Project 4

Jose Alfaro

Due: March 27, 2019

Question: 1

Part: A

A scatterplot of GPA Vs. ACT Scores can be seen in *figure 1*. Here, we observe that there is a correlation of 0.2695 between the two variables. This implies that GPA and SAT Scores are not highly correlated. Moreover, we superimposed the regression line onto the scatterplot to emphasize the weak correlation between the two variables since the data points do not follow the linear pattern produced by the regression line.

Part: B

By using the bootstrap technique, we were able to take 10,000 bootstrap samples from our data (with replacement) and calculate all 10,000 correlation coefficients. From here, we calculated the point estimate for ρ (the population correlation between GPA and ACT Scores), the bootstrap estimates for bias and standard error of the point estimate, and the 95% confidence interval (computed using percentile bootstrap). These findings are demonstrated in *table 1*. Here we see that the point estimate was 0.2694818 which was extremely close to the sample correlation calculated in part a. We also saw that the bias and standard error were 0.00438085 and 0.1047863 respectively. This means that there is a very small amount of bias with good accuracy. Lastly, the 95% confidence interval produced for the point estimate was (0.0673, 0.4778). Since this confidence interval does not contain 0, we say it is significant.

Part: C

A simple linear regression model was fit which predicted GPA on the basis of ACT scores. From here, we were able to obtain the least squares estimates of the regression coefficients, the standard errors of the estimates, and a 95% confidence interval of the coefficients. Note, these values can be found on *table 2*. Furthermore, when checking our model assumptions, we found two major issues. The first issue can be seen in the residual plot of the data (*Figure 2*). Here, we see that although the residuals are somewhat centered around 0, there is a clear trend in the residuals. In other words, there is an issue of non-constant variance among the residuals. Furthermore, the second issue that we found was due to normality. When observing the QQPlot (*Figure 2*), we see that the data mainly closely follows the quantile plot in the middle of the quantiles, however, once there are deviations in the tails that hint that the data may have an issue with normality.

Part: D

We used the bootstrap technique to compute the standard errors and 95% confidence intervals for the regression coefficients. The point estimates and the standard errors produced through the bootstrap procedure can be found in *table 3*. Moreover, when comparing the regression coefficients produced by the bootstrap technique with those calculated by least squares method, we see that the coefficients are identical. However, there are a couple of significant differences. The first difference is that the standard errors are not the same. The standard error for the intercept produced by the non-parametric bootstrap approach was 0.3596, while the linear model produced a standard error of 0.3209. Moreover, the standard error for the slope (ACT variable) produced by the bootstrap approach was 0.0146, while the linear model produced a standard error of 0.0128.

It is important to note that both of the standard errors produced via the bootstrap method are slightly larger than those produced by the least squares method. The second noticeable difference between the two methods was that the 95% confidence intervals produced via the bootstrap technique were slightly larger than those produced via the least squares method.

Question: 2

Part: A

StoreID, STORE, and Store7 were examined to see if any of these variables are to be included in the model. To check this, we observed our data and realized that these variables were all identification predictors that correspond to various stores. Since identification predictors aren't informative pieces of information when attempting to predict orange juice sales, we decided that none of these variables should be included in the model. Moreover, boxplots for each of these variables can be seen in *figure 3* where one can observe that none of these predictors have a clear relationship with the purchase of orange juice (our response variable). However, due to the directions, we kept StoreID in our dataset as a potential categorical predictor and removed STORE and Store7.

Part: B

The full logistic regression model was:

$$log(\frac{P(y=1)}{1-P(y=1)}) = \beta_0+\beta_1*WeekofPurchase+\beta_2*StoreID2+\beta_3*StoreID3+\beta_4*StoreID4+\beta_5*StoreID7+$$

$$\beta_6*PriceCH+\beta_7*PriceMM+\beta_8*DiscCH+\beta_9*DiscMM+\beta_{10}*SpecialCH+\beta_{11}*SpecialMM+\beta_{12}*LoyalCH+$$

$$\beta_{13}*SalePriceMM+\beta_{14}*SalePriceCH+\beta_{15}*PriceDiff+\beta_{16}*PctDiscMM+\beta_{17}*PctDiscCH+\beta_{18}*ListPriceDiff$$

Alternatively, this logistic regression model could be writen with the coefficients as:

$$log(\frac{P(y=1)}{1-P(y=1)}) = 4.910-0.011*WeekofPurchase-0.047*StoreID2-0.213*StoreID3-0.532*StoreID4-$$

$$0.648*StoreID7+4.565*PriceCH-3.685*PriceMM+11.272*DiscCH+25.443*DiscMM+0.263*SpecialCH+$$

$$0.314*SpecialMM-6.249*LoyalCH+NA*SalePriceMM+NA*SalePriceCH+NA*PriceDiff-$$

$$48.557*PctDiscMM-28.261*PctDiscCH+NA*ListPriceDiff$$

It is important to note that when looking at the logistic regression model with the coefficients in place, there are four variables that have "NA" as their coefficients. This is because these variables are linearly related to one or more of the other variables already in the model. This led us to perform model selection procedures to determine which variables are useful and which can be excluded.
After performing forward, backward, and stepwise model selection procedures based on AIC values, we see that the backward and stepwise procedures concluded in the same model with the lowest AIC score of 840.6. A table of the model selecion procedure can be found in *table 4*. The simplified logistic regression model was:

$$log(\frac{P(y=1)}{1-P(y=1)}) = \beta_0+\beta_1*SpecialMM+\beta_2*PctDiscMM+\beta_3*DiscMM+\beta_4*PriceCH+\beta_5*PriceMM+$$

$$\beta_6*PctDiscCH+\beta_7*LoyalCH$$

Alternatively, the simplified logistic regression model could be written with its coefficients as:

$$log(\frac{P(y=1)}{1-P(y=1)}) = 4.680+0.390*SpecialMM-52.533*PctDiscMM+27.190*DiscMM+3.610*PriceCH-$$

$$4.157 * PriceMM - 8.039 * PctDiscCH - 6.419 * LoyalCH$$

Once we obtained a simplified logistic regression model, we computed the confusion matrix which can be seen in *table 5*. We saw that:

$$Sensitivity = \frac{317}{317 + 100} = \frac{317}{417} = 76.02\%$$

$$Specificity = \frac{581}{581 + 72} = \frac{581}{653} = 88.97\%$$

$$MR = \frac{72 + 100}{72 + 100 + 317 + 581} = \frac{172}{1070} = 16.07\%$$

We then plotted the ROC curve which can be found in *figure 4* and saw that the area under the curve was 0.9026. Which implies that the test is accurate. Lastly, we calculated an estimate of 0.1226 for the test error rate using 10-fold cross-validation (Note: A table of test error rates calculated using 10-Fold CV can be found in *table 9*).

Part: C

We repeated all of part B except this time we used all of the predictors using the LDA method. The confusion matrix can be found in *table 6*. From here we can see that:

$$Sensitivity = \frac{318}{99 + 318} = \frac{318}{417} = 76.26\%$$

$$Specificity = \frac{572}{81 + 572} = \frac{572}{653} = 87.60\%$$

$$MR = \frac{81 + 99}{81 + 99 + 318 + 572} = \frac{180}{1070} = 16.82\%$$

We then plotted the ROC curve which can be found in *figure 4* and saw that the area under the curve was 0.9017 which implies that LDA is fairly accurate. Lastly, we calculated an estimate of 0.1682 for the test error rate using 10-fold cross-validation.

Part: D

Once again, we repeated all of part B except this time we used all the predictors using the QDA method. The confusion matrix can be found in *table 7*. From here we can see that:

$$Sensitivity = \frac{325}{92 + 325} = \frac{325}{417} = 77.94\%$$

$$Specificity = \frac{540}{540 + 113} = \frac{540}{653} = 82.70\%$$

$$MR = \frac{92 + 113}{92 + 113 + 325 + 540} = \frac{205}{1070} = 19.16\%$$

We then plotted the ROC curve which can be found in *figure 4* and saw that the area under the curve was 0.8660 which implies that QDA is fairly accurate, but not as accurate as LDA. Lastly, we calculated an estimate of 0.1972 for the test error rate using 10-fold cross-validation.

Part: E

We repeated all of part B, but this time we used all of the predictors using the KNN method. To implement this method, we used 10-fold cross-validation to find our optimal k value which ended up being k = 12. From here, we were able to run the KNN algorithm using K = 12 and computed he confusion matrix, which can be found in *table 8*. From here we can see that:

$$Sensitivity = \frac{319}{98+319} = \frac{319}{417} = 76.50\%$$

$$Specificity = \frac{565}{88+565} = \frac{565}{653} = 86.52\%$$

$$MR = \frac{88+98}{88+98+319+565} = \frac{186}{1070} = 17.38\%$$

We then plotted the ROC curve which can be found in *figure 4* and saw that the area under the curve was 0.8163 which implies that QDA is fairly accurate, but not as accurate as LDA. Lastly, we calculated an estimate of 0.1695 for the test error rate.

Part: F

Based on the results found in parts A-E, I would recommend the logistic regression classifier above all of the other ones tested. This is simply because the area under the curve for the ROC curve in the logistic regression method was 0.9026, which was the highest among all of the other methods studied. A comparison of AUC values between the different methods can be seen in *figure 4*. Furthermore, when observing the test error rates for all of the methods, we see that the logistic regression classifier has the lowest error rate of 0.1226. A comparison of test error rates among all methods studied can be seen in *table 9*. Lastly, one can see that the logistic regression classifier has the lowest misclassification rate among all the methods studied. A comparison of misclassification rates between all methods studied can be seen in *table 10*

Question: 3

Part: A

We performed an exploratory data analysis on the Auto dataset and found that displacement and cylinders are highly positively correlated. We also saw that weight is also highly correlated with cylinders and displacement. This immediately hints that we may not need all of the variables in the model since a lot of them are highly correlated and probably explain a lot of the same information. (Note: A scatterplot matrix containing correlation values as well as a density plot can be seen in *figure 5*).

Part: B

After conducting our exploratory data analysis, we fit a multiple linear regression model using the least squares method. For this model, we used mpg as our response variable and all other variables as predictors (except name). The resulting multiple linear regression model was:

$$mpg = \beta_0 + \beta_1 * cylinders + \beta_2 * displacement + \beta_3 * horsepower + \beta_4 * weight + \beta_5 * acceleration + \beta_6 * year + \beta_7 * origin$$

Alternatively, we can substitute the regression coefficients in the model and rewrite it as:

$$mpg = -17.218 - 0.493 * cylinders + 0.020 * displacement - 0.017 * horsepower - 0.006 * weight +$$

$$0.080576 * acceleration + 0.751 * year + 1.426 * origin$$

This model resulted in an R^2 value of 0.8215 and an R^2_{adj} value of 0.8182.

Part: C

From our exploratory data analysis, we knew that we had to drop some of the variables in the model since a lot of them were strongly correlated. Therefore, we preformed best-subset selection to find the best model. After running best subset selection, we saw that different selection criteria chose different number of variables. This can be seen in *figure 6* where subsets of variables are chosen by RSS, R^2_{adj}, C_p, and BIC. The most important thing to note in *figure 6*, 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 3 predictors. This means that the most parsimonious model is one with only three 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 an extra variable (or more) 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 11*. From here, we saw that since the optimal number of predictors to include in the model was 3, then by using best subset selection the model chosen is:

$$MPG = \beta_0 + \beta_1 * Weight + \beta_2 * Year + \beta_3 * Origin$$

. Lastly, we tested to see if the interactions between these three predictors were significant by fitting a "full" model which contained all main effects as well as all possible interactions and fitting a "reduced" model which contained only the main effects. Then, we conducted an F-Test to see if we could drop the interaction terms since these two models were nested. As a result, we kept the reduced model and concluded that that the interactions were not significant. Thus, $MPG = \beta_0 + \beta_1 * Weight + \beta_2 * Year + \beta_3 * Origin$ is the final and has an R^2 value of 0.8175 and an R^2_{adj} value of 0.8160.

Part: D

After running forward selection, we saw that all of the model selection criteria (RSS, R^2_{adj}, C_p, and BIC) had an elbow at around 3 predictors. This can be seen in *figure 7*. This was the same conclusion as in part C. Moreover, the forward selection procedure also produced 7 "optimal" models which corresponded to the size of the model. Interestingly, the optimal models were exactly the same as in part C and can be found in *table 12*. This means that the same analysis took place as in part C in which the interactions were not significant and the final model chosen by forward selection was $MPG = \beta_0 + \beta_1 * Weight + \beta_2 * Year + \beta_3 * Origin$. This model resulted in an R^2 value of 0.8175 and an R^2_{adj} value of 0.8160.

Part: E

After running backward selection, we saw that all of the model selection criteria (RSS, R^2_{adj}, C_p, and BIC) had an elbow at around 3 predictors. This can be ssen in *figure 8*. This was the same case as in parts C and D. Moreover, the forward selection procedure also produced 7 "optimal" models which corresponded to the size of the model. Interestingly, the optimal models were exactly the same as in parts C and D and are shown in *table 13*. This means that the same analysis took place as in parts C and D in which the interactions were not significant and the final model chosen by forward selection was $MPG = \beta_0 + \beta_1 * Weight + \beta_2 * Year + \beta_3 * Origin$. This model resulted in an R^2 value of 0.8175 and an R^2_{adj} value of 0.8160. This can further be examined in *figure 9* where we compare the three selection procedures from parts C, D, and E to ensrue that three variables is the optimal model size choice based on the elbow.

Part: F

Since the multiple linear regression model in part B contained all of the variables, it produced a high R^2_{adj} value of 0.8182. On the other hand, parts C, D, and E all produced the same model, which only contained 3 predictors, and produced an R^2_{adj} value of 0.8160. Furthermore, since parts C, D, and E all produced the

same model, then they all shared the same R^2 value of 0.8175 so there is no need to pick a model of the same size since all three of the models are identical. Although the multiple linear regression model produced a slightly higher R^2_{adj} value, we would prefer the model chosen by parts C, D, and E since it still has a high R^2_{adj} that is very similar to the full model.

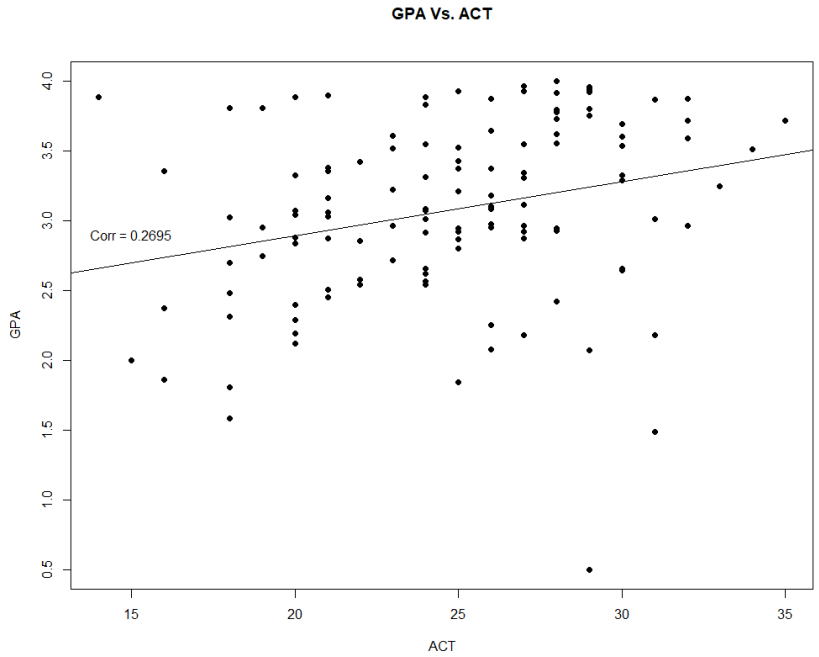

Figure 1: Scatterplot of GPA Vs. ACT Scores

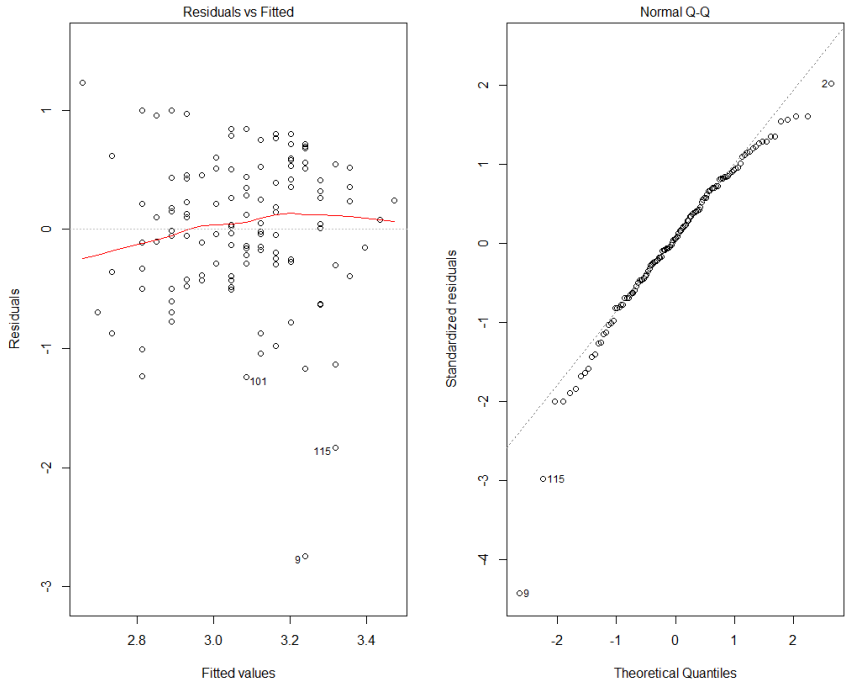

Figure 2: Model Diagnostics for Linear Model GPA ACT

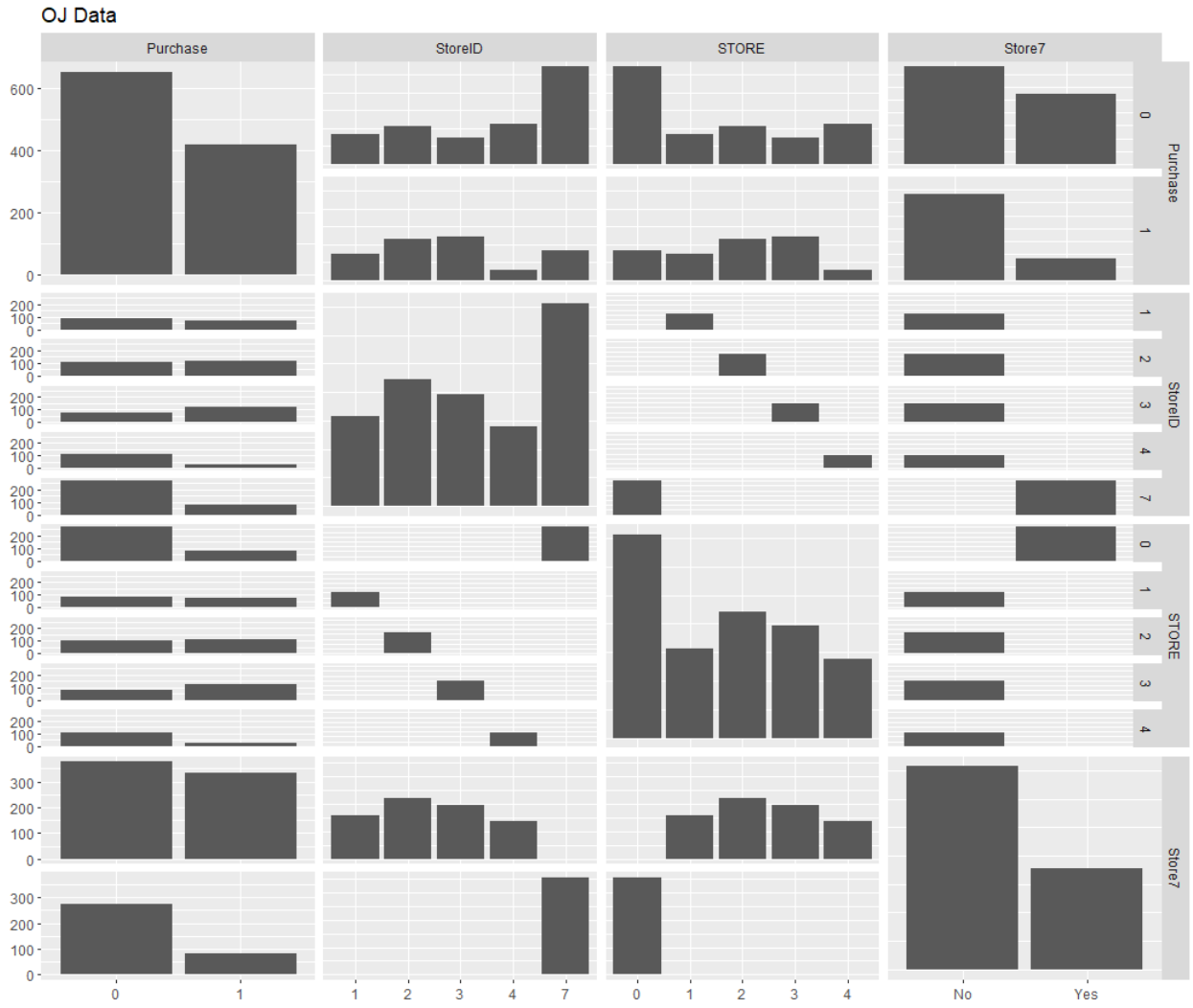

Figure 3: Barplots for StoreID, STORE, and Store7

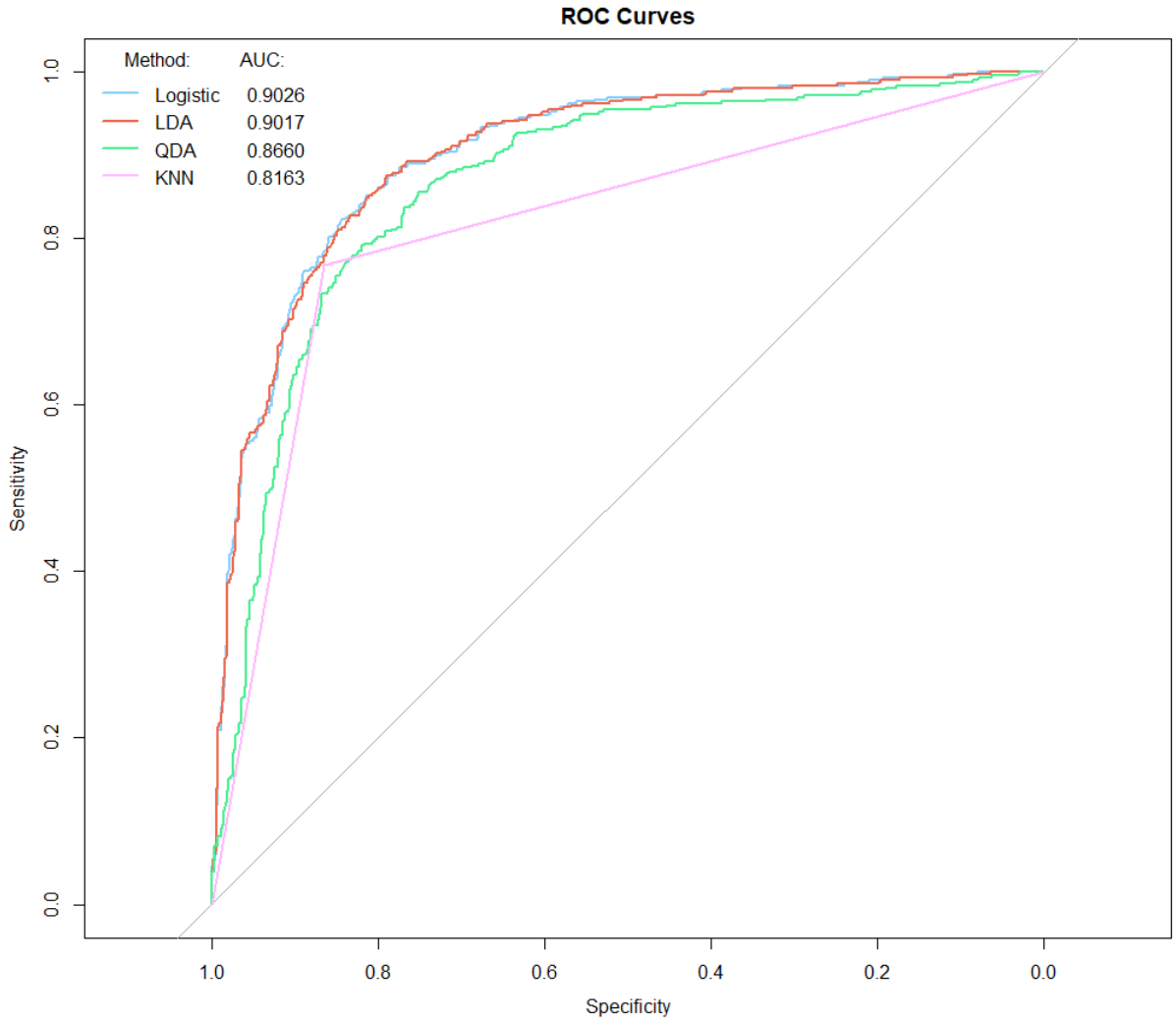

Figure 4: ROC Curves for all Methods

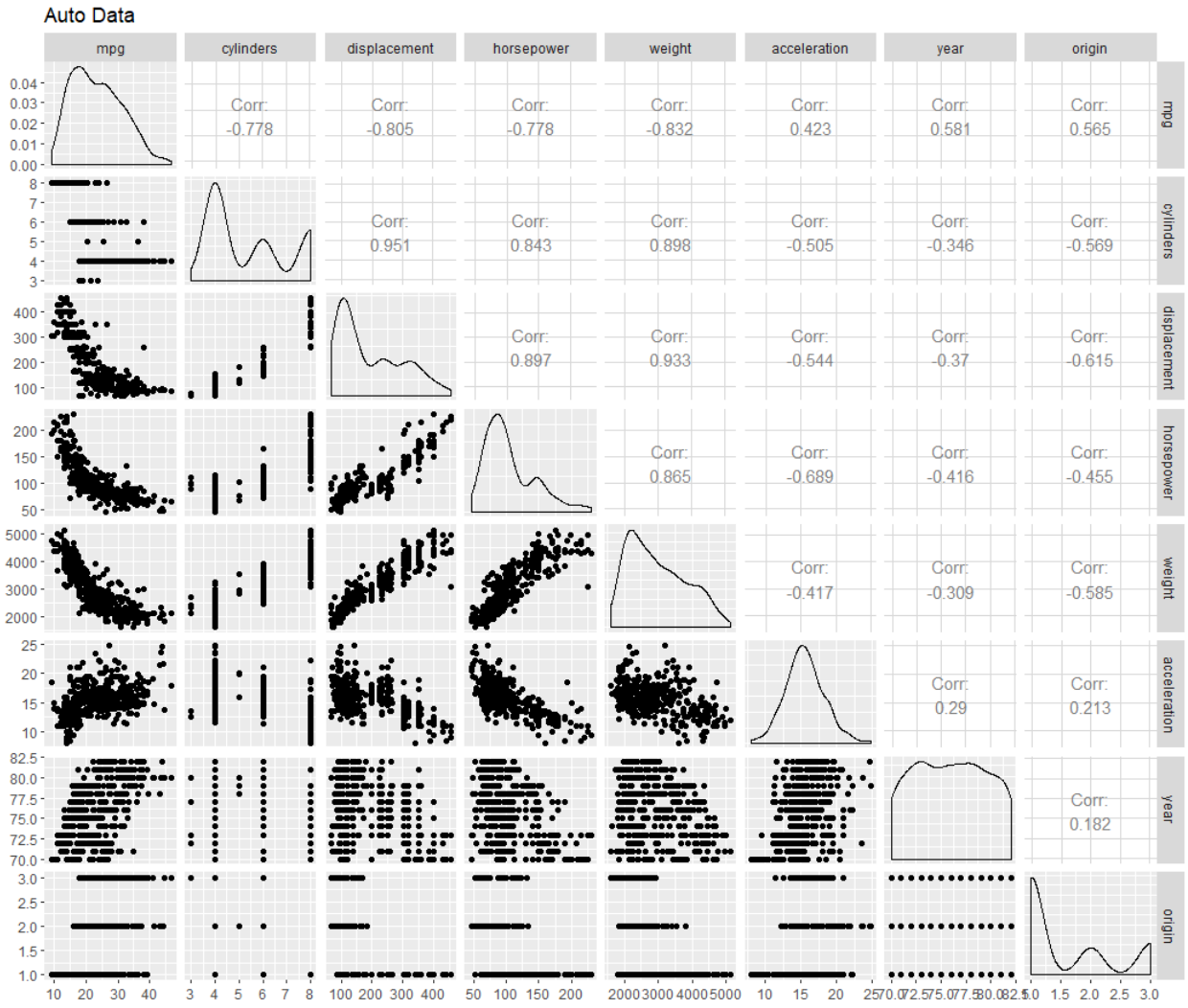

Figure 5: Scatterplot Matrix for Auto Dataset

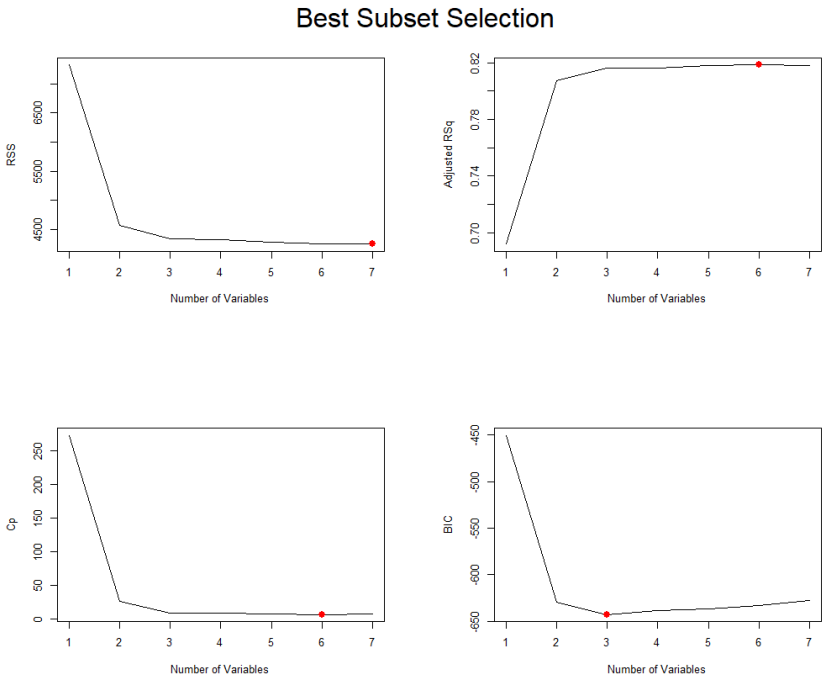

Figure 6: Best Subset Selection with Different Selection Criteria

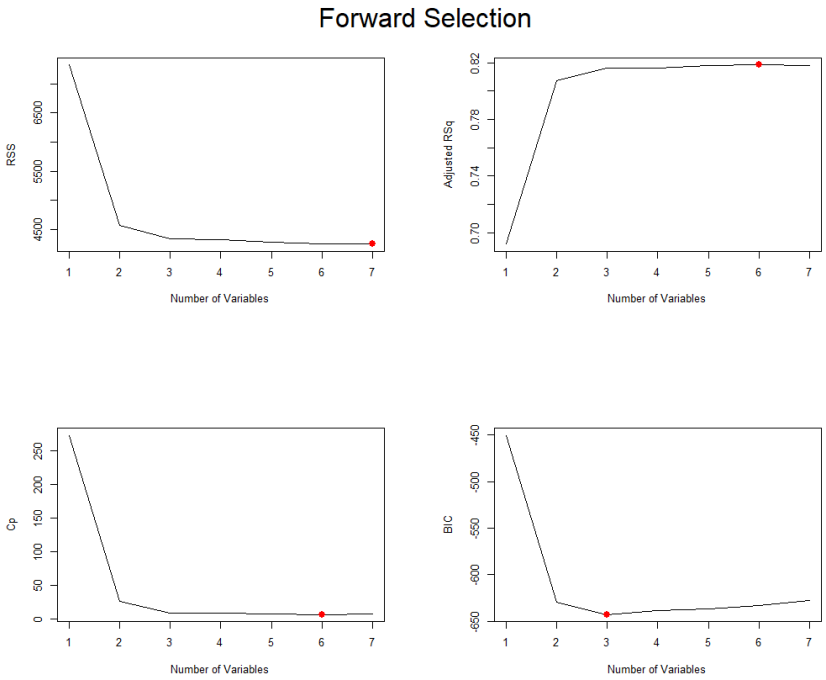

Figure 7: Forward Selection with Different Selection Criteria

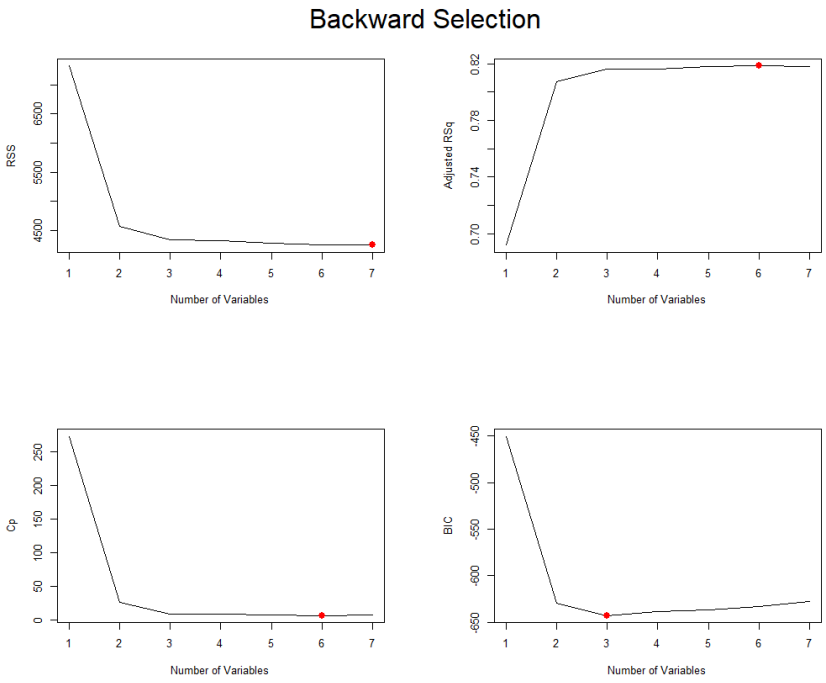

Figure 8: Backward Selection with Different Selection Criteria

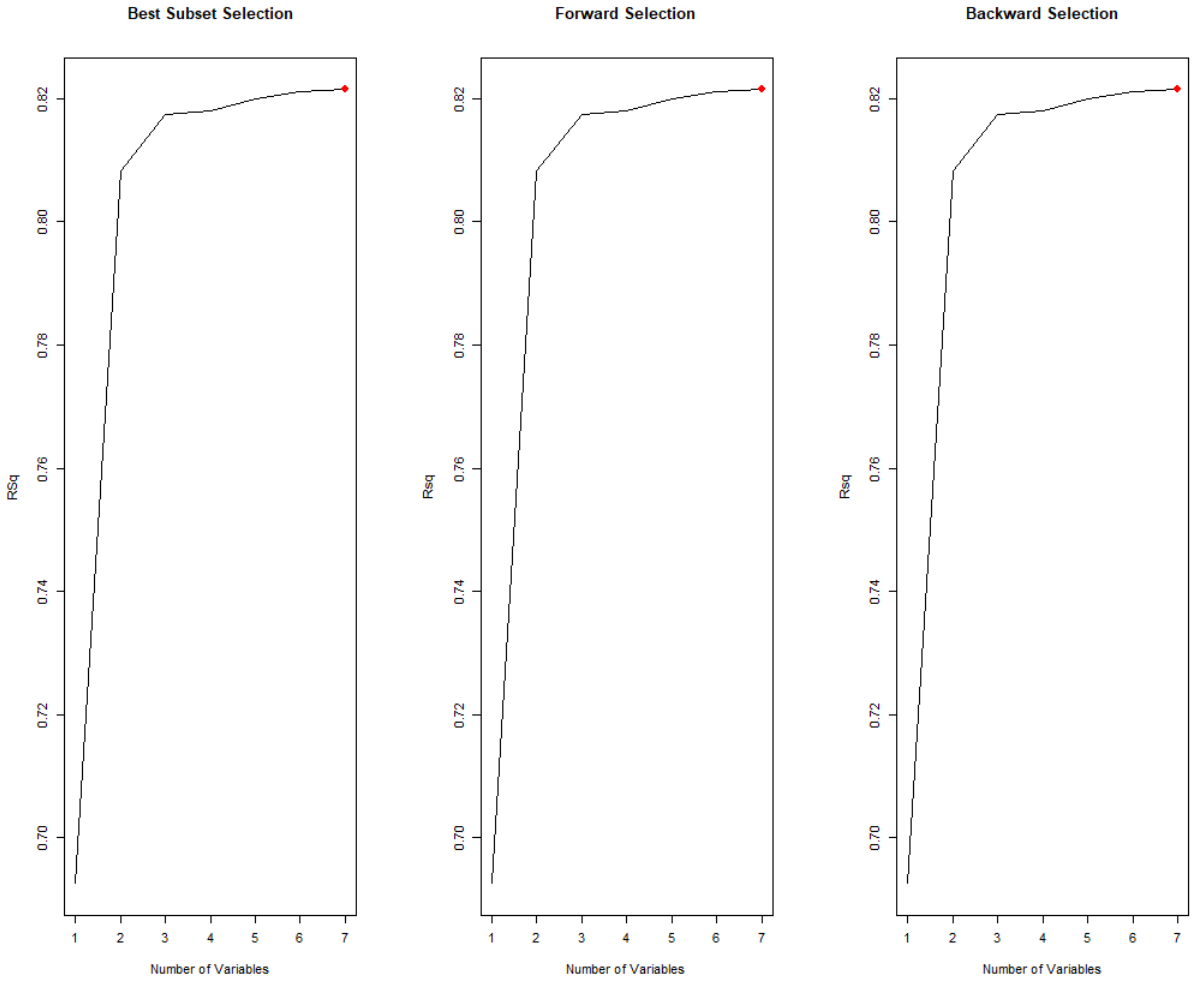

Figure 9: R2 Comparison for Best Subset, Forward, and Backward Selection

Tables for Report

Jose Alfaro

March 24, 2019

Bootstrap Statistic (Correlation)	Bias	Standard Error	95% Confidence Interval (Correlation)
0.2694818	0.004846546	0.1049546	(0.0673, 0.4778)

Table 1: Boostrap Correlation Statistic, Bias, Standard Error, and 95% Percentile Confidence Interval

	Estimate	Standard Error	t-Value	P-Val	95% Confidence Intervals
Intercept	2.11405	0.32089	6.588	1.3e-09	(1.4786, 2.750)
ACT	0.03883	0.01277	3.040	0.00292	(0.0135, 0.0641)

Table 2: Linear Model Output (Regression Coefficients)

	Bootstrap Statistics (Reg Coeff)	Bias	Standard Error	95% Confidence Interval
Intercept	2.11404929	-0.0108361368	0.35962727	(1.387, 2.812)
ACT	0.03882713	0.0004350197	0.01455027	(0.0110, 0.0674)

Table 3: Boostrap Output (Regression Coefficients)

Direction	Variables	AIC
Forward	WeekofPurchase, StoreID, PriceCH, PriceMM, SpecialCH, SpecialMM, LoyalCH, SalePriceMM, SalePriceCH, PriceDiff, PctDiscMM, PctDiscCH, ListPriceDiff	846.2
Backward	PriceCH, PriceMM, DiscMM, SpecialMM, LoyalCH, PctDiscMM, PctDiscCH	840.6
Stepwise	PriceCH, PriceMM, DiscMM, SpecialMM, LoyalCH, PctDiscMM, PctDiscCH	840.6

Table 4: Logistic Regression Model Selection

		Actual	
		CH	MM
Predicted	CH	581	100
	MM	72	317

Table 5: Logistic Regression Confusion Matrix

		Actual	
		CH	MM
Predicted	CH	572	99
	MM	81	318

Table 6: LDA Confusion Matrix

		Actual	
		CH	MM
Predicted	CH	540	92
	MM	113	325

Table 7: QDA Confusion Matrix

		Actual	
		CH	MM
Predicted	CH	565	98
	MM	88	319

Table 8: KNN Confusion Matrix

Method	Test Error Rate
Logistic Regression	0.1226
LDA	0.1682
QDA	0.1972
KNN	0.1695

Table 9: Test Error Rates for all Methods

Method	Misclassification Rate (Percent)
Logistic Regression	16.07
LDA	16.82
QDA	19.16
KNN	17.38

Table 10: Misclassification Rate for all Methods

Model Size	Cylinders	Displacement	Horsepower	Weight	Acceleration	Year	Origin
1				*			
2				*		*	
3				*		*	*
4		*		*		*	*
5		*	*	*		*	*
6	*	*	*	*		*	*
7	*	*	*	*	*	*	*

Table 11: Best Subset Selection Visual

Model Size	Cylinders	Displacement	Horsepower	Weight	Acceleration	Year	Origin
1				*			
2				*		*	
3				*		*	*
4		*		*		*	*
5		*	*	*		*	*
6	*	*	*	*		*	*
7	*	*	*	*	*	*	*

Table 12: Forward Selection Visual

Model Size	Cylinders	Displacement	Horsepower	Weight	Acceleration	Year	Origin
1				*			
2				*		*	
3				*		*	*
4		*		*		*	*
5		*	*	*		*	*
6	*	*	*	*		*	*
7	*	*	*	*	*	*	*

Table 13: Backward Selection Visual