Project 1 - Modeling Diabetes

Jacob Walsh

October 2, 2016

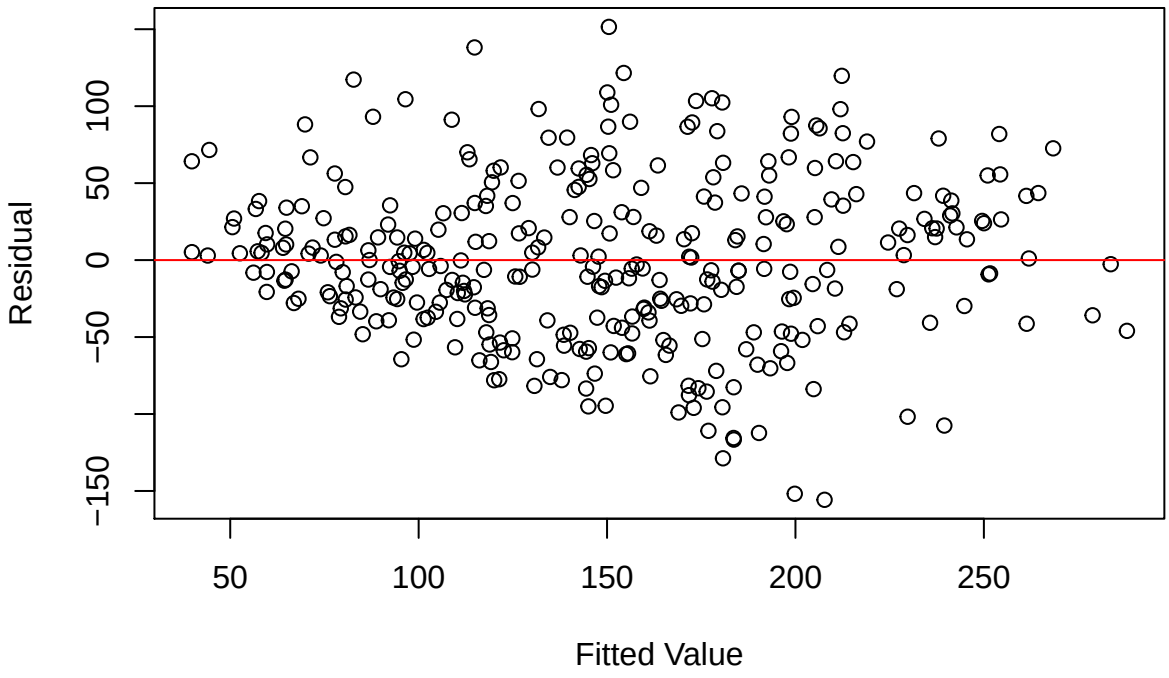
Introduction

In a diabetes study referenced in the report, "Least Angle Regression", by Efron et. all in 2004, data was gathered on a response of interest, the progression of diabetes one year after baseline tests. The study included 442 diabetes patients. Baseline prediction variables included age, sex, body mass index, blood pressure and six serum measurements. In this report, models using ordinary least squares regression, best subsets using BIC and 10 fold cross validation, ridge regression, and lasso regression will be built. The data will be split and models will be built based on the training data set and then compared. The mean squared error for the test data will be the basis for comparison of the models. The model with the lowest MSE will be chosen as the "best" model for the prediction of diabetes progression.

Analysis

The Linear Model

In order to make interpretations on a linear model, we first must assume the reasonability of applying a linear model. A residual plot is made to verify the linearity and equal variance of errors conditions.



After inspection of the residuals there doesn't appear to be any violations of the assumptions of linearity and

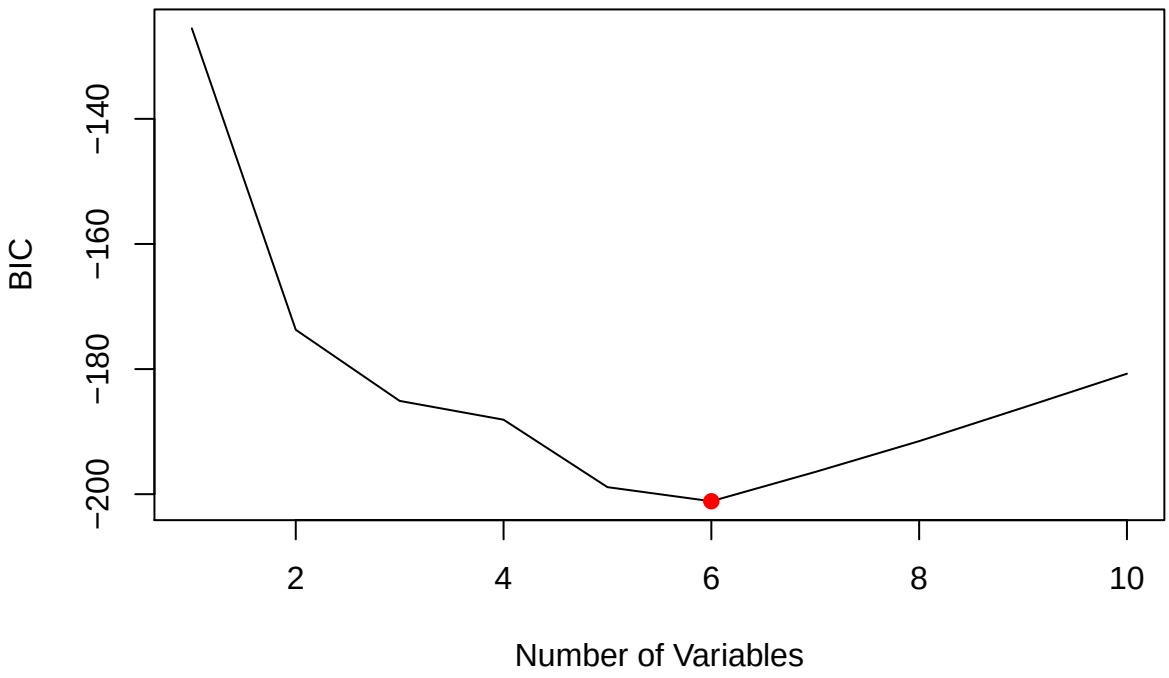
equal variance among the errors. A regression model is then built using the 332 observations in the training data, including all 10 predictors. The coefficients are as follows:

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   149.92       2.98   50.38     0.00
## age           -66.76      68.95   -0.97     0.33
## sex          -304.65      69.85   -4.36     0.00
## bmi           518.66      76.57    6.77     0.00
## map           388.11      72.75    5.33     0.00
## tc           -815.27     537.55   -1.52     0.13
## ldl           387.60     439.16    0.88     0.38
## hdl           162.90     269.12    0.61     0.55
## tch           323.83     186.80    1.73     0.08
## ltg           673.62     206.89    3.26     0.00
## glu            94.22      79.59    1.18     0.24
```

The test MSE for the model is 3111.265. Upon further inspection of the significance of the predictors in the model, it can be seen that there are several predictors that yield large p-values. Age, and the blood serum measurements of ldl, hdl, and glu all have very large p-values. If we were to use forward, backward, or subset selection methods, these predictors may not be included in the best model. Since there are several predictors that might not be beneficial to include, it is reasonable to want to improve the model for diabetes progression.

The Best Subsets Model

A best subsets model is generated by comparing the MSE for each model containing from one to ten predictors of diabetes regression. The criteria used for model selection in this case is the Bayesian information criterion (BIC). We will select the model that contains the number of predictors which yield the smallest BIC.



The BIC is clearly at its minimum value for the 6 predictor model.

```
## (Intercept)         sex         bmi         map          tc         tch
##    150.1166   -306.0420    538.8274    389.0673   -379.0379    332.6735
##         ltg
##    527.5658
```

The resulting analysis shows that the model that yields the smallest BIC is a six predictor model with the above coefficients. The resulting test MSE for the six predictors model is 3095.48. This best subset model is an improvement over the full linear model as it has reduced over fitting by a great deal. The predictors age, ldl, hdl, and glu have been removed from the model, which is in line with what we suspected about them based on their p-values in the full model.

10-Fold Cross-Validation model

The 10 fold cross validation model yields the same result to the BIC selection criterion. The same six predictor model is chosen. This model is chosen because the validation test error was smallest with 6 predictors. Since the predictors are not changed in our best subset from the BIC method, the coefficients and test MSE will remain the same. The test MSE is 3095.48, and the coefficients will also remain the same.

```
## (Intercept)         sex         bmi         map          tc         tch
##    150.1166   -306.0420    538.8274    389.0673   -379.0379    332.6735
##         ltg
##    527.5658
```

Ridge Regression Model

The ridge regression model using the criterion of selecting the largest lambda value, such that the cross-validation error is wtihin 1 standard error of the minimum yields a larger MSE than the six predictor model using the 10-fold cross-validation best subsets method. The test MSE for the ridge regression model is 3070.636. Another weakness of this model is that it includes all 10 predictors. The ridge regression model is not much of an improvement over the original full linear model. The following coefficients are produced from the ridge regression model:

```
## (Intercept)         age         sex         bmi         map          tc
##   149.99086   -11.25502  -156.90281   374.44565   264.86245   -32.09103
##         ldl         hdl         tch         ltg
##   -66.97779  -173.82190   124.03502   307.72524
```

Lasso Model

The lasso model uses the same criterion as the ridge regression model. The lasso regression method has the advantage of removing predictors from the model and in this case does yield a stronger model.
The lasso model gives a test MSE of 2920.08 which is the smallest of any of the models. The resulting coefficients of the model are:

```
## (Intercept)         age         sex         bmi         map          tc
##    149.9530      0.0000   -119.6489    501.4859    270.9240      0.0000
##         ldl         hdl         tch         ltg
##      0.0000   -180.3035      0.0000    390.5745
```

Results

In summary, we have built five models for predicting the progression of diabetes with the given 10 predictors. The resulting MSE values are:

```
##                                         Model Test.Error
## 1                       Full Least Squares Model    3111.26
## 2                       Best Subsets Model (BIC)    3095.48
## 3 Best Subsets Model (10-fold Cross-Validation)    3095.48
## 4                         Ridge Regression Model    3070.64
## 5                                   Lasso Model    2920.08
```

Conclusion

While the best subsets model which uses 10-fold cross-validation and the lasso method using 10-fold cross-validation both yield a six predictor model, based on model selection through the lowest MSE, the lasso regression model will be chosen as the "best" model for the estimation of diabetes progression. Five of the six predictors in the two models are the same but the predictor hdl is included in the lasso model but tc is not. The predictors of age, tc, ldl, and tch have been reduced to zero, and so eliminated from the model that estimates diabetes progression.

Appendix - R Code

```
library(lars)
data(diabetes)
data.all <- data.frame(cbind(diabetes$x, y=diabetes$y))
n <- dim(data.all)[1] # sample size = 442
set.seed(1306) # set random number generator seed to enable
 # repeatability of results
test <- sample(n, round(n/4)) # randomly sample 25% test
data.train <- data.all[-test,]
data.test <- data.all[test,]
x <- model.matrix(y~.,data=data.all)[,-1] # define predictor matrix
 # excl intercept col of 1s
x.train <- x[-test,] # define training predictor matrix
x.test <- x[test,] # define test predictor matrix
y <- data.all$y # define response variable
y.train <- y[-test] # define training response variable
y.test <- y[test] # define test response variable
n.train <- dim(data.train)[1] # training sample size
n.test <- dim(data.test)[1] # test sample size
```

Linear Model

```
lm.train=lm(y~.,data=data.train)
coef(summary(lm.train)) #Linear model coefficients
mean((data.test$y-predict(lm.train, data.test))^2) #Test MSE linear model.
plot(lm.train$fitted.values, resid(lm.train), xlab = "Fitted Value", ylab="Residual")
abline(a=0.0, b= 0.0, col="red")
```

Best Subsets BIC Model

```
library(leaps)
subset.model= regsubsets(y~., data=data.train, nvmax=10)
reg.summary=summary(subset.model)
plot(reg.summary$bic, xlab="Number of Variables", ylab="BIC", type='l')
points(6, reg.summary$bic[6], col="red", pch=19)
coef(subset.model, which.min(reg.summary$bic))
test.mse=model.matrix(y~.,data=data.test)
val.errors=rep(NA,10)
  for(i in 1:10){
  coefi=coef(subset.model,id=i)
  pred=test.mse[,names(coefi)]%*%coefi
  val.errors[i]=mean((data.test$y - pred)^2)
  }
val.errors[6] #Test MSE for best subsets BIC Selection model.
```

10 Fold Cross Validation model

```
predict.regsubsets=function(object,newdata,id,...){
   form=as.formula(object$call[[2]])
   mat=model.matrix(form,newdata)
   coefi=coef(object,id=id)
   xvars=names(coefi)
   mat[,xvars]%*%coefi
  }

k=10
set.seed(1306)
folds = sample(1:k,nrow(data.train),replace=TRUE)
cv.errors=matrix(NA,k,10, dimnames=list(NULL,paste(1:10)))

for(j in 1:k){
   best.fit=regsubsets(y~., data=data.train[folds!=j,], nvmax=10)
   for(i in 1:10){
      pred=predict.regsubsets(best.fit,data.train[folds==j,], id=i)
      cv.errors[j,i] = mean((data.train$y[folds==j] - pred)^2)
   }
}

mean.cv.errors=apply(cv.errors,2,mean)
cvmin=which.min(mean.cv.errors)
val.errors[cvmin]
coef(subset.model, cvmin)
```

Ridge model

```
library(glmnet)
grid =10^seq (10,-2, length=100)
set.seed(1306)
cv.out = cv.glmnet(x.train, y.train, alpha=0)
ridge.mod = glmnet(x.train,y.train,alpha=0, lambda=grid)
ridge.pred = predict(ridge.mod,s=cv.out$lambda.1se, newx=x[test,])
mean((y.test-ridge.pred)^2) #Ridge MSE
predict(cv.out, type="coefficients", s=cv.out$lambda.1se)[1:10,] #ridge coef
```

Lasso Model

```
library(glmnet)
grid =10^ seq (10,-2, length =100)
set.seed(1306)
cv.out=cv.glmnet(x.train,y.train,alpha=1)
lasso.mod=glmnet(x.train,y.train, alpha=1, lambda=grid)
lasso.pred=predict(lasso.mod, s=cv.out$lambda.1se, newx=x[test,])
mean((y.test-lasso.pred)^2) #Test MSE
predict(cv.out, type="coefficients", s=cv.out$lambda.1se)[1:10,]
```