Daren Purnell Jr.
Predict 422_Sec 55
Project 1

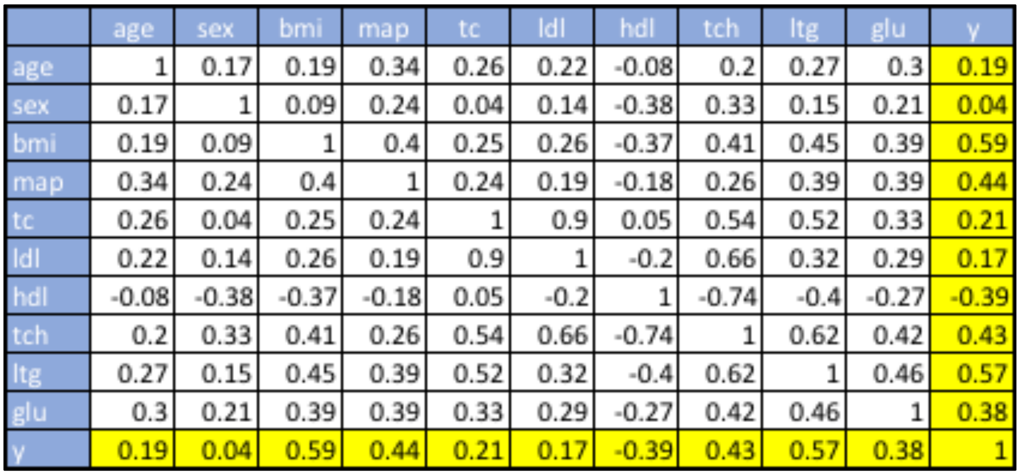
Introduction

The purpose of this project is to use diabetes data from the “lars” library to examine the effect of ten human biological predictor variables upon a quantitative response variable that is the measure of disease progression after one year. During this project, we will examine the data using exploratory data analysis, employ machine learning techniques to fit linear regression, ridge regression, and lasso models, and incorporate best subset selection and cross-validation. We will then present our results and draw conclusions on our analysis and the applicability of the various machine learning techniques upon the diabetes data.

Exploratory Data Analysis

The diabetes data consists of the response variable (y) and ten biological predictor variables age, sex, body mass index (bmi), average blood pressure (map), and six blood serum measurements (tc, ldl, hdl, tch, ltg, glu). There are 442 diabetes patients in this data set with no missing records. The response variable (y) and eight of the predictor variables appear to be continuous. The remaining two predictor variables, sex and tch, appear to be categorical. A data correlation matrix and scatterplot of the predictors with the response variable (y) appear to show a strong positive linear relationship between the progression of the disease and body mass index (bmi), average blood pressure (map), and the blood serum measurements tch, ltg, and glu. The Pearson correlation coefficients associated with these predictors suggest that as bmi, map, tch, ltg, and glu increase so does the progression of the disease. Conversely, the blood serum measurement hdl appears to show a strong negative linear relationship with the response variable that suggest that as hdl increases the progression of the disease decreases.

Diabetes Data Correlation Matrix



	age	sex	bmi	map	tc	ldl	hdl	tch	ltg	glu	y
age	1	0.17	0.19	0.34	0.26	0.22	-0.08	0.2	0.27	0.3	0.19
sex	0.17	1	0.09	0.24	0.04	0.14	-0.38	0.33	0.15	0.21	0.04
bmi	0.19	0.09	1	0.4	0.25	0.26	-0.37	0.41	0.45	0.39	0.59
map	0.34	0.24	0.4	1	0.24	0.19	-0.18	0.26	0.39	0.39	0.44
tc	0.26	0.04	0.25	0.24	1	0.9	0.05	0.54	0.52	0.33	0.21
ldl	0.22	0.14	0.26	0.19	0.9	1	-0.2	0.66	0.32	0.29	0.17
hdl	-0.08	-0.38	-0.37	-0.18	0.05	-0.2	1	-0.74	-0.4	-0.27	-0.39
tch	0.2	0.33	0.41	0.26	0.54	0.66	-0.74	1	0.62	0.42	0.43
ltg	0.27	0.15	0.45	0.39	0.52	0.32	-0.4	0.62	1	0.46	0.57
glu	0.3	0.21	0.39	0.39	0.33	0.29	-0.27	0.42	0.46	1	0.38
y	0.19	0.04	0.59	0.44	0.21	0.17	-0.39	0.43	0.57	0.38	1

Models

In the modeling phase, we employed several machine learning techniques to fit models using linear regression, ridge regression, and lasso while incorporating best subset selection and cross-validation techniques. Each model was initially fit on training data that consisted of 332 patient records to estimate the coefficients, then tested for fit against 110 separate records to calculate Mean Standard Error (MSE) and Standard Error (SE) as measures of fit. Implementing best subset selection using BIC and 10-fold cross-validation both yielded the same six variables as the subset of choice.

Summary of Coefficient Values & Goodness of Fit Metrics

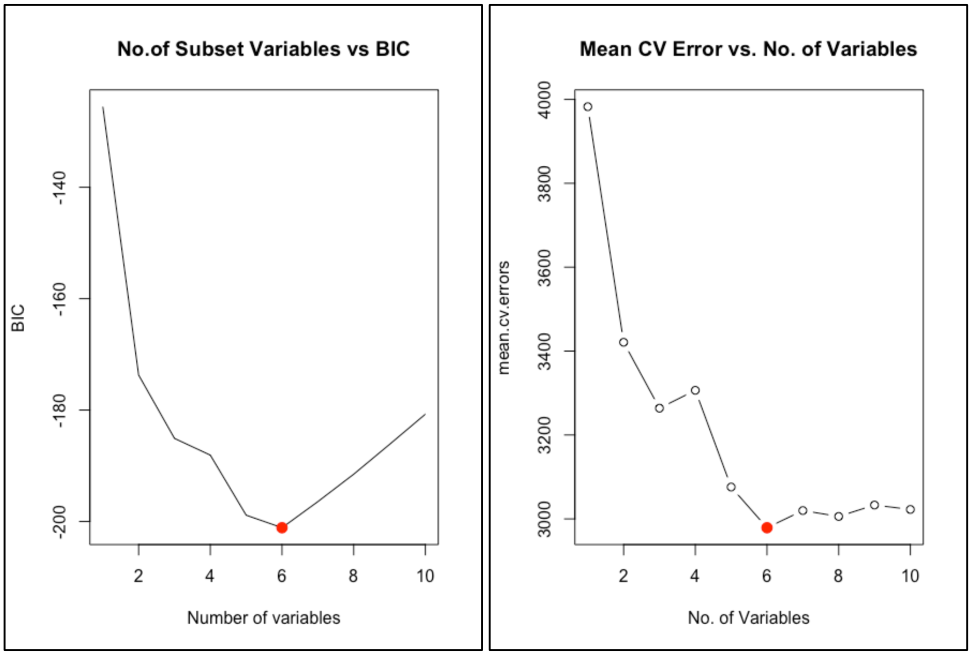
		Model Types				
Predictor	Coefficient	Least squares regression model using all ten predictors	Best subset selection using BIC to select the number of predictors (6 variables)	Best subset selection using 10-fold cross-validation to select the number of predictors (6 variables)	Ridge regression modeling using 10-fold cross-validation to select the largest value of λsuch that the cross-validation error is within 1 SE of the minimum	Lasso model using 10-fold cross-validation select the largest value of λsuch that the cross-validation error is within 1 SE of the minimum
Intercept	B0	149.92029	150.1166	150.7686	152.13348	149.95298
Age	B1	-66.75836	-306.042	-318.1613	21.26351	0
Sex	B2	-304.65071	538.8274	487.0493	-126.07041	-119.66533
BMI	B3	518.66346	389.0673	431.9678	375.81014	501.47816
MAP	B4	388.1113	-379.0379	-375.9508	240.56639	270.93815
TC	B5	-815.26815	332.6735	343.4045	-12.81517	0
LDL	B6	387.60431	527.5658	491.7793	-55.48969	0
HDL	B7	162.90253			-172.6269	-180.33591
TCH	B8	323.83151			121.76859	0
LTG	B9	673.62035			321.71154	390.53399
GLU	B10	94.21867			111.45002	16.62376
SE		361.1	369.75	372.17	339.57	346.228
MSE		3111.265	3095.48	3136.85	2947.51	2920.08

Analysis

We developed five different models that fitted our diabetes test data with varying levels of accuracy. Our initial ten-predictor least squares regression model fared the worst against the test set with a SE of 361.1 & MSE of 3111.265 P-values > 0.05 for several of the predictors (Age, Tc, LDL, HDL, GLU) indicated that they were not significant and contributed to the test set MSE. The Predictors with insignificant P-values also had signs that were not consistent with the implied relationship of their associated Pearson Correlation Coefficients (R-value). For example, the Predictor Age has a positive R-value of 0.19 with the response variable "y" but its coefficient in the least squares regression model is negative. I believe this is because the relationship between the response variable and predictor is not strongly linear.

Implementing best subset selection using BIC and 10-fold cross-validation yielded similar results. Both models selected six variables as the subset resulting in a SE of 369.75 & MSE of 372.17 for BIC and a SE of 372.17 & MSE of 3136.85 for 10-fold cross validation.

Comparison of BIC and 10-Fold Cross-Validation Subset Selection



Ridge regression using 10-fold cross-validation to select the largest value of λ such that cross-validation error is within 1 standard error of the minimum, yielded a λ-value of 41.67209 with a SE of 339.57 & MSE 2947.51. We observed that ridge regression with a good choice of λ significantly out performs our least squares regression model despite having the same number of predictors. It is interesting to note that ridge regression, except for GLU, attempted to shrink the coefficient values of the least squares predictors with insignificant P-values.

Lasso modeling using 10-fold cross-validation to select the largest value of λ such that cross-validation error is within 1 standard error of the minimum, yielded a λ-value of 4.791278 with a SE of 346.228 & MSE 2920.08. There was a small improvement in MSE that corresponded with a small increase in SE. It is interesting to note that that the predictors with high P-values in the least squares regression model coincided with shrunken ridge regression coefficients and zeroed-out lasso model coefficients. The lasso model with λ-value of 4.791278 was the highest performing model in our analysis.

Conclusion

During this project, we used diabetes data from the "lars" library to examine the effect of ten human biological predictor variables upon a quantitative response variable that is the measure of disease progression after one year. Employing the machine learning techniques of least squares, ridge, and lasso regression with best subset selection and cross-validation enabled us to identify our lasso model with λ-value of 4.791278 as the best performing. If we had more time to perform our analysis we would attempt to implement forward and backward stepwise variable selection to possibly achieve a better MSE.