Project 5

Jose Alfaro

Due: April 10, 2019

Question: 1

Part: A

It is good idea to standardize the predictors before performing any analysis whenever the variables are not on the same scale. This way, variables with larger values do not dominate those with smaller values when conducting any sort of analysis. In the Caravan dataset, we saw that many predictors contained small values, but there were a handful with larger values that would overshadow the other predictors. For this reason, we standardized the dataset.

Note: A preview of the dataset can be seen below in which the first five variables and the last five variables are displayed to emphasize the different scales between predictors.

Part: B

The data was standardized and split into training and testing sets in which the first 1,000 observations were assigned to be part of the "testing" dataset and the remaining observations were assigned as the "training" dataset.

Part: C

A logistic regression model was fit in which the estimated probability for the cut-off value was altered from 0.5 to 0.2. This was done to better represent the Caravan dataset provided since only about 20% of the individuals in the dataset purchased the caravan insurance policy. From this model, we were able to calculate the confusion matrix which can be found in *table 1*. From this table, we computed the following:

$$Sensitivity = \frac{18}{18+41} = \frac{18}{59} = 30.51\%$$

$$Specificity = \frac{904}{37+904} = \frac{904}{941} = 96.07\%$$

$$MR = \frac{37+41}{18+37+41+904} = \frac{78}{1000} = 7.80\%$$

$$TestMSE = 0.0501$$

Part: D

Logistic regression was performed again, however, this time the model was fit using a penalized maximum likelihood method (ridge method). Using this method, we found our optimal lambda value to be $\lambda = 0.1377$. We then refit the ridge regression model using our optimal lambda value in order to extract the coefficients for the model which can be seen in *tables 4 and 5*. It is important to note that most of the variable coefficients were greatly reduced, some of which were 100 times smaller than in the original logistic regression from part C. Most of the coefficients were very close to 0. From this model, we managed to calculate the confusion matrix which can be found in *table 2*. From this table, we computed the following:

$$Sensitivity = \frac{4}{4+55} = \frac{4}{59} = 6.78\%$$

$$Specificity = \frac{931}{10+931} = \frac{931}{941} = 98.94\%$$

$$MR = \frac{10 + 55}{4 + 10 + 55 + 931} = \frac{65}{1000} = 6.5\%$$
$$TestMSE = 0.0526$$

Part: E

We conducted logistic regression one more time, but this time the model was fit using a different penalized maximum likelihood method (lasso method). Using this method, we found the optimal lambda value to be $\lambda = 0.0029$. Note that the optimal lambda value using the lasso technique was far smaller than the optimal lambda value that was used for the ridge regression. Once again, it is important to note that the coefficients were significantly smaller using the lasso shrinkage method as opposed to the regular logistic regression coefficients (*tables 4 and 5*). Moreover, the lasso technique shrunk the coefficients of 73 variables down to 0 and essentially removed them from the model. Therefore, the final model obtained using the lasso technique consisted of only 13 out of the original 86 predictors. From this final model, we calculated the confusion matrix which can be found in *table 3*. From this table, we computed the following:

$$Sensitivity = \frac{4}{4 + 55} = \frac{4}{59} = 6.78\%$$

$$Specificity = \frac{933}{8 + 933} = \frac{933}{941} = 99.15\%$$

$$MR = \frac{8 + 55}{4 + 8 + 55 + 933} = \frac{63}{1000} = 6.3\%$$
$$TestMSE = 0.0527$$

Part: F

If our goal is to predict whether or not an individual will purchase a caravan insurance policy, then we recommend using the logistic regression model found in part C since it had the highest sensitivity rate of 30.51% and the lowest test MSE rate of 0.0501. Although the lasso regression model has the highest specificity percentage of 99.15%, and is the most parsimonious model, it had a very low sensitivity percentage of 6.78%. This suggests that the lasso regression model would do the best job in predicting if a customer **will not** buy the insurance policy, but the logistic regression model does the best job in predicting if a customer **will** buy the caravan insurance policy. Since our goal is to predict whether not a customer will buy the insurance policy, then we recommend the logistic model.

Question: 2

Part: A

Exploratory data analysis was conducted on the Mali dataset to get an idea of what the data is portraying. Specifically, we constructed two pairwise scatterplots (*figure 1*). The first plot was Family Vs. DistRD in which we saw that the majority of the data was clustered around towards the bottom right corner of the plot. We also noticed that there were 3 obvious outliers in the data that did not belong. The second plot was DistRD Vs. Cattle in which we noted that a majority of the data was heavily concentrated in the bottom left-hand corner of the plot. Again, we noticed two obvious outliers in the data.

There were a total of five outliers detected visually. These observations were identified and removed from the dataset in order to get a more insightful analysis. The two pairwise scatterplots were replotted with the omission of the five outliers and can be seen in *figure 2*. In the plot consisting of Family vs. DistRD, we noted that the observations were more spread out than they were originally. We also note that with the omission of the outliers, the range of the x and y axis greatly reduced. In the plot consisting of DistRD Vs. Cattle, we saw that with the omission of the outliers the data was more spread out, but there was still a heavy concentration of observations in the bottom left corner. Once again, the omission of the outliers greatly reduced the x and y axis of the plot.

Part: B

After exploring our data in greater detail, we observed that not all of the variables had similar scales. This meant that standardization of the predictors was needed so that variables with larger values do not dominate those with smaller values.

Part: C

The predictors were standardized in order to get all of the variables on a similar scale. Then, we performed principle component analysis in order to visualize the variation present in the Mali dataset within its 9 variables. After conducting PCA, we plotted a scree plot (*figure 3*) and a cumulative proportion of variance explained plot (*figure 4*). From the scree plot, we see that after two principle components, we do not see much more variance explained as the number of components increases. Therefore, we recommend the use of two principle components for the rest of the analysis.

Part: D

After conducting PCA, we were able to put together a table showing the correlations of the standardized variables with the components(*figure 8*). Here we see that the first two columns of the table represent the first two principle components. We also created a table that demonstrates the cumulative percentage of the total variability explained by the first two components which can be found in *figure 7*. Note, in this case, the first two rows correspond to the first two principle components. Furthermore, we note that most of the variables are highly positively correlated with the first component. We saw that the highest correlation here corresponded to the cotton predictor which had a correlation value of 0.9127, and the lowest correlation corresponded to the DistRD predictor which had a correlation of 0.0155. Moreover, in the second column we see that the correlation drops for most predictors and in some cases the predictors become negatively correlated. Here, we saw the strongest correlation between a predictor and the second principle component corresponded to the sorg variable which had a correlation of -0.7239. Furthermore, we saw that the lowest correlation between a predictor and the second component corresponded to the cotton variable which had a correlation value of 0.0107.

Note: These observations can be seen more graphically in *figure 6* in which we visualized the correlation between the predictors and the principle components. Keep in mind that we are only taking account the first two columns which deal with the first two principle components.

Lastly, we displayed the scores on the two components and the loadings on them using a biplot which can be found in *figure 5*. The biplot provides a visual which depicts how strongly each characteristic (vector) influences the principle components. Here we see that Maize, Cattle, Bull, Cotton, and Family have large positive loadings on component 1. Therefore, this component focuses on a farmer's overall farm size with respect to the size of the crop fields, cattle size numbers, and overall number of family members. On the other hand, we saw that Sorg and Millet have large negative loadings while DistRD and Maize had a large positive loading on component 2. Therefore, this component focuses on a farmer's crops and distance to road.

Part: E

As mentioned in part D, the first component may be interpreted as a "Farm Size" component since it has high loadings for the cattle and crop variables. We also stated that the second component could be interpreted as a "crops to distance to road" component since it has high loadings for variables corresponding to crops and the distance to road variable. We used *figures 6 and 8* to help us interpret the third, fourth, and fifth principle components. The third principle component may be interpreted as a "goats and distance to road" component since it has high correlation values with those variables. The fourth component may be interpreted as a "millet and distance to road" component since it is highly correlated to those predictors. Lastly, the fifth component may be interpreted as a "sorghum and millet" since these are the highest correlation values between this principle component and the predictors.

Question: 3

Part: A

Before fitting a linear model using the usual least squares method, we looked at a preview of the dataset and concluded that we needed to standardize the data since the predictors were on drastically different scales. After standardizing the data, we fit a model via the least squares method which resulted in the following model:

$$MPG = 22.415 - 0.835 * Cylinders + 2.509 * Displacement - 0.670 * Horsepower - 5.670 * Weight + 0.218 * Acceleration$$

$$+ 2.862 * Year + 2.630 * Origin2 + 2.8532 * Origin3$$

However, upon viewing the digagnostic plots for variance and normality, we stated that this model violates both of these assumptions since there is not constant variance, the variance is not centered around 0 and the QQ Plot does not follow the 45 degree line closely (*figure 9*).

Part: B

From our initial exploration of the data, we had a feeling that we could drop some of the variables in the model. Therefore, we performed the best-subset selection procedure to find a more parsimonious model. After running best subset selection, we saw that different selection criteria chose different number of predictors. This can be seen in *figure 11* where subsets of predictors are chosen by RSS, R^2_{adj}, C_p, and BIC. The most important thing to note her, is that although each of these four sub-plots describe a different selection criterion and chose a different "optimal" number of predictors, they all have an "elbow" at around the 2 predictors mark. This means that the most parsimonious model is one with only two predictors since adding any more would only increase the R^2_{adj} by a small amount. In other words, the amount of information explained by adding more predictors into the model is not worth the additional level of complexity that comes with adding the additional variable(s). Lastly, the best subset selection gave us a list of 7 optimal models corresponding to the number of variables. This is graphically displayed in *table 6*. From here, we saw that since the optimal number of predictors to include in the model was 2, then by using best subset selection the model chosen is:

$$MPG = \beta_0 + \beta_1 * Weight + \beta_2 * Year$$

This model may also be written as:

$$MPG = 23.446 - 5.633 * Weight + 2.790 * Year$$

However, upon viewing the digagnostic plots for variance and normality, we stated that this model violates both of these assumptions since there is not constant variance, the variance is not centered around 0 and the QQ Plot does not follow the 45 degree line closely (*figure 10*).

Part: C

We implemented the ridge regression technique in order to fit a penalized maximum likelihood model. Using this method, we found our optimal lambda value to be $\lambda = 0.7120$. We then refit the ridge regression model using our optimal lambda value in order to extract the coefficients for the model which can be seen in *table 7*. It is important to note that most of the variable coefficients were reduced in comparison to both the least square estimate coefficients and the best subset model coefficients. From this model, we calculated the test error to be 11.444.

Part: D

We implemented the lasso regression technique in order to fit a penalized maximum likelihood model. Using this method, we found our optimal lambda value to be $\lambda = 0.0046$. We then refit the lasso regression model using our optimal lambda value in order to extract the coefficients for the model which can be seen in *table 7*. It is important to note that most of the variable coefficients were reduced in comparison to both the least

square estimate coefficients and the best subset model coefficients. Also, one should note that the lasso regression technique did not perform model selection in this procedure since none of the coefficients shrunk to a value of 0. From this model, we calculated the test error to be 10.685.

Part: E

We implemented principal component regression (PCR) on the dataset to compare test error rates between different models. Instead of regressing the dependent variable on the explanatory variables, the principle components of the predictors are used as regressors. We fit the PCR model using a built in 10-fold cross validation technique that would help us in computing test errors. We then used these test estimates (computed via 10-fold cross-validation) to plot the principle components against their error rates to produce *figure 12*. Here we saw that there was a clear "elbow" located at the 1 component value which lead us to choose 1 as our principal component. We then, examined the regression coefficients associated with this value of the principle component and calculated the test MSE which can be found in *table 7*.

Part: F

We implemented partial least squares regression (PLS) on the dataset to compare test error rates between different models. This method combines features from PCA and multiple regression and allows us to predict dependent variables from a large set of independent variables. We fit the PLS model using a built in 10-fold cross validation technique that would help us in computing test errors. We then used these test estimates to plot the principle components and produce *figure 13*. Here we saw that there was a clear "elbow" located at the 1 component value which lead us to choose 1 as our principal component. We then, examined the regression coefficients associated with this value of the principle component and calculated the test MSE which can be found in *table 7*.

Part: G

Strictly based off *table 7*, we can see that the least squares, the best subset, the ridge regression and the lasso regression models had the lowest test MSEs out of all of the models tested. However, the best subset model and the least squares model violated the normality and constant variance assumptions which led to these models being invalid. From here, we shifted our attention to the ridge and lasso regression models and saw that the model obtained via the lasso technique had the lowest test MSE between the two models. Therefore, we recommend the lasso regression model due to its small test MSE.

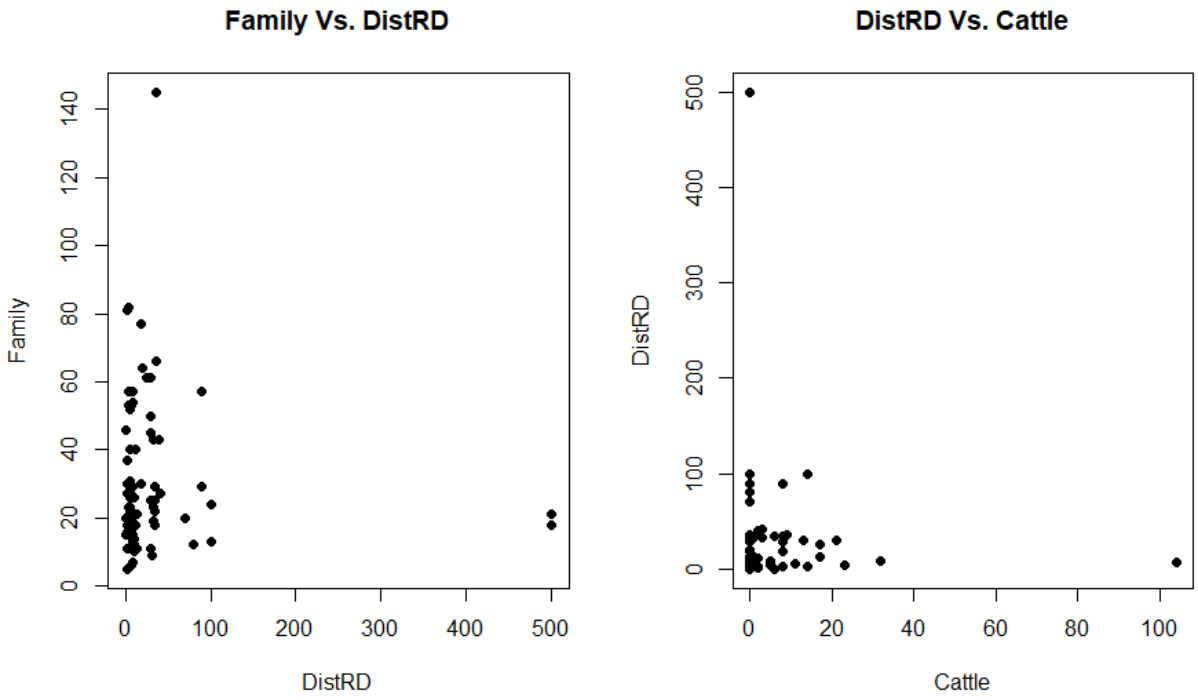

Figure 1: Scatterplots of Family Vs. DistRD and DistRD Vs. Cattle (With Outliers)

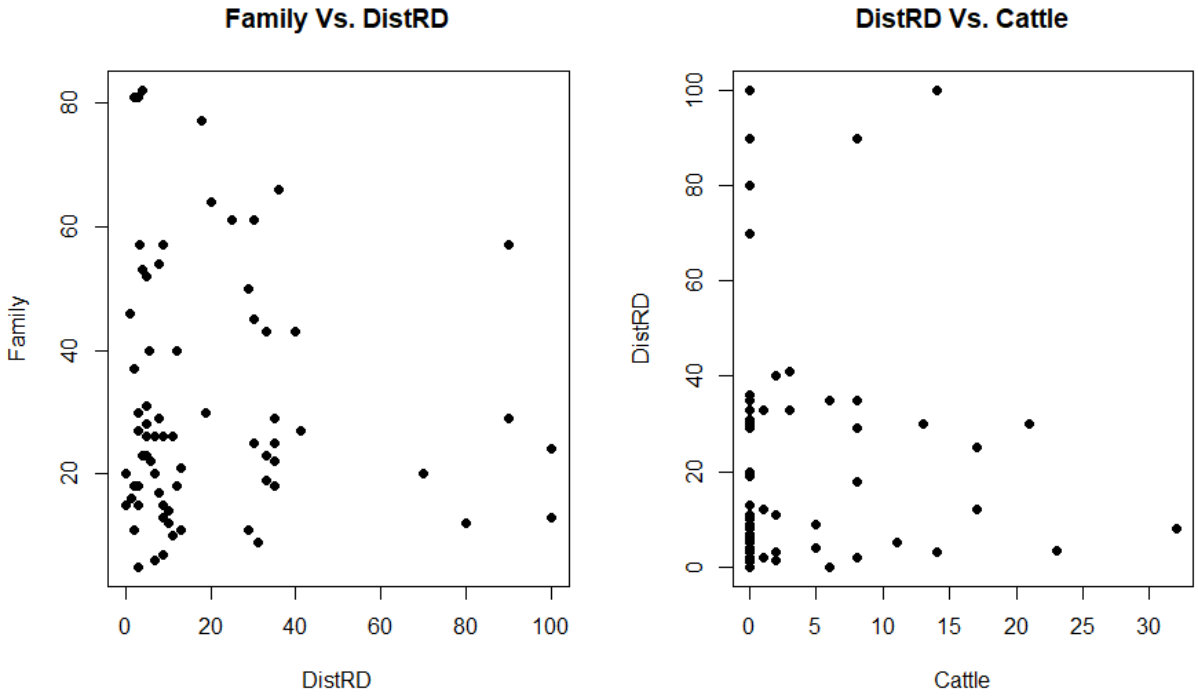

Figure 2: Scatterplots of Family Vs. DistRD and DistRD Vs. Cattle (Without Outliers)

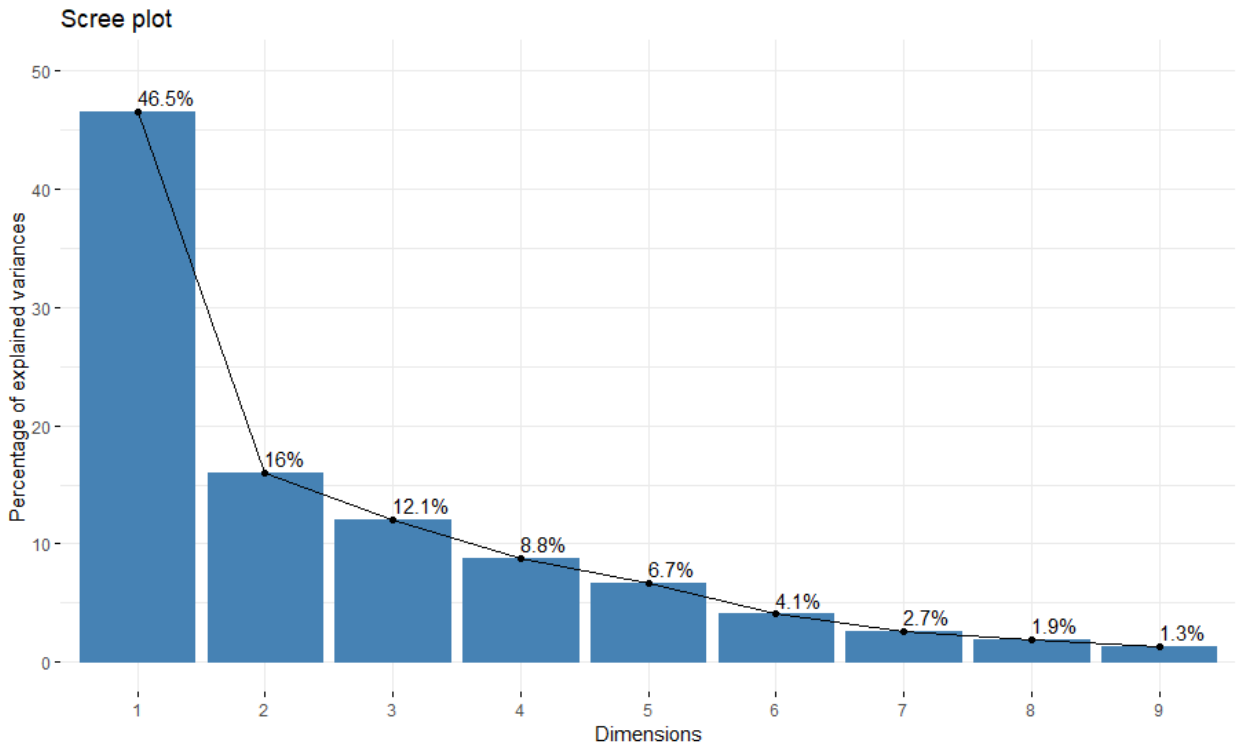

Figure 3: Scree Plot for Mali Dataset

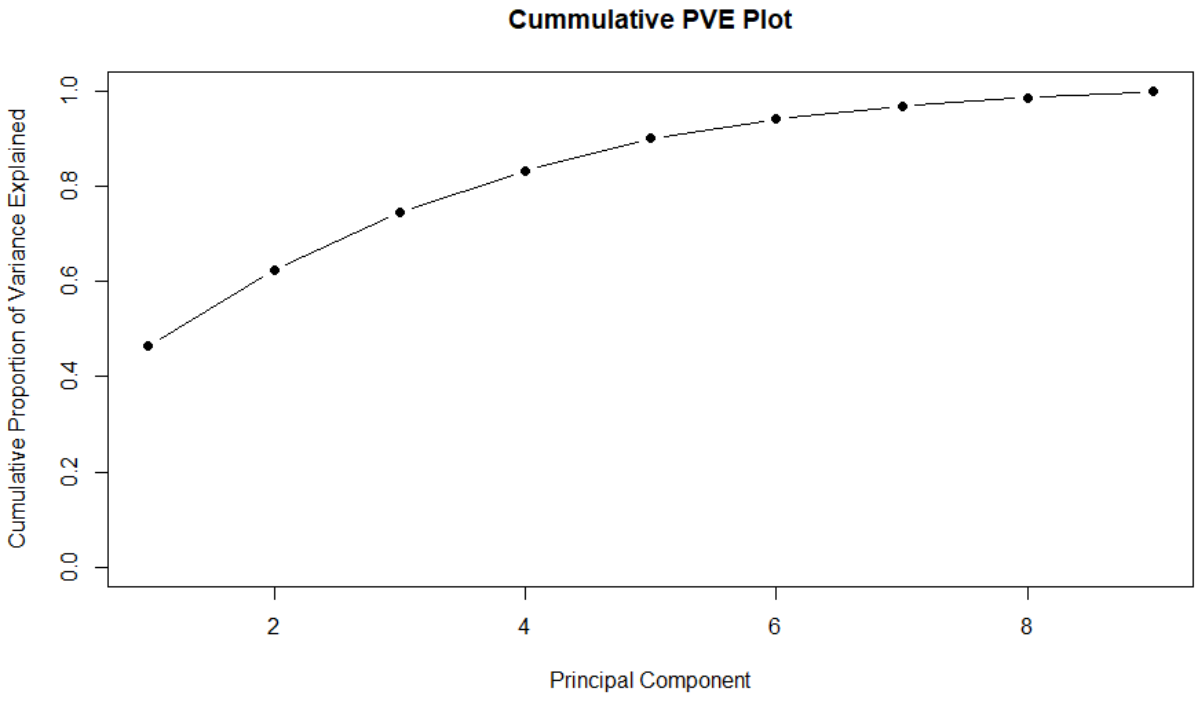

Figure 4: Cumulative Percent of Variance Explained Plot for Mali Dataset

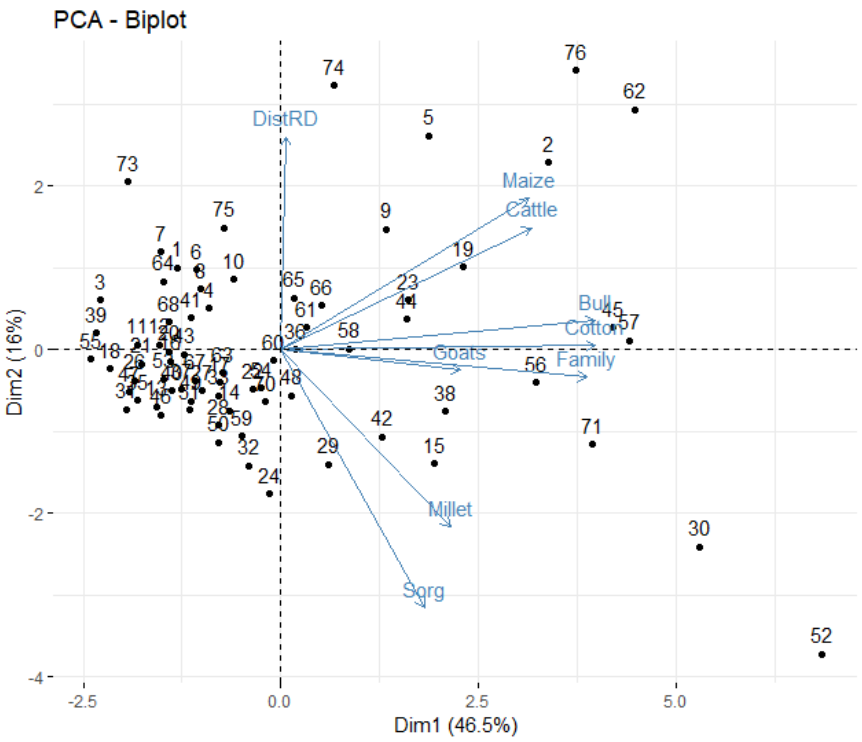

Figure 5: Biplot for Mali Dataset

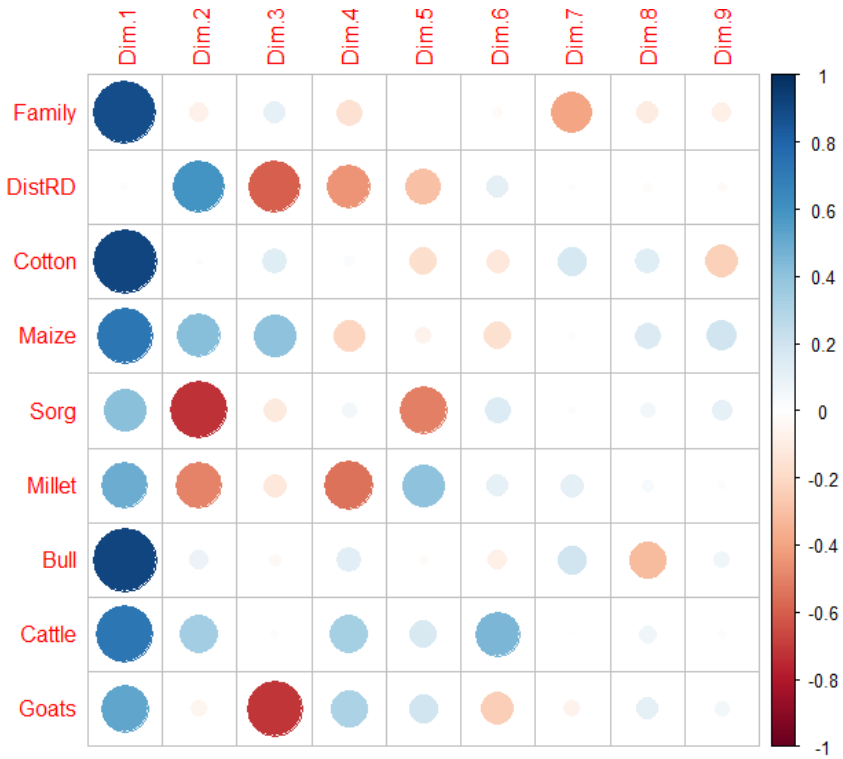

Figure 6: Correlation Plot for Mali Dataset

Principle Component	PVE	PVE (Cumulative)
1	0.4650146	0.4650146
2	0.1597874	0.6248020
3	0.1205000	0.7453020
4	0.0879797	0.8332817
5	0.0671472	0.9004289
6	0.0406818	0.9411107
7	0.0266693	0.9677800
8	0.0190917	0.9868717
9	0.0131283	1.0000000

Figure 7: Table of Principle Components, Percent of Variation Explained, and Cumulative PVE

	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5	Dim.6	Dim.7	Dim.8	Dim.9
Family	0.8875378	-0.0780545	0.1024734	-0.1523421	0.0088055	-0.0241907	-0.3906933	-0.1089396	-0.0854605
DistRD	0.0155212	0.5956090	-0.5920952	-0.4410146	-0.2935929	0.1128993	0.0103225	-0.0198556	-0.0222372
Cotton	0.9126957	0.0106936	0.1375858	0.0243255	-0.1700199	-0.1208284	0.1769833	0.1365773	-0.2320947
Maize	0.7205745	0.4228046	0.4042725	-0.2137443	-0.0615714	-0.1653183	-0.0116742	0.1504593	0.1974594
Sorg	0.4165618	-0.7239195	-0.1161075	0.0520936	-0.5010837	0.1488449	-0.0101016	0.0520485	0.1008530
Millet	0.4917210	-0.4978604	-0.1207597	-0.5484331	0.4096551	0.1093877	0.1179272	0.0319730	0.0164850
Bull	0.9109228	0.0815967	-0.0316457	0.1295536	-0.0219998	-0.0811149	0.1941165	-0.3110991	0.0651813
Cattle	0.7270865	0.3411413	0.0143987	0.3318518	0.1691056	0.4592951	-0.0052107	0.0699134	0.0130850
Goats	0.5207471	-0.0583631	-0.7153907	0.3121457	0.1933128	-0.2433571	-0.0640248	0.1134457	0.0513409

Figure 8: Correlation of the Standardized Values with the Components

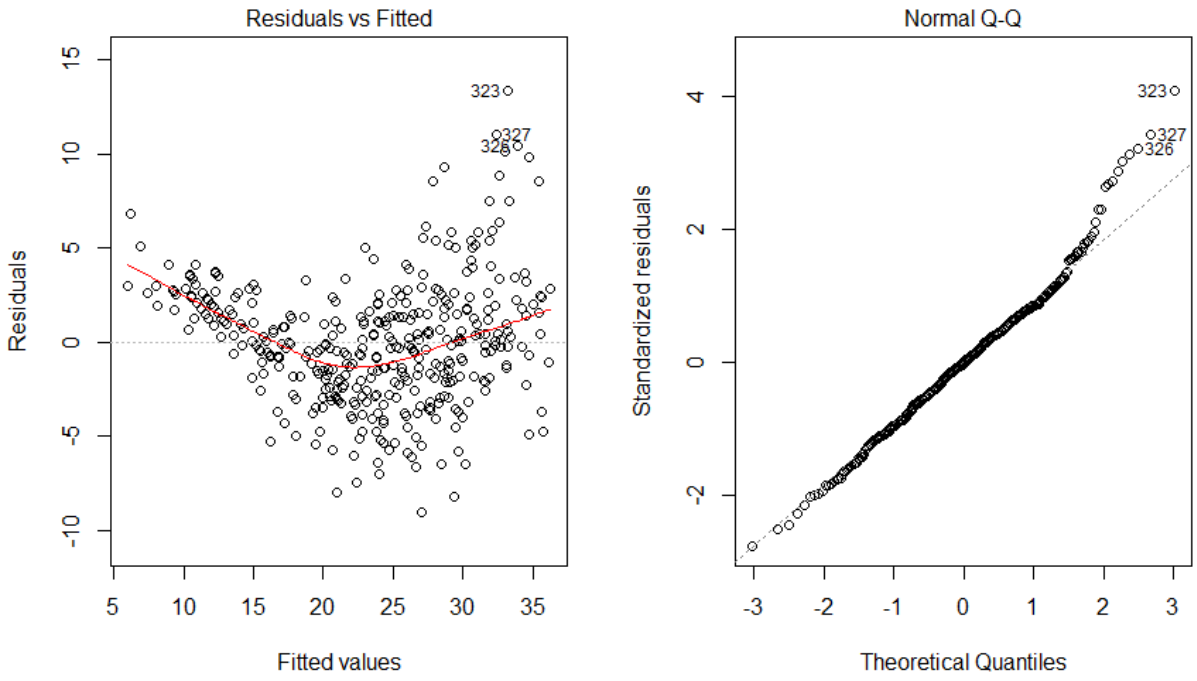

Figure 9: Full Model Diagnostics

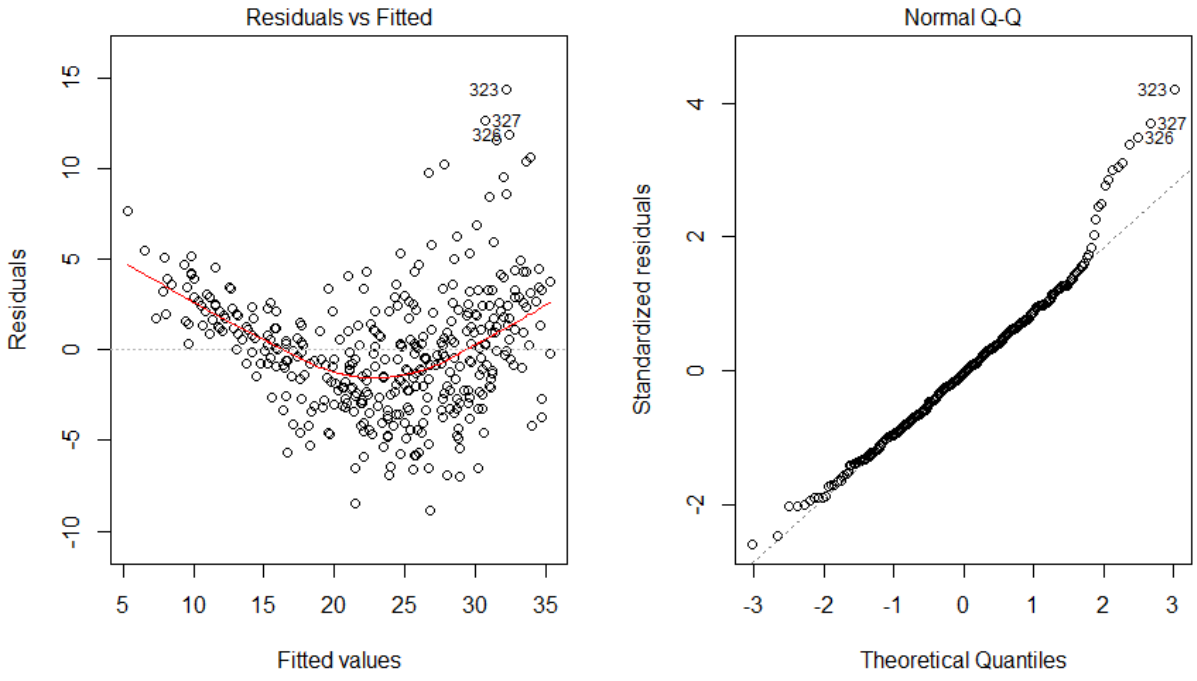

Figure 10: Best Subset Model Diagnostics

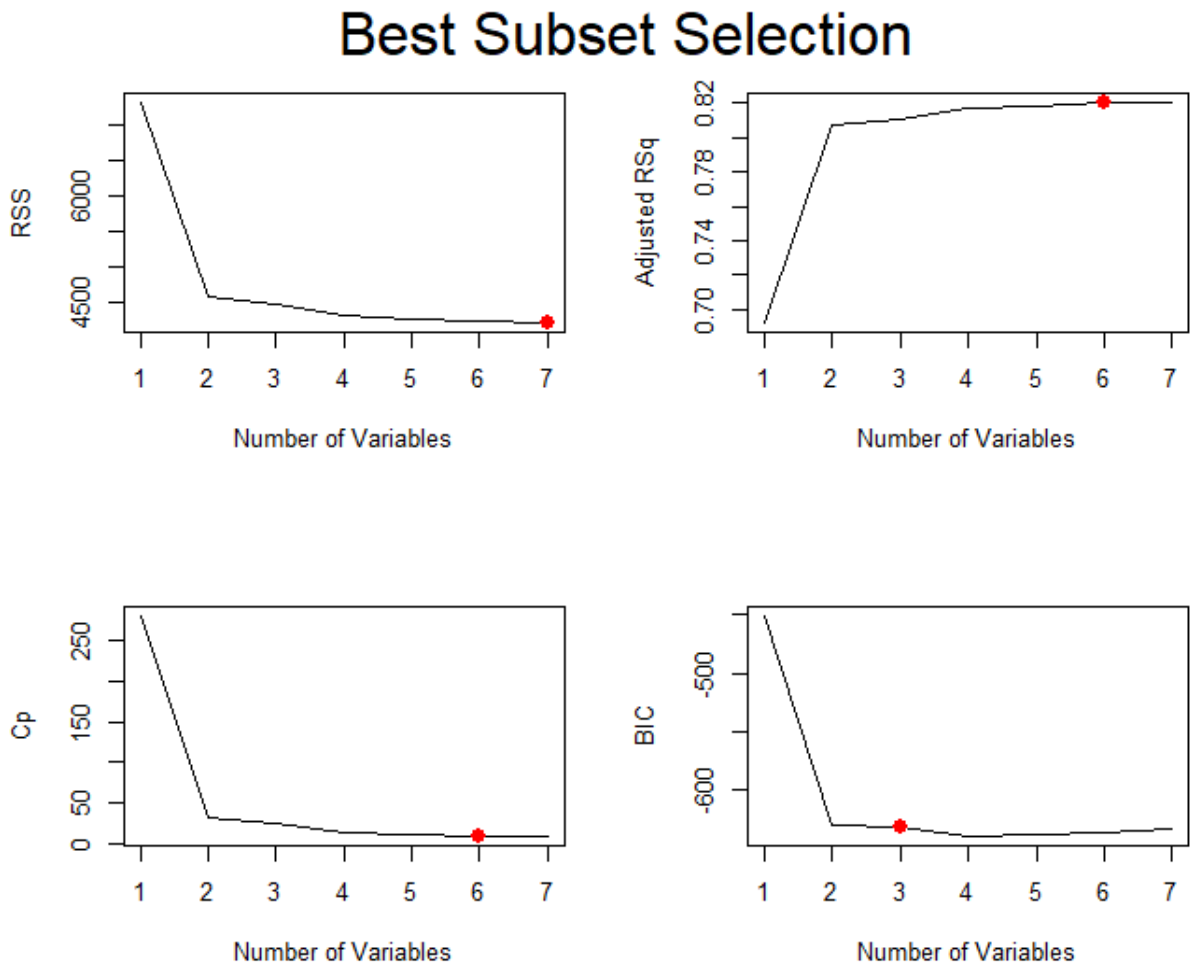

Figure 11: Correlation of the Standardized Values with the Components

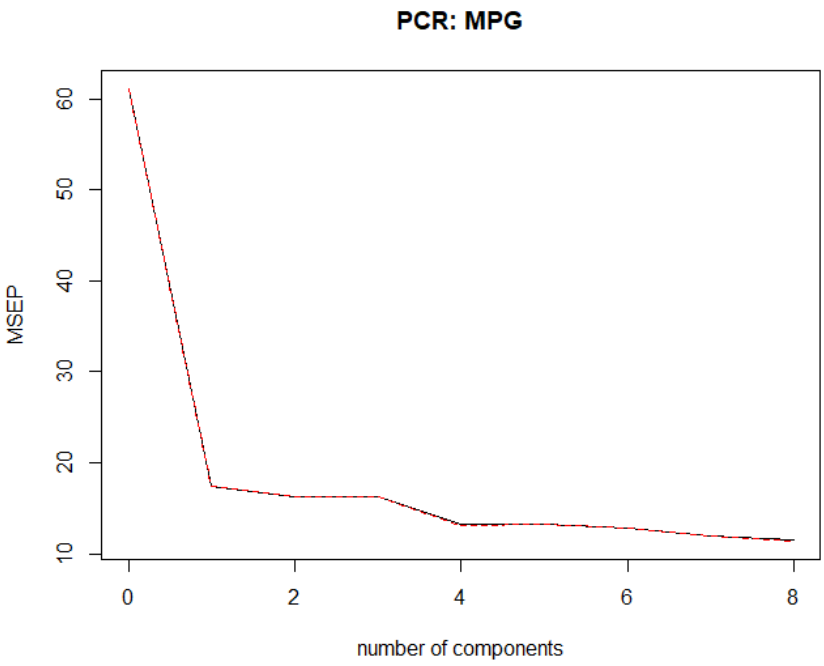

Figure 12: Correlation of the Standardized Values with the Components

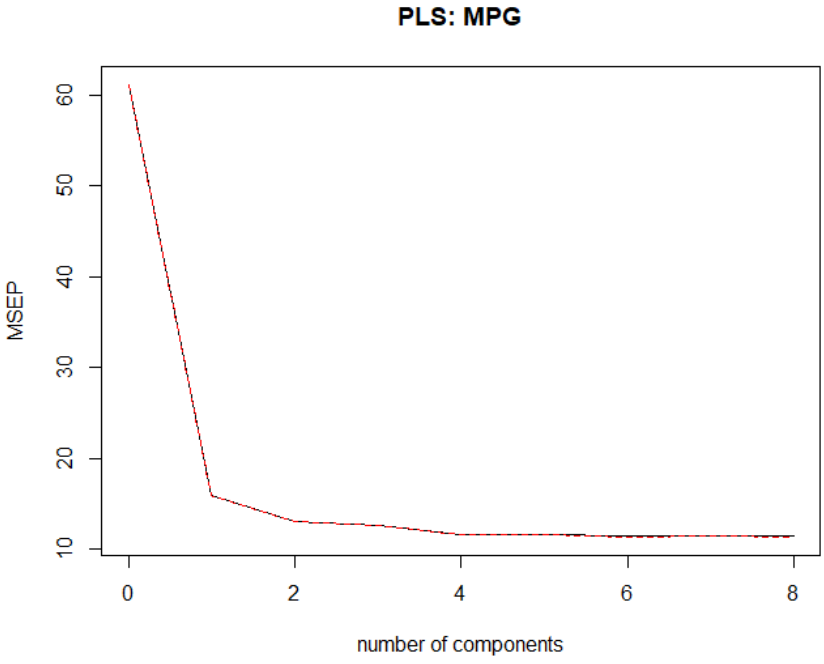

Figure 13: Correlation of the Standardized Values with the Components

Tables

Jose Alfaro

April 7, 2019

		Actual	
		No	Yes
Predicted	No	904	41
	Yes	37	18

Table 1: Confusion Matrix for Logistic Regression

		Actual	
		No	Yes
Predicted	No	931	55
	Yes	10	4

Table 2: Confusion Matrix for Ridge Regression

		Actual	
		No	Yes
Predicted	No	933	55
	Yes	8	4

Table 3: Confusion Matrix for Lasso Regression

Predictors	Logistic	Ridge	Lasso
Intercept	-3.776	-4.176e-02	0.007
MOSTYPE	0.845	-3.368e-05	
MAANTHUI	-0.074	-5.573e-03	
MGEMOMV	-0.021	1.045e-03	
MGEMLEEF	0.171	6.922e-03	

Table 4: First Five Predictors for Logistic, Ridge, and Lasso Regression

Predictors	Logistic	Ridge	Lasso
AZEILPL	0.238	-3.105	
APLEZIER	0.205	1.584e-01	0.153
AFIETS	0.049	2.031e-02	0.001
AINBOED	0.176	2.573e-02	

Table 5: Last Four Predictors for Logistc, Ridge, and Lasso Regression

# Of Predictors	Cylinders	Displacement	Horsepower	Weight	Acceleration	Year	Origin2	Origin3
1				*				
2				*		*		
3				*		*		*
4				*		*	*	*
5		*		*		*	*	*
6		*	*	*		*	*	*
7	*	*	*	*		*	*	*
8	*	*	*	*	*	*	*	*

Table 6: Best Subset Selection Visual

Term	LS	Best Subset	Ridge	Lasso	PCR	PLS
Intercept	22.415	23.446	22.656	22.442	23.446	23.446
Cylinders	-0.835		-0.650	-0.714	-1.370	-1.354
Displacement	2.509		-0.259	2.250	-1.416	-1.402
Horsepower	-0.670		-1.084	-0.654	-1.372	-1.356
Weight	-5.670	-5.633	-3.128	-5.632	-1.346	-1.449
Acceleration	0.218		-0.158	0.207	0.942	0.737
Year	2.862	2.790	2.443	2.852	0.668	1.011
Origin2	2.630		1.706	2.552	0.511	0.425
Origin3	2.853		2.454	2.784	0.661	0.786
Test Error	10.682	11.655	11.444	10.685	17.138	15.603

Table 7: Estimated coefficients and test error results, for different subset and shrinkage methods applied to the Auto data. The blank entries correspond to variables omitted