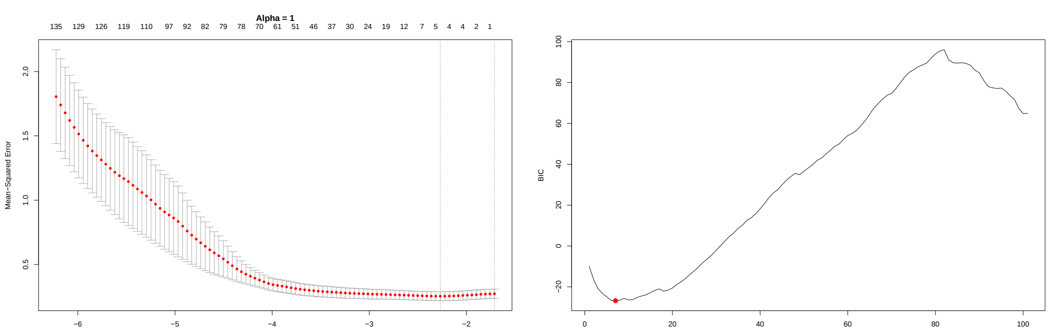
Elasticnet	Forward Selection
cantonSan Cristobal (+)	AGUAAGUA ENTUBADA PRIVADA (-)
CTransitoriosPIMIENTO (-)	cantonSan Cristobal (+)
CTransitoriosSANDIA (-)	CTransitoriosTOMATE RINON (-)
PastosKING GRASS (-)	PastosMIEL O SETARIA (+)
pc4None (+)	pc6 (-)

Table 7: Modeling of Nonzero Invasive , Top 5 features for each method



(a) Cross-validation plot for Elasticnet (b) Forward Selection plots

Figure 9: Invasive Binary Variable Selection. The Elasticnet plot shows adding variables does not improve just predicting the majority class. In Forward Selection the BIC very quickly levels off.



(a) Cross-validation plot for Elasticnet (b) Forward Selection plots

Figure 10: Nonzero Invasive Variable Selection. Similar to Figure 9 the lack of benefit from adding additional variables demonstrates a weak linear relationship.

For linear regression the optimal models for Elasticnet included 4 variables with a RMSE still of 0.25 and for Forward Selection included 7 variables with a R^2 slightly increasing to 0.35. 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.

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 thought of progressing from land that has been most worked by the farmer to least worked. A primary concern is that unworked land may be most susceptible to invasive species growth. For this reason we sort the farms in ascending order first by percent invasive species, and then by percent brush, percent pasture, etc. and graph the stacked bar plots in Figure 11. Since percent brush is given precedence in sorting the purple area stands out more than the rest. Even so, it can be seen that there are farms primarily made up of each of the categories as well as farms fairly evenly split between the categories.

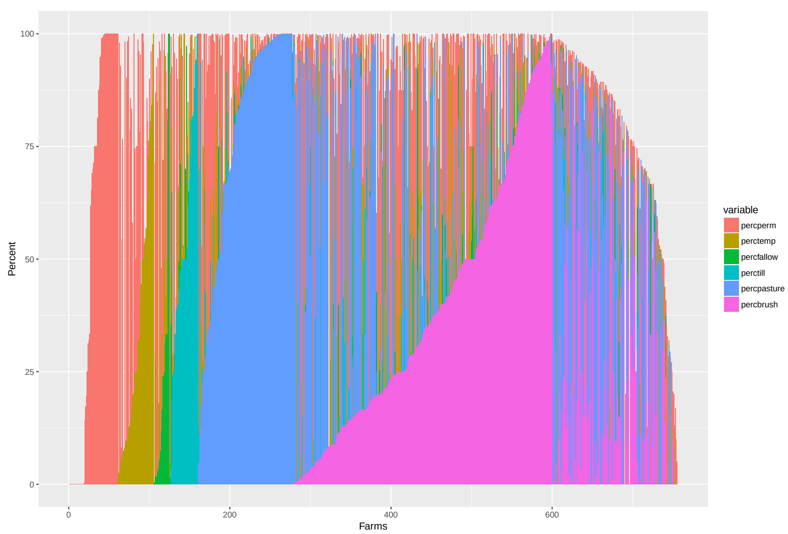


Figure 11: Stacked Barplot of Land Use. Farms are sorted in ascending order by percent invasive and then by percbrush, percpasture, etc. Some bars do not total 100% because percent invasive and percent other categories are not included. There appear to be different classes of farms, some primarily permanent crops, some primarily temporary crops, etc.

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 8.

Majority Category	Number of Farms
Percperm	209
Perctemp	50
Perctill	36
Percpasture	271
Percbrush	189

Table 8: Landuse Category Count

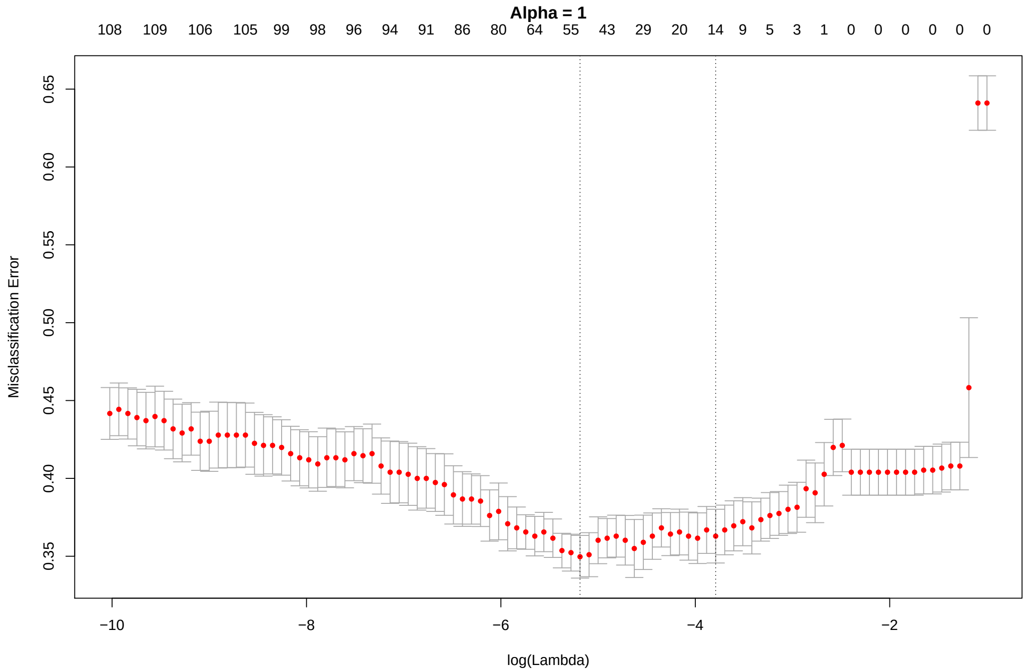


Figure 12: Land Use Cross-validation plot for Elasticnet. The Elasticnet model shows a large improvement over just predicting the majority class. Since there are now 5 categories, such a strategy would result in a 0.65 misclassification rate.

Using this derived categorical variable we used multinomial logistic regression with Elasticnet 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 9. Different categories may have varying number of coefficients and because Perctill had no variables included at this lambda level, it is left out of the table.

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

Table 9: Modeling of Landuse , Top features for each category

The plot of the Elasticnet cross-validation is shown in Figure 12. The optimal model achieves 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 16 and 17.

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 much sense.

This calls for the use of logistic regression, 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 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 Elasticnet 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 $\beta\mathbf{X}$). 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 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.

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

The technique that is used in this analysis is a combination of the Ridge and LASSO penalties, called ElasticNet. 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. ElasticNet, 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 Elasticnet 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. First, is that relationships discovered in this analysis are correlational 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 Elasticnet 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 [3]). While there are 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., $x_1 x_2$) 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. 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] Friedman, J., Hastie, T., & Tibshirani, R. (2009). glmnet: Lasso and elastic-net regularized generalized linear models. R package version, 1(4).

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

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

[5] 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.

[6] 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.

[7] 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

[8] 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	Elasticnet	Forward Selection
Intercept	3.1334	3.3349
CPermanentesOTROS BANANOS	0.1866	0.2765
CPermanentesPLATANO	0.1366	0.2575
CTransitoriosTOMATE RINON	0.1167	0.2057
CTransitoriosYUCA	0.0119	NA
numcultivo	$5.9400 \cdot 10^{-06}$	$8.7700 \cdot 10^{-06}$
PastosBRACHIARIA	-0.1595	-0.2626
PastosELEFANTE	-0.0916	-0.2081
PastosKING GRASS	-0.0648	NA
pc4None	0.0507	NA
pc6	-0.0017	-0.0020
ArbolesGUABA	0.0804	0.1415
ArbolesGUANABANA	0.0001	NA
ArbolesGUINEO	0.0904	0.3332
ArbolesLIMON MANDARINA	0.0631	0.2042
ArbolesNARANJA	0.0281	NA
ArbolesPLATANO	0.0176	NA
ad11	$1.8400 \cdot 10^{-05}$	NA
produccionenlibrasproductovendido	$7.7400 \cdot 10^{-06}$	$2.0000 \cdot 10^{-05}$
v3	$-9.6300 \cdot 10^{-05}$	NA
v30a	-0.0013	NA
c12	$3.0100 \cdot 10^{-06}$	$1.1900 \cdot 10^{-05}$
a7a	$2.7500 \cdot 10^{-06}$	NA
ga9Si	0.0235	NA
ga9a	$7.4700 \cdot 10^{-06}$	0.0001
ga15cualADECUACION UPA	-0.2915	-1.5240
ga15cualMANTENIMIENTO DE CAFdb	-0.2769	-1.9386
ga15cualPACHETE	0.6444	1.8037
e30	0.0176	NA
percperm	0.0018	NA
perctemp	0.0043	NA
perctill	-0.0009	-0.0081
percpasture	-0.0071	-0.0109
percinv	-0.0027	-0.0105
percbrush	-0.0052	-0.0096
percinv2	-0.0023	NA
d3Si	-0.0261	NA
ReclassCONSERVACION	-0.1550	-0.3375
ReclassPECUARIO	-0.0675	-0.1137
ABANDONEDTRUE	-0.0649	NA
CONSERVATIONTRUE	-0.0503	NA
FORESTRYTRUE	-0.0849	-0.1455
LODGINGTRUE	-0.0135	-0.1298
ENERGIAELENERGIA SOLAR PRIVADA	-0.6549	-1.4008
VIASDEACASFALTADA	-0.0637	-0.1335
RELIEVEABRUPTO	-0.0638	NA
RELIEVEPLANO	NA	0.1413

Table 10: Full coefficient list for Production model

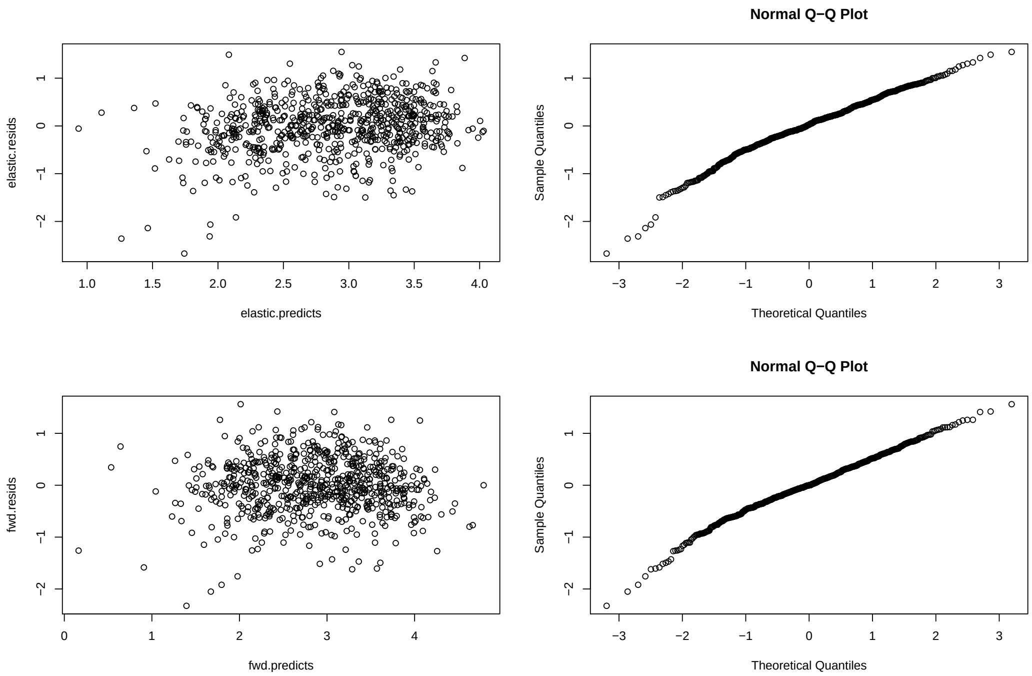


Figure 13: Diagnostic Residual plots for Production. The plots on the left show the predicted values vs. the residuals and both demonstrate a desirable lack of pattern. The plots of the right examine the normality of the residuals, both staying close to the desired straight diagonal pattern.

Variable	Elasticnet	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 11: Full coefficient list for Net Income model

Variable	Elasticnet	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 12: Full coefficient list for Binary Workers model

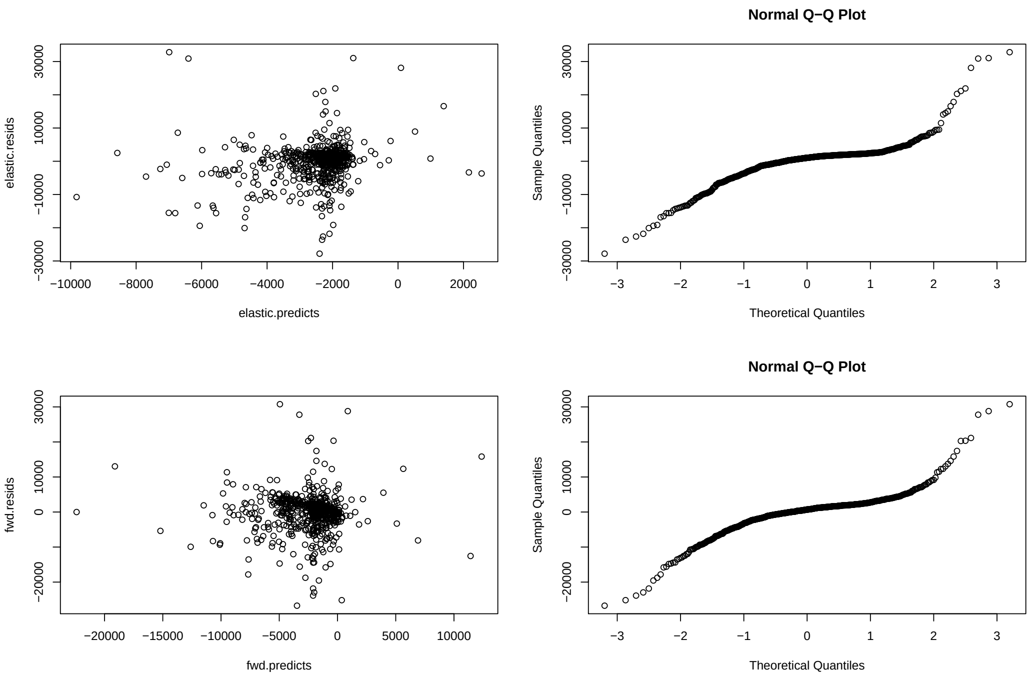


Figure 14: Diagnostic Residual plots for Net Income. The plots of the predicted values vs. the residuals both demonstrate a desirable lack of pattern. The plots examining the normality of the residuals do not follow a diagonal line as closely as would be hoped, but are not an extreme departure.

Variable	Elasticnet	Forward Selection
Intercept	0.7029	0.5775
s4	0.0012	0.0029
v3	0.0006	NA
v44	$1.8600 \cdot 10^{-05}$	NA
GastosAgricolas	$3.4900 \cdot 10^{-06}$	$3.9600 \cdot 10^{-05}$
biokSi	0.0355	0.3634
CTransitoriosCILANTRO	NA	0.2809
ArbolesLIMON MANDARINA	NA	0.2497
a24f	NA	$5.3300 \cdot 10^{-05}$
ga15cualLIMPIEZA DE TERRENO	NA	0.5315
FARMINGTRUE	NA	-0.1272

Table 13: Full coefficient list for Nonzero Workers model

Variable	Elasticnet	Forward Selection
Intercept	-1.3373	-1.0550
cantonSan Cristobal	NA	0.1935
cantonSanta Cruz	NA	-1.7710
cantonFloreana	NA	-16.5993
v30a	NA	0.0076
CTransitoriosMAIZ SUAVE CHOCLO	NA	0.9028
CPermanentesNARANJA	NA	-0.6912

Table 14: Full coefficient list for Binary Invasive model

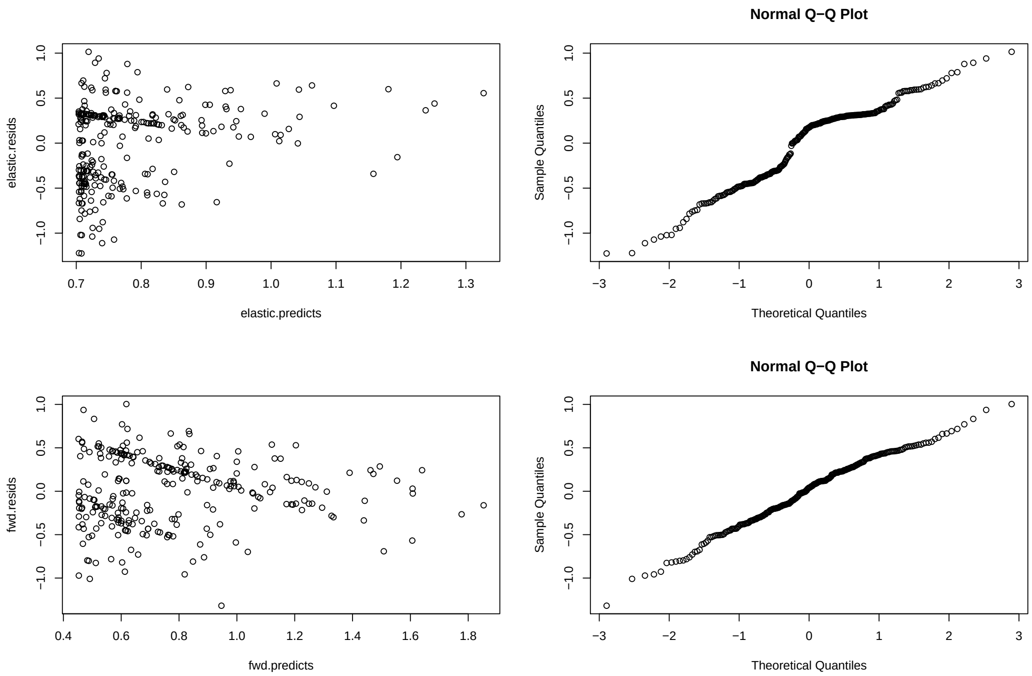


Figure 15: Diagnostic Residual plots for Nonzero workers. The plots of predictions versus residuals shows points bunch at the left end, but the QQ plot closely follows a diagonal line.

Variable	Elasticnet	Forward Selection
Intercept	1.0880	1.3634
cantonSan Cristobal	0.0362	NA
CTransitoriosSANDIA	−0.1068	−0.5569
PastosKING GRASS	−0.2365	−0.6140
pc4None	0.0729	0.2020
CPermanentesPAPAYA	NA	−0.3602
a7c	NA	−0.0035
AGUAAGUA DE POZO PUBLICA	NA	−1.2438
TELEFONOFNO TIENE	NA	−0.2369

Table 15: Full coefficient list for Nonzero Invasive model

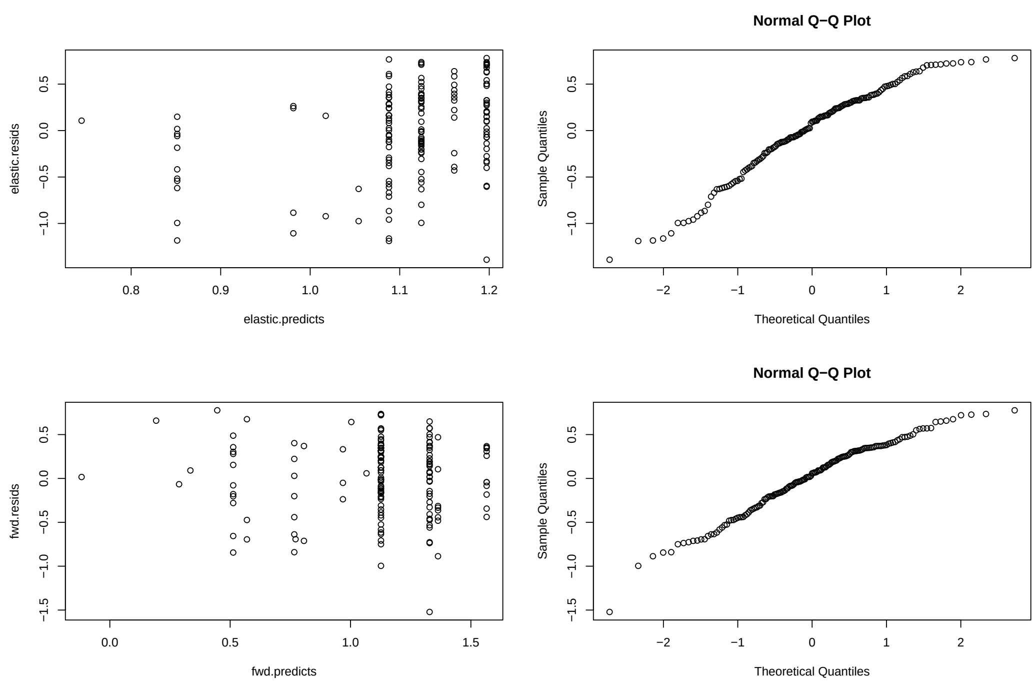


Figure 16: Diagnostic Residual plots for Nonzero Invasive. Due to only a few variables being included in these models and some of them being discrete, the predictions occur at a limited number of distinct points.

Variable	Percperm	Perctemp	Perctill
Intercept	0.0701	-1.0310	-0.9879
s4	-0.0006	NA	NA
CPermanentesAGUACATE	0.0272	NA	NA
CPermanentesCAFE	0.9865	NA	NA
CPermanentesGUABA	0.2531	NA	NA
CPermanentesMANDARINA	0.0107	NA	NA
CPermanentesNARANJA	0.8371	NA	NA
CPermanentesOTROS BANANOS	0.1066	-0.1408	NA
p9	0.0007	NA	NA
VentaLibras	$1.0800 \cdot 10^{-05}$	NA	NA
CTransitoriosAPIO	-0.1273	NA	NA
CTransitoriosBROCOLI	-0.3912	NA	NA
CTransitoriosFREJOL TIERNO	-0.2355	NA	NA
CTransitoriosMAIZ DURO CHOCLO	-0.0389	0.9800	NA
CTransitoriosTOMATE RINON	-0.0138	NA	NA
CTransitoriosVAINITA	-0.0365	NA	NA
pc4None	0.2780	NA	NA
ArbolesMANDARINA	0.0896	NA	NA
ga15cualADECUACION UPA	0.1027	NA	NA
to57e	0.0009	NA	NA
biokSi	0.2364	NA	NA
ENERGIAELNO TIENE	-0.0610	NA	NA
INTERNETINTERNET PRIVADO	-0.3405	NA	NA
ALCANTARILALCANTARILLADO PUBLICO	0.1760	NA	NA
ALCANTARILSIN INFORMACION	-0.3244	NA	NA
RELIEVEONDULADO	-0.0505	NA	NA
CTransitoriosCOL	NA	0.1142	NA
CTransitoriosHIERVITA	NA	0.0422	NA
CTransitoriosMELON	NA	0.6251	NA
CTransitoriosNABO	NA	1.7678	NA
CTransitoriosPIMIENTO	NA	0.2115	NA
librasvendida	NA	$1.4800 \cdot 10^{-05}$	NA
ga15cualPERSONAL PARA SEMBRAR	NA	2.3489	NA
e30f	NA	0.9146	NA
o4f	NA	NA	0.0012
ga15cualGASTOS HERRAMIENTAS	NA	NA	1.5543
ga15cualHERRAMIENTA DE TRANAJO	NA	NA	1.4120
ga15cualPARA PREVERCION MEDICINA	NA	NA	1.6878
GastosPecuarios	NA	NA	$1.2500 \cdot 10^{-06}$

Table 16: Full coefficient list for Landuse model

Variable	Percpasture	Percbrush
Intercept	1.4584	0.4904
cantonSan Cristobal	-0.0682	NA
CTransitoriosCILANTRO	0.0156	NA
PastosBRACHIARIA	0.1537	NA
PastosKING GRASS	0.1639	NA
pc6	0.0033	NA
ArbolesPAPAYA	-0.2863	NA
v3	0.0076	NA
v30a	0.0130	NA
v45	0.0003	NA
o4b	0.0002	NA
e29f	0.0006	NA
d3Si	0.1251	NA
ENERGIAELENERGIA SOLAR PRIVADA	0.0258	NA
cantonSanta Cruz	NA	-0.4764
c10Si	NA	-0.0209
c14	NA	0.0051
r3	NA	0.0731
CPermanentesNARANJILLA	NA	0.2271
CPermanentesPINA	NA	0.1021
t28	NA	$1.4900 \cdot 10^{-05}$
ArbolesAGUACATE	NA	0.3216
ArbolesGUAYABA	NA	0.7167
ArbolesNARANJA	NA	0.0358
ArbolesNARANJA AGRIA	NA	-0.3328
oe	NA	0.0004
ForestalAGUACATE	NA	0.0396
ForestalCEDRELA	NA	0.4316
ga15cualHERRAMIENTAS DE TRABAJO	NA	0.5174
ga15cualLIMPIEZA DE LA FINCA	NA	0.2210
ga15cualTRABAJADORES	NA	0.4183
GastosAgricolas	NA	$-2.9500 \cdot 10^{-05}$
AGUAAGUA ENTUBADA PUBLICA	NA	-0.0656
AGUANO TIENE	NA	0.0913
ALCANTARILPOZO SEPTICO O CIEGO PRIVADO	NA	0.2335
VIASDEACASFALTADA	NA	0.1182
VIASDEACSENDERO	NA	0.2151

Table 17: Full coefficient list for Landuse model (cont.)