

Diabetes data analysis

Project 1 – STAT 897



September 29, 2017

Daljeet singh

Penn State

Table of Contents

[Introduction 1](#_Toc494209850)

# 

# Introduction

This report analyses the effectiveness of using different linear regression methods to examine diabetes progression. We will be using data collected from 442 patients. Specifically, we will examine whether age, sex, body mass index, average blood pressure, and six blood serum measurements are useful in explaining progression in a quantitative measure of disease progression one year after a baseline measurement. The data is sourced from Efron et al. (2003).

# Analysis

We start off with partitioning the patients into two groups: training (~75%) and test (~25%).

The first step is fitting a regular least squares multiple linear regression model using all the ten variables. The coefficients are:

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 151.885 3.065 49.555 < 2e-16 \*\*\*

age -6.387 71.461 -0.089 0.928837

sex -257.173 72.951 -3.525 0.000485 \*\*\*

bmi 513.830 78.892 6.513 2.84e-10 \*\*\*

map 335.714 77.309 4.342 1.89e-05 \*\*\*

tc -779.357 507.431 -1.536 0.125550

ldl 481.739 407.534 1.182 0.238047

hdl 85.036 262.514 0.324 0.746203

tch 262.487 197.443 1.329 0.184650

ltg 649.500 205.962 3.153 0.001766 \*\*

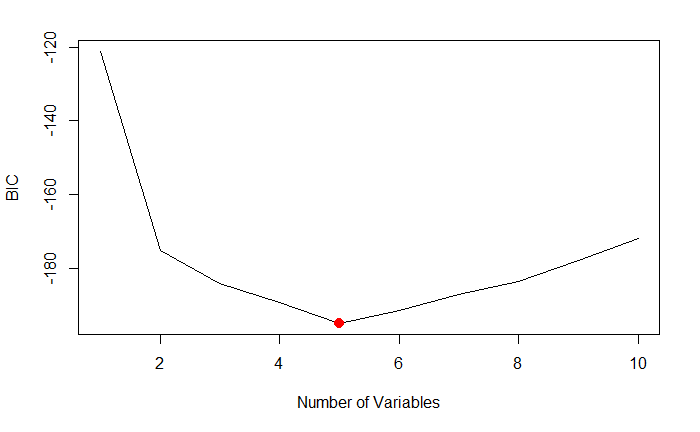
glu 117.226 76.871 1.525 0.128252

Interestingly we find that age, tc, ldl, hdl, tch and glu are not significant predictors of disease progression. This is interesting because these variables intuitively appear that they should be significant for instance age. The plots of the residuals:

|  |  |
| --- | --- |
|  |  |

**Test MSE for the full linear regression** model is: 2511.981

Next we will apply the best subset selection using BIC to select the number of predictors.



The lowest and therefore the best BIC is for 5 predictors. The BIC value for a model with 5 predictors is -194.7 and their respective coefficients are:

|  |  |
| --- | --- |
|  | Coefficients:  Estimate Std. Error t value Pr(>|t|)  (Intercept) 151.801 3.067 49.491 < 2e-16 \*\*\*  sex -242.577 72.001 -3.369 0.000845 \*\*\*  bmi 530.590 76.902 6.900 2.72e-11 \*\*\*  map 346.631 74.405 4.659 4.64e-06 \*\*\*  hdl -353.984 80.206 -4.413 1.38e-05 \*\*\*  ltg 426.250 78.251 5.447 1.01e-07 \*\*\* |

**Test MSE for the best subset** model based on BIC with 5 parameters is: 2506.565

When we use the cross validation approach with 10 folds, the training data is divided into 10 parts and the approach loops through each time using 1 fold as validation and the remaining as training data. With this approach, too we get the minimum cross validation error for the same 5 parameter model as above. The coefficients are listed below (they will be same as the last output).

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | |  |  | | --- | --- | | (Intercept) | 151.8015 | | sex | -242.577 | | bmi | 530.5897 | | map | 346.6307 | | hdl | -353.984 | | ltg | 426.25 | |

**Test MSE for the best subset** model based on 10-fold cross validation is: 2506.565

We will now move to the ridge regression approach. We used 10-fold cross validation and selected the largest value of λ such that the cross-validation error is within 1 standard error of the minimum.

The coefficients with ridge are:

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| (Intercept) | age | sex | bmi | map | tc | ldl | hdl | tch | ltg | glu |
| 152.10 | 29.72 | -105.00 | 333.38 | 221.14 | 4.64 | -20.74 | -178.61 | 148.64 | 255.54 | 133.29 |

**Test MSE for the ridge regression** model is: 2852.122

In the last step, we will use lasso model. We used 10 fold cross validation and selected the largest value of λ such that the cross-validation error is within 1 standard error of the minimum.

The coefficients with ridge are:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| (Intercept) | bmi | map | hdl | ltg | glu |
| 152.332 | 491 | 192.21 | -146.9 | 381.2 | 24 |

**Test MSE for the lasso regression** model is: 2599.714

# Results

|  |  |
| --- | --- |
| Model | Test MSE |
| Full linear regression | 2511.981 |
| Best subset model based on BIC | 2506.565 |
| **Best subset model based on 10-fold cross validation** | **2506.565** |
| Ridge Regression | 2852.122 |
| Lasso Regression | 2599.714 |

# Conclusion

Based on the Test Mean squared error, we find that the best subset models behave the best. Lasso and full linear regression are very close while ridge regression performs the worst.

# Appendix