Andrea Bruckner
Predict 422, Sec 56

Individual Project #1

Introduction

In this report I am using the diabetes data set from Efron et al. (2003) to predict the effects that ten predictor variables, including age, sex, BMI, average blood pressure, and various blood measurements, have on the progression of diabetes one year after baseline. The data set contains 442 observations, each representing a unique patient. I am using various machine learning modeling techniques, including least squares regression and best subset selection, as well as 10-fold cross validation in conjunction with best subset selection, ridge regression, and lasso techniques. Prior to creating any models, I partitioned the data so that I could train each model on approximately 75 percent of the data and test it on the remaining 25 percent.

Models and Analysis

Model 1: Least squares regression model using all ten predictors

Coefficient Estimates

Variable	Estimate	Std. Error	t value	Pr(>|t|)	Signif.
(Intercept)	149.920	2.976	50.382	< 2e-16	***
age	-66.758	68.946	-0.968	0.33364	
sex	-304.651	69.847	-4.362	1.74E-05	***
bmi	518.663	76.573	6.773	6.01E-11	***
map	388.111	72.755	5.335	1.81E-07	***
tc	-815.268	537.549	-1.517	0.13034	
ldl	387.604	439.162	0.883	0.37811	
hdl	162.903	269.117	0.605	0.54539	
tch	323.832	186.803	1.734	0.08396	.
ltg	673.62	206.888	3.256	0.00125	**
glu	94.219	79.59	1.184	0.23737	
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1					

Test Errors

Test MSE	Standard Error of Test MSE
3111.265	361.0908

In the coefficient estimates table above, we can see that only five of the ten predictors are statistically significant: sex, bmi, map, tch, and ltg. If I were to explore this data further, I might want to try simplifying the least squares regression model by fitting only the significant variables.

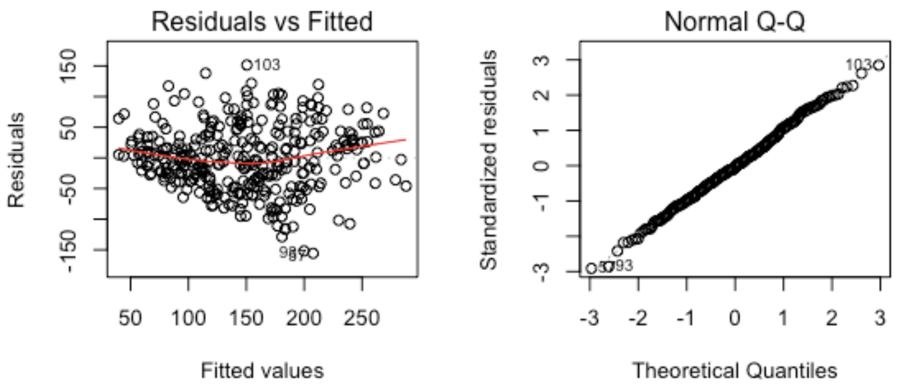

Figure 1: Diagnostic Plots

The Residuals vs Fitted plot in **Figure 1** reveals that the residuals for Model 1 appear randomly scattered about a mean of zero, and the Normal Q-Q plot shows that the residuals fall nearly in a straight line. Both plots suggest that the data do not violate the assumptions of normality.

Model 2: Best subset selection model using BIC to select the number of predictors

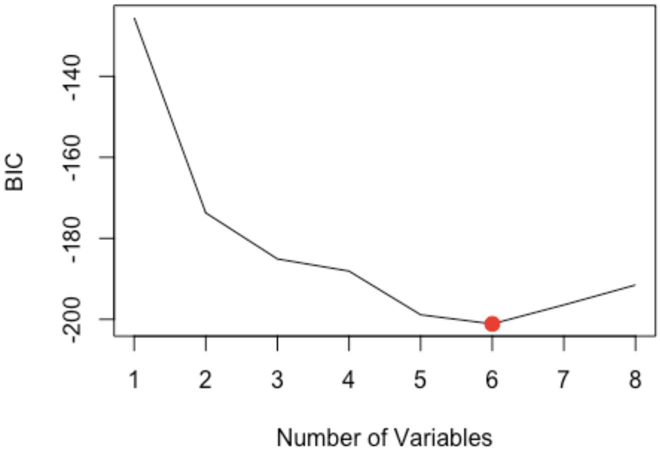

Figure 2: Best Subset Selection Plot

Coefficient Estimates

Variable	Estimate
(Intercept)	150.1166
sex	-306.042
bmi	538.8274
map	389.0673
tc	-379.0379
tch	332.6735
ltg	527.5658

Test Errors

Test MSE	Standard Error of Test MSE
3095.483	369.7526

In order to determine the best subset selection model, I looked for the model with the smallest BIC value. **Figure 2** shows that the model with six variables has the smallest BIC value.

Model 3: Best subset selection model using 10-fold cross-validation to select the predictors

For Model 3, I used 10-fold cross-validation to find the best subset selection model with the lowest mean cross-validation error. **Figure 3** shows that the model with six variables has the lowest mean cross-validation error.

As the tables show, the coefficient estimates, test MSE, and standard error of the test MSE are identical to the values in Model 2.

Coefficient Estimates

Variable	Estimate
(Intercept)	150.1166
sex	-306.042
bmi	538.8274
map	389.0673
tc	-379.0379
tch	332.6735
ltg	527.5658

Test Errors

Test MSE	Standard Error of Test MSE
3095.483	369.7526

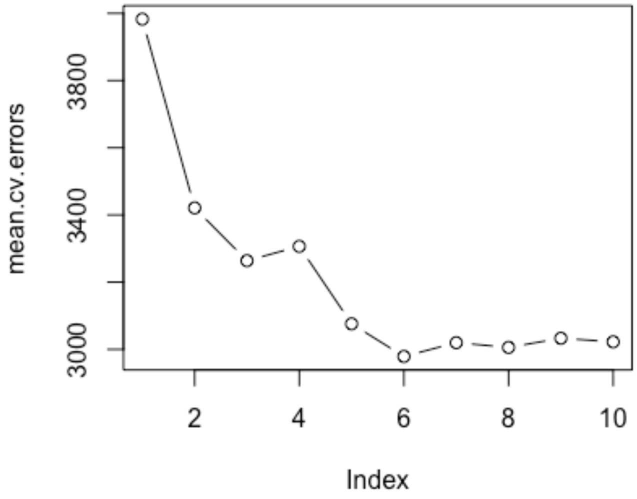

Figure 3: Best Subset Selection Using Cross-Validation Plot

Model 4: Ridge regression model using 10-fold cross-validation to select the largest value of λ such that the cross-validation error is within 1 standard error of the minimum

Coefficient Estimates

Variable	Estimate
(Intercept)	149.99068
age	-11.33162
sex	-156.91053
bmi	374.44939
map	264.89998
tc	-31.96990
ldl	-66.89724
hdl	-174.01202
tch	123.97204
ltg	307.68646
glu	134.48120

Test Errors

Test MSE	Standard Error of Test MSE
3070.870	350.5467

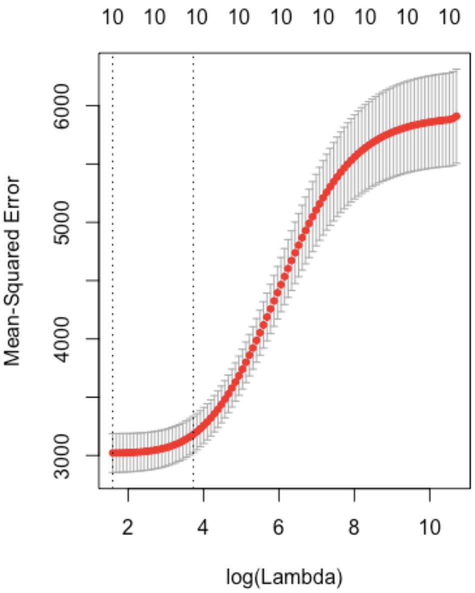

Figure 4: Cross-Validation Plot

In **Figure 4**, we can see the MSE is smallest when log(Lambda) is smallest. The first vertical line from the left of the plot signifies the smallest MSE, while the second vertical line represents one standard deviation from the minimum MSE. The number 10 at the top of the plot indicates that all ten predictor variables are present in the model regardless of the lambda value. I used lambda = 41.67209 to estimate the coefficients and calculate the test errors since this was the largest value of lambda where the error was still within one standard error of the minimum MSE.

Model 5: Lasso model using 10-fold cross-validation to select the largest value of λ such that the cross-validation error is within 1 standard error of the minimum

Coefficient Estimates

Variable	Estimate
(Intercept)	149.95298
age	.
sex	-119.62207
bmi	501.56473
map	270.92613
tc	.
ldl	.
hdl	-180.29436
tch	.
ltg	390.55001
glu	16.58881

Test Errors

Test MSE	Standard Error of Test MSE
2920.041	346.2248

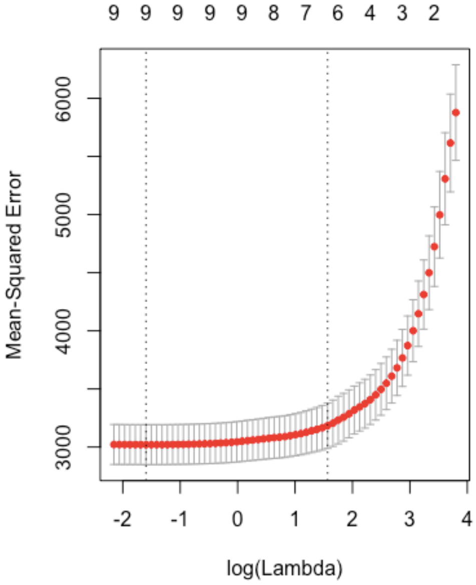

Figure 5: Cross-Validation Plot

Looking at the second vertical line from the left in **Figure 5**, we can see that the number of predictors in the model is six, as denoted by the number at the top of the plot. Although Models 2 and 3 also use only six predictors, the predictors for Model 5 are slightly different. All three models use the sex, bmi, map, and ltg predictors, but Model 5 uses hdl and glu instead of tc and tch as used in the other two models. To estimate the coefficients and calculate the test errors, I used lambda = 4.791278 since this was the largest value of lambda where the error was still within one standard error of the minimum MSE.

Results

Model	Test MSE	Standard Error of Test MSE
1	3111.265	361.0908
2	3095.483	369.7526
3	3095.483	369.7526
4	3070.870	350.5467
5	2920.041	346.2248

Model 1 (least squares regression containing all ten predictors) resulted in the worst test MSE, although its test MSE standard error was slightly better than the test MSE standard error for Models 2 and 3. As mentioned earlier, Model 2 (best subset selection using BIC) and Model 3 (best subset selection using 10-fold cross-validation) resulted in the same six-variable model with identical coefficient estimates and test errors. Model 4 (ridge regression using 10-fold cross-validation) performed slightly better than Models 1, 2, and 3. It, however, is a more complex model than Models 2 and 3 because it uses all ten predictors rather than just six. Model 5 (lasso using 10-fold cross-validation) resulted in the smallest test MSE and test MSE standard error of all the models. It contains only six predictors. Because of its small test errors and relative simplicity, Model 5 can be considered the best model.

Conclusion

After fitting various machine learning modeling techniques, including least squares regression, best subset selection using BIC, best subset selection using 10-fold cross validation, ridge regression using 10-fold cross validation, and lasso using 10-fold cross validation, I discovered that the lasso model using 10-fold cross validation performed best in predicting the progression of diabetes one year after baseline. This model contained the predictor variables sex, bmi, map, hdl, ltg, and glu. Before deploying the model, I recommend further model testing and consulting a diabetes expert to determine if the predictors included in the model make medical sense. If the model performs well in additional tests and receives the approval of a medical expert, it should be safe to deploy the model on new data.