Survival of the Fittest: Variable Selection on Galapagos Agricultural Data

Michael Bostwick
Department of Statistics and Operations Research
University of North Carolina at Chapel Hill
mgb2188@live.unc.edu

Abstract

This paper...

1 Introduction

1.1 Background

The Galapagos 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 Galapagos Islands than other systems. Yet the Galapagos 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 Galapagos, but with limited interaction parameters between agents, particularly in regards to farm success walsh. 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 sort through a large number of possible relationships and clarify the empirically most significant ones for future study and incorporation into simulations.

1.2 Data

The data available to study the dynamics between agricultural measures and related factors primarily comes from the Census of Agricultural Production Units (UPA) of Galapagos (Censo de las Unidades de Producción Agropecuaria (upa) de Galápagos). This is a self-reported survey with data from 755 farms detailing the production and sale of crops and livestock, agricultural expenses, and land use decisions.

In addition to the census, data is also available from ... including information on water, energy and road access.

In total, under the direction of the client 240 variables were selected for consideration in modeling relationships between predictors and the outcome variables of interest.

2 Modeling

2.1 Challenges to address

The primary challenge in this analysis is the vast amount of potential predictor variables. This challenge is twofold; a.) when the number of predictors is large the calculation of a reliable model is difficult and b.) 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 several variable selection techniques and a comparison of their performance.

These variable selection techniques aim to build a linear model with a subset of the available variables that still maintains a strong predictive performance.

Secondarily, when performing standard linear regression the outcome variable is assumed to be roughly normally distributed. When this is not true, as is the case for several outcome variables in this study, a poorly fitting model will be found with unreliable 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 brief overview of the statistical methods used is presented here to allow for understanding of results, but for further details see Section refstatmethods Statistical Methods. For each of the outcome variables of interest we build a set of linear models using the appropriate subset of predictors. Each relationship is modeled using Best Subset, Forward Selection and ElasticNet regression. Best Subset considers all variable combinations to find the linear model of best fit. 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. ElasticNet regression fits a linear model by constraining 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 sparser model, but variable selection can be carefully tuned as optimization is done in a more continuous way.

3 Results

3.1 Farm Success

We define farm success with three different measurements: 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 difference between the earnings from all products sold and total expenses. Number of workers supported is calculated as the total labor expenditures divided by a standard worker's salary.

3.1.1 Productivity

Response variable plots and discussion

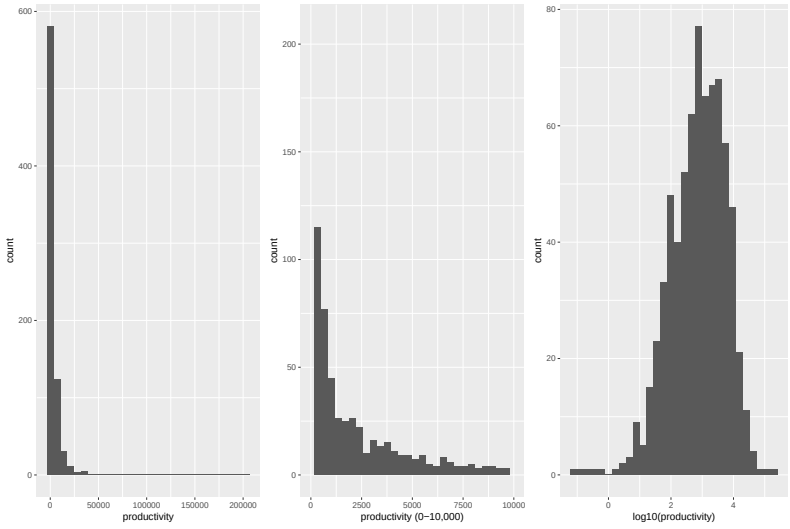

Figure 1: Histogram of Productivity...

Bootstrap plots

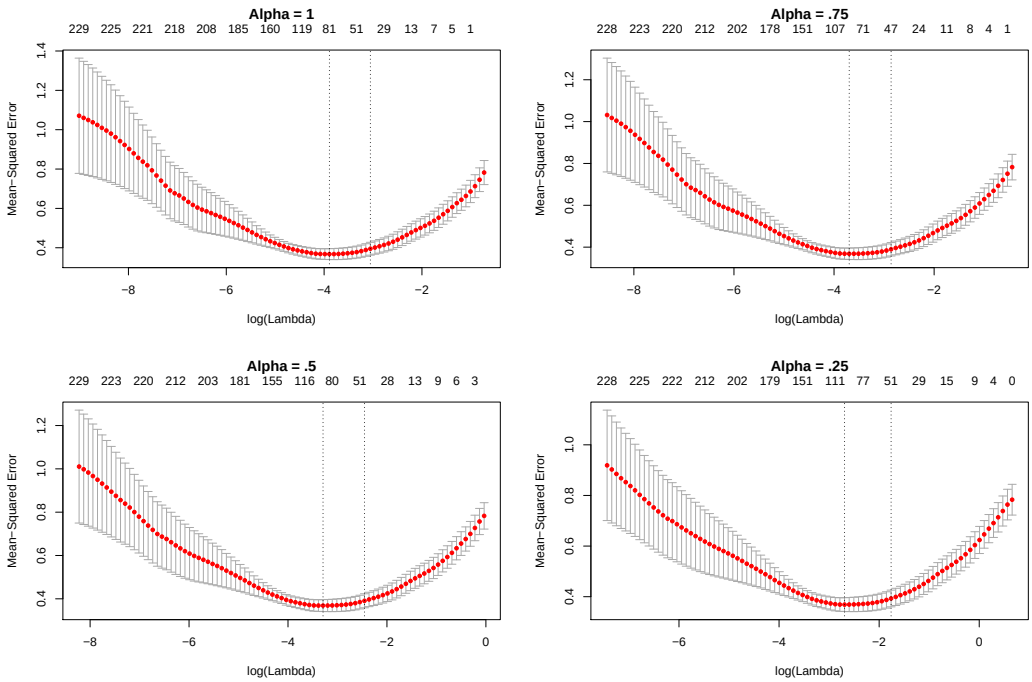

Figure 2: Cross Validation plots for Elasticnet...

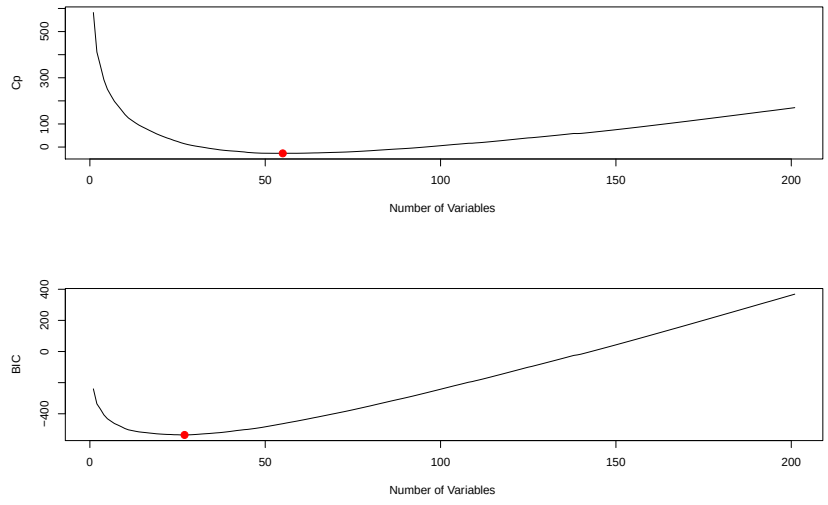

Figure 3: Number of variables Forward Selection plots...

Elasticnet	Forward Selection	Best Subset
pc4None	Arboles AGUACATE	Arboles AGUACATE
pc6	CPermanentes PAPAYA	CPermanentes PAPAYA
percpasture	percbrush	e29
percperm	percinv	percbrush
v30 a	percpasture	percinv

Table 1: Modeling of Productivity, Top 5 features for each of the three methods

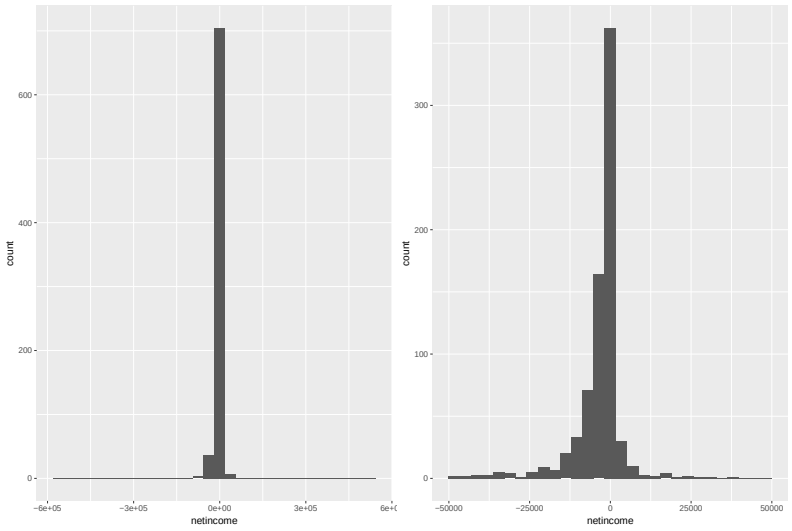

Figure 4: Histogram of Net Income...

Elasticnet	Forward Selection	Best Subset
e30	e30 f	e30 f
s4	Forestal AGUACATE	Forestal AGUACATE
v3	produccion en libras producto vendido	Pastos TANZANIA
v44	productivity	productivity
v45	v30 a	v30 a

Table 2: Modeling of Net Income, Top 5 features for each of the three methods

3.1.2 Net Income

3.1.3 Number of Workers Supported

3.2 Invasive Species

3.3 Land use choices

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 equation 1 above, where $\mathbf{Y}$ is a $n \times 1$ vector of the outcome variable, $\mathbf{X}$ is a $n \times p$ matrix of the predictor variables, β is a $p \times 1$ vector of variable coefficients and ϵ is the error term. The standard linear model assumes the outcome variable $\mathbf{Y}$ has a normal distribution, and therefore takes continuous values. When the outcome variable is continuous, but not normal shaped (perhaps 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 labels of 1 and 0 denoting absence/presence of agricultural

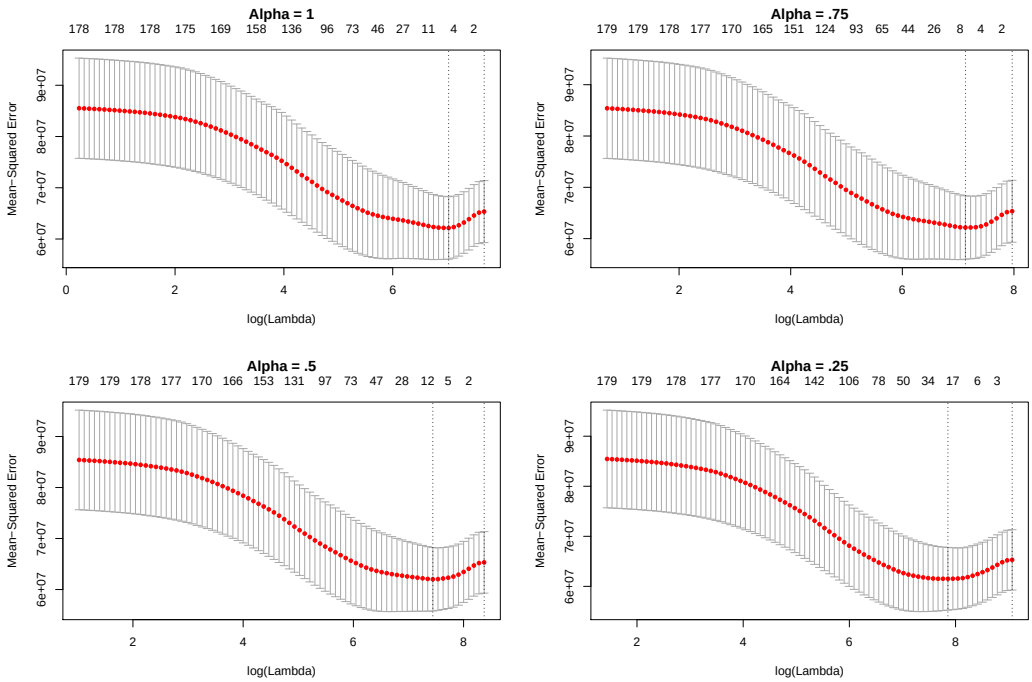

Figure 5: Cross Validation plots for Elasticnet...

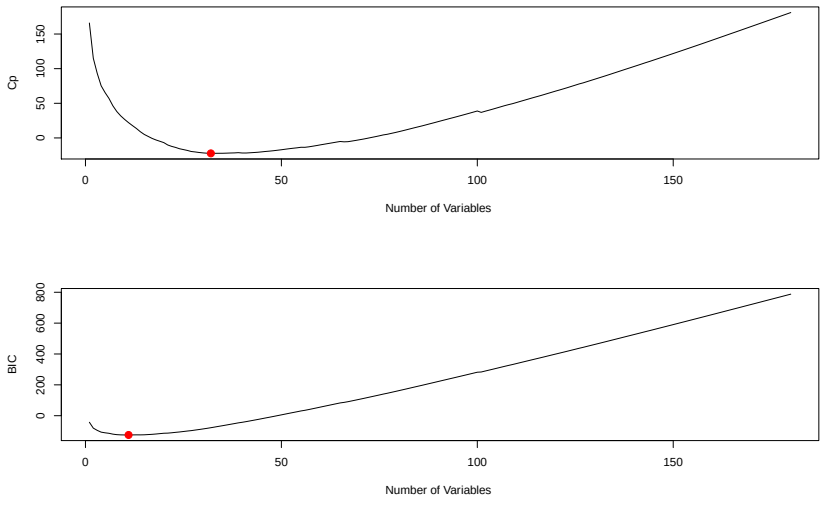

Figure 6: Number of variables Forward Selection plots...

land use) 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 the equation 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. 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.

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

This equation can be rewritten as below

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

Multiclass Logistic Regression

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

4.2 Performance Measures

There are many measures of the fit of linear models. Since 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) and Mallow's C_p (C_p) can used here. BIC and C_p are 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 and C_p, that is the model with the best balance of small size and goodness of fit. BIC puts a large penalty on the inclusion of additional variables, so it will be our primary metric.

Another approach is to use cross-validation. In this technique the dataset is first split into n equally sized sets. Then a model is fit using $n-1$ of the sets (training sets) and evaluated on the remaining 1 (test) set. This is repeated n times, each time reserving a different 1 test set, and then results across the n 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.

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. Recent advances have expanded the problem sizes that can be efficiently computed using Mixed Integer Optimization. Using this technique we find the best 5 variable models for each of the outcomes, but are not able to search all possible model sizes to find the optimal number of variables. In addition, as marked in the results above, sometimes the algorithm can verify that it has found the optimal solution, while other times it returns the best solution found in the allotted time limit (set to 60 seconds), that may or may not be the overall optimal solution.

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 is most beneficial. 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

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

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

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

As shown in equation ? in the standard linear model, we try to find the β 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. Further explanations of this can be found in ???

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

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 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 deal with correlated predictor variables. Both the λ and the α can be set using cross-validation (as discussed above) to be appropriate values for the particular dataset.

4.5 Comparison of methods

5 Conclusion

References

6 Appendix

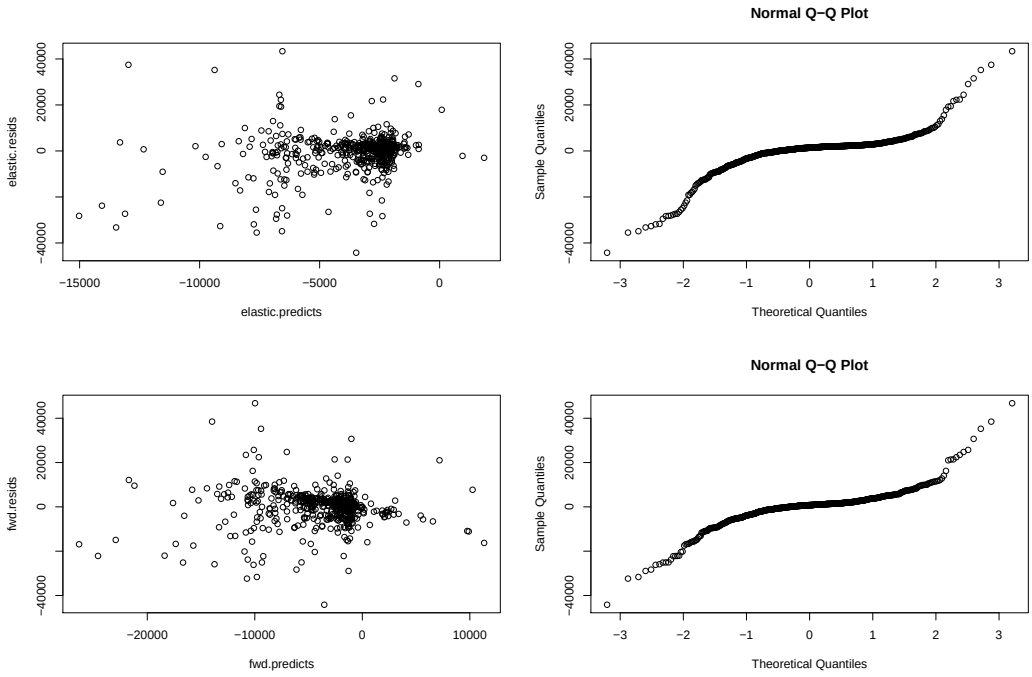

Figure 7: Diagnostic Residual plots for Net Income...

Variable	Elasticnet	Forward Selection
(Intercept)	3.080041181	3.313333818
‘CPermanentes OTROS BANANOS‘	0.182086236	0.276145942
CPermanentes PLATANO	0.132925981	0.259550932
‘CTransitorios TOMATE RINON‘	0.117299266	0.196286142
CTransitorios YUCA	0.014129158	NA
num cultivo	5.74E-06	1.09E-05
Pastos BRACHIARIA	-0.158926563	-0.236139401
Pastos ELEFANTE	-0.082979328	-0.245193118
‘Pastos KING GRASS‘	-0.072805602	NA
pc4None	0.079852208	NA
pc6	-0.00165915	-0.002374897
Arboles GUABA	0.078692352	0.18178855
Arboles GUANABANA	0.006025003	NA
Arboles GUINEO	0.089675012	0.289877302
‘Arboles LIMON MANDARINA‘	0.061406497	0.21056596
Arboles NARANJA	0.027558501	NA
Arboles PLATANO	0.020059222	NA
ad11	2.14E-05	NA
produccion en libras producto vendido	6.99E-06	2.03E-05
v3	-0.000255643	NA
v30 a	-0.001316099	NA
c12	3.13E-06	1.10E-05
a7 a	5.02E-06	NA
ga9Si	0.022969822	NA
ga9 a	9.25E-06	NA
ga15 cualADECUACION UPA	-0.287706991	NA
ga15 cualMANTENIMIENTO DE CAF<db>	-0.275504129	NA
ga15 cualPACHETE	0.627506972	NA
e30	0.019677747	NA
percperm	0.002121402	NA
perctemp	0.004543186	NA
perctill	-0.000540854	NA
percpasture	-0.005457599	-0.010533656
percinv	-0.002085349	-0.010890599
percbrush	-0.004802799	-0.009326194
percpasture2	-0.000971254	NA
percinv2	-0.002518628	NA
d3Si	-0.027935711	NA
ReclassCONSERVACION	-0.151841745	NA
ReclassPECUARIO	-0.069947042	NA
ABANDONEDTRUE	-0.064830728	NA
CONSERVATIONTRUE	-0.052017735	NA
FORESTRYTRUE	-0.085830115	NA
LODGINGTRUE	-0.013661245	-0.17332584
‘ENERGIA ELENERGIA SOLAR PRIVADA‘	-0.644529267	NA
VIAS DE ACASFALTADA	-0.063469331	NA
RELIEVEABRUPTO	-0.061862044	NA
ga7 a	NA	3.57E-06
ga15 cualAGUA POTABLE	NA	-0.355650778
ga15 cualMATA RATAS	NA	-0.006609286
ga15 cualPALAS	NA	-0.37928575
percperm2	NA	7.77E-05
ReclassFORESTAL	NA	-0.157748359
‘ReclassSIN APROVECHAMIENTO‘	NA	-0.074688418
‘AGUAAGUA DE POZO PRIVADA‘	NA	0.322600326
‘ENERGIA ELENERGIA SOLAR PUBLICA‘	NA	0.470839499
‘VIAS DE ACCAMINO DE HERRADURA‘	NA	0.025360849
o4 c	NA	0

Table 3: Full coefficient list for Production model

Variable	**Elasticnet**	**Forward Selection**
(Intercept)	-3239.252011	-1037.984538
s4	NA	-26.38665266
CTransitorios BROCOLI	NA	-3884.785218
CTransitorios ZAPALLO	NA	-4182.207378
Pastos SABOYA	NA	4214.218204
produccion en libras producto cosechado autoconsumo	NA	0.123258405
v3	NA	-39.552126
v45	NA	-115.922115
v53 a	NA	55.98292904
e29	NA	-420.5227497
e30	NA	-5313.724658
bio kSi	NA	-8228.217856

Table 4: Full coefficient list for Net Income model