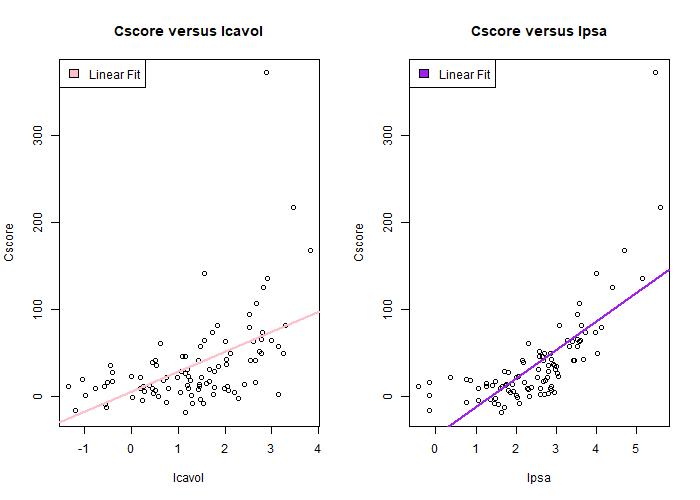
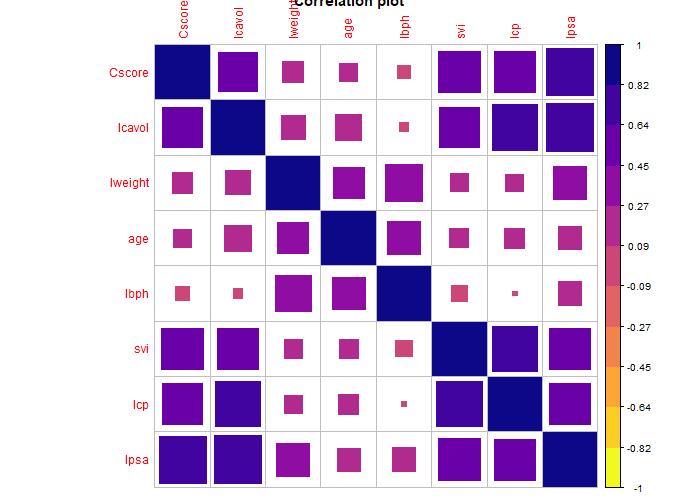
Data analysis began with a general survey of the data. The response variable, Cscore, and all the predictor variables except svi are continuous variables. This was seen when summarizing the dataset, svi took values of 0 or 1. In addition, a scatterplot of the data showed there may be a non-linear relationship between the response variable, Cscore, and lcavol (Figure 1) and between Cscore and lpsa (Figure 2). These figures show a clear non-linear trend that will be determined later. This may hold significance when trying to predict because linear models will not be able to capture this relationship.



Figure 1: Plots displaying the non-linear relationship between Cscore and lcavol(left) and Cscore and lpsa(right

The correlation between each of the variables was calculated and visualized. Figure 3 showed a strong correlation between Cscore and lcavol, svi,lcp and lpsa which can be visualized by the darker purple squares from Figure 2.

This indicates that these predictors could be important in explaining/ predicting Cscore.

Figure 2: Correlation plot displaying the high correlation between Cscore and 3 other predictor variables in addition to a high correlation between lcavol and other predictors

The correlation values were also used to display that exists a high correlation (closer to 1) for many predictor variables from Figure 3. The values highlighted in pink show that svi, lcp, and lpsa show a high correlation with lcavol. This might be a case of multi-collinearity. In addition, many variables are highly correlated to the response variable.

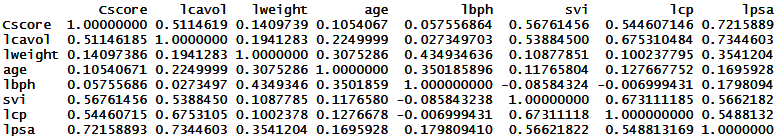


Figure 3: Correlation values to quantize the relationship

Next, the shape of the distribution(histograms) and linearity assumptions were checked. Figure 4 shows a right-skew in the distribution of the data. Due to the fact that this is not normally distributed, further step that could be taken is to use a gamma distribution to transform the data into a normal distribution.

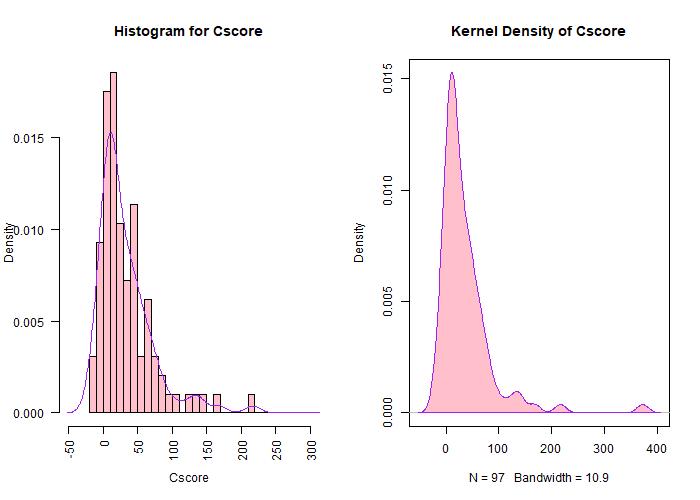


Figure 4: Showing a right-skew distribution that can be further understood using different link functions

The next step was to apply a shrinkage/ regularization technique to shrink the coefficients estimates to exactly 0 in order to determine the variables that contribute the most to the model. This is done by applying a penalty to the coefficients estimates. Ideally, applying a penalty aims to reduce the variance at the expense of a small and negligible increase in bias. In addition, the penalty is formulated on the basis of the size of the coefficients and therefore the scale of the data should be standardized in order to ensure comparability. Luckily, glmnet function will standardize the data.

The data was split into the training and testing dataset in order to derive an optimal tuning parameter by Cross Validation and determine the test MSE. A vector was created describing the possible values for Lambda (tuning parameters) that are allowed in this scenario.

Upon visualization of figure 5, the variables that remained after fitting a lasso regression on the training set were: lpsa,svi,lcp,lweight,and lcavol. Figure 5 shows that lcavol was one of the last variabes to enter the model and the first to leave in the log lambda graph. However, the importance of the variables is not derived by the magnitude of the coefficients but the order in which they enter and exit the model. This is seen in the Figure to the right. For example, in the range of [-5,0] log lambda, Lcavol has a higher absolute value of coefficient in this range as compared to variables such as lweight or lcp. However, lcavol shows a steep plummet towards 0 and leaves the model earlier than the other variables. In addition, if the coefficients of svi1 and lcavol are assumed to be the same in terms of magnitude (absolute value), svi1 is able to stay in the model for a longer amount of time. Therefore, the importance of variables does not lie in its magnitude of coefficients but in the order in which they leave the model at high log lambda. Lpsa shows strong importance in the model and this is shown not by the magnitude of the coefficient but by its ability to be retained in the model even at large penalty values.

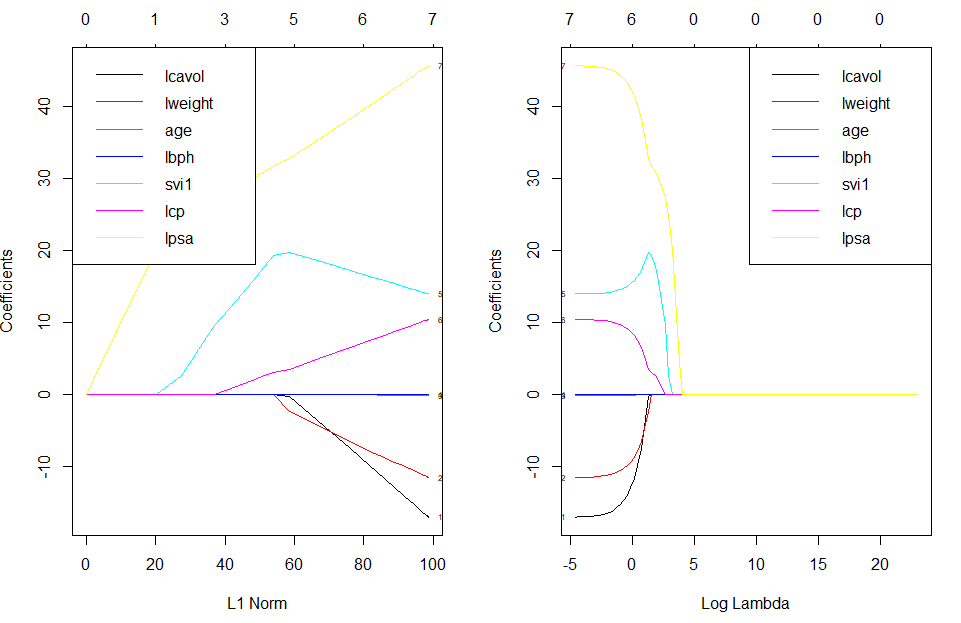


Figure 5: Plots used to show the importance of certain variable and their contribution towards the model fitted using LASSO

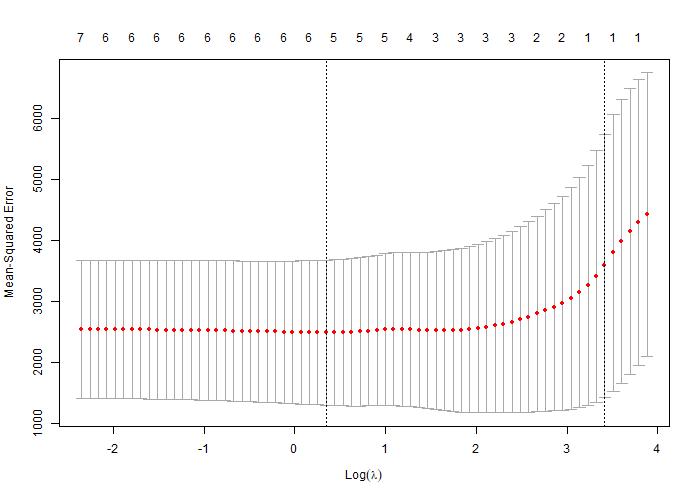
To find the appropriate value of the tuning parameter to balance bias and variance, a grid of possible values was given. The best lambda was determined by cross validation to be 1.41. The graph shows that the no further shrinkage of coefficients improved the mean-square error and 5 coefficients were left in the model.

Figure 6: Plot displaying the amount of coefficients left in the model that contribute to a low training MSE

The testing set was used to determine the test mean-square error of 717.7531. LASSO then fit a regression based on the shrinkage of the coefficients:

If a linear model is fitted using the features selected by LASSO, the following model is generated:

According to this shrinking method, using a tuning parameter of 1.41, a unit increase in lcavol will lead to a 3.02 decrease in Cscore while holding all other variables constant. LASSO is a feature selection technique that works by removing predictors (setting their coefficients to 0) that do not contribute to the model. From figure 5, lcavol was one of the variables that remained in the model even after the penalty was added. Penalties are used to shrink variables that do not explain much variability in the model to 0. Therefore, due to the fact that lcavol remained in the model shows its importance in its ability to predict Cscore (but not as important as lpsa that stayed in the model even at larger penalties).  
Because this is an underlying linear function, the link function would be the Identity function and the error function would be the mean squared error(MSE).

As an introduction to the last part of the assignment (question 4), I used a makeshift cross validation set up to determine the best non-linear fit of lcavol and lpsa which were shown in earlier parts to have a non-linear relationship with the response variable. Possible values were taken to be between 1 and 10. The test MSE was checked for each to find the minimum polynomial with the lowest MSE (Figure7)

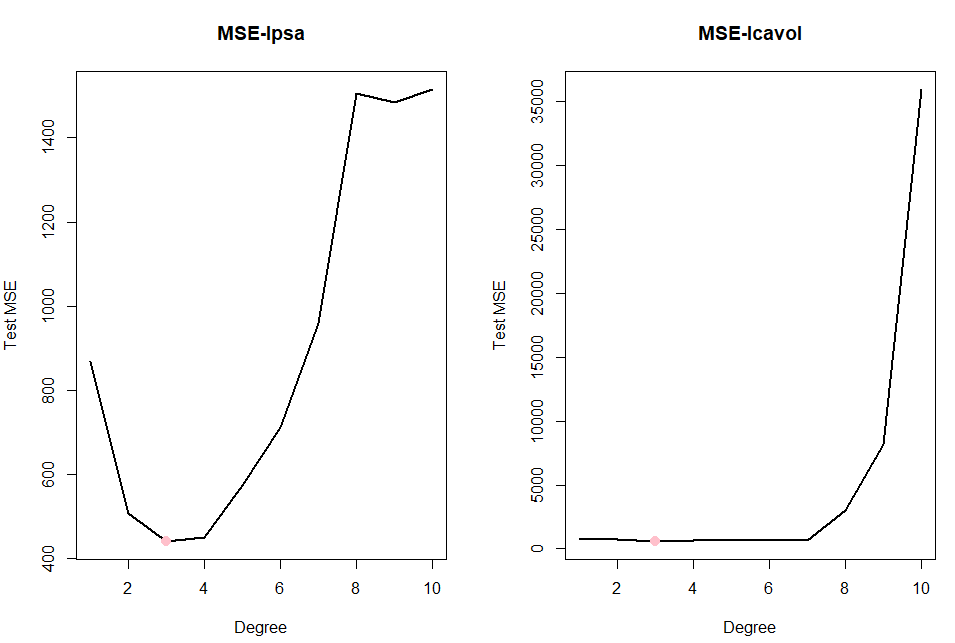
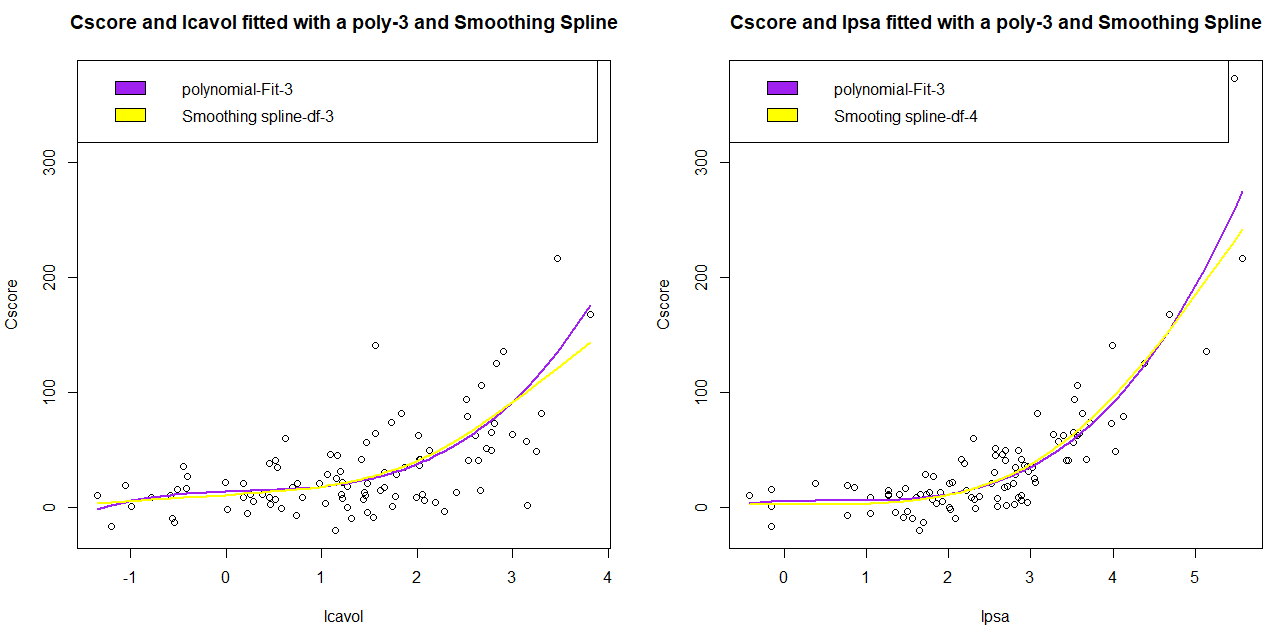


Figure 7: Plot displaying the optimal degree of polynomial(pink) with respect to the lowest Test MSE.

The minimum polynomial was found to be 3 which will be the underlying base for the rest of the problem (cubic spline).

  
Figure 8: Plots displaying the non-linear relationship between lcavol an lpse with respect to the response variable Cscore. In purple is a cubic polynomial (purples)found using cross validation and smoothing spline(yellow)

From these figures, we do see a data-sparse region at the extremities which lead me to want to explore smoothing splines. Cross-validation resulted in the optimum amount of degrees of freedom for smoothing splines of models with lcavol (df=3.8) and lpsa(df=4) displayed in yellow. In order to have a full model, the generalized additive model was used to combine the effects of the 2 smoothing splines and the rest of the variables and the test MSE was derived. I was also interested in seeing whether or not age and lpha, which were removed from the lasso model, would contribute to a reduced test MSE when used with the generalized additive model.

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Effect | Test MSE |
| GAM+smoothing spline | gam.m | Cscore~s(lcavol,df=3.842969)+s(lpsa,df=4.004397)+svi+lweight+lcp+age+lbph | 594.16 |
| gam.m1 | Cscore~s(lcavol,df=3.842969)+s(lpsa,df=4.004397)+svi+lweight+lcp+age | 599.32 |
| gam.m2 | Cscore~s(lcavol,df=3.842969)+s(lpsa,df=4.004397)+svi+lweight+lcp+lbph | 635.92 |
| gam.m3 | (Cscore~s(lcavol,df=3.842969)+s(lpsa,df=4.004397)+svi+lweight+lcp | 645.86 |
| lasso | out\_lasso | -6.94-3.02\*lcavol+27.78\*lpsa +18.85\*svi -6.80\*lweight +5.13\*lcp | 717.75 |
| lm | ls.fit | -9.27-17.16\*lcavol+45.68\*lpsa+13.84\*svi-11.98\*lweight+10.56\*lcp | 826.47 |

Table 1: Table displaying the Test MSE of the models. Highlighted in yellow are the interesting models to compare

The linear model allows the easy interpretation on the effect of a unit increase on the response. However, the majority of world problems are not modelled as linear problem. For example, having the response variable that does not follow a Gaussian distribution or that the relationship between the response variable and the predictor variables are non-linear. In these cases, one might opt for a model that can capture this relationship- polynomial regression or splines. Smoothing splines, somewhat like lasso, applies a smoothing parameter and penalty to control the variability and flexibility of the model. The Advantages of this method lie in its ability to capture non-linear trends and allow predictions. However, a possible disadvantage would be that the relationship between a predictor and the response can no longer be summarized into an exact quantity as an ordinary least squares regression is able to do. For this reason, the model may lack in some interpretability. For the models designed in this assignment, the generalized additive model with all the predictors, gam.m, had the lowest test MSE compared to that generated by LASSO and OLS and therefore performs better in this dataset. This may be due to its ability to capture the non-linear trends that were mentioned in part 1 of the assignment in figure1 and figure 8. However, the difference between the 3 model’s MSE is not visually substantial.

The MSE that were calculated are dependent on the method in which the data was split. In my assignment, the dataset was split into half training and half testing set. If it was partitioned as 80/20 or 70/30 the test MSE would be different.

For this reason, Akaike information criterion was used to compare the models.

|  |  |
| --- | --- |
| Model | AIC |
| Ls.fit-Linear model of features selected by LASSO | 506.05 |
| Gam.m | 462.63 |
| Gam.m1 | 460.60 |
| Gam.m2 | 460.80 |
| Gam.m3 | 458.76 |

Table 2: table displaying the Akaike information criterion for 5 models

Therefore, if GAM.M3 is taken as the model to be used.

When comparing using the Akaike information criterion, gam.m3, the generalized additive model using smoothing spline but having removed age and lbph, had a lower AIC than the full model. This is intuitive because AIC penalizes the addition of extra variables, therefore having a simpler model is more important. If the effects of the nonlinear terms are to be interpreted, they must be interpreted in terms of the basis functions and the linear effects can be interpreted normally. The mean square error was used as an evaluation of the models generated. The MSE is the average of the squared differences observed between the predicted/estimated values and the true values. However, the MSE is relative. A smaller MSE, relative to others, will indicate that the model provides estimates of the coefficients that are more accurate as compared to other models and their estimates. The MSE generated from GAM had a lower MSE when compared to models generated by LASSO or ordinary least squares. Therefore, they have more predictive power and can be assumed to be more accurate in its abilities to capture trends found in the dataset. However, GAM often suffer from the caveat of overfitting and this must be taken into account when using this model for predictions.

In conclusion, both models performed well in terms of the Test MSE. However, in order to compare the models, a question must be asked. The choice between the models is dependent on the problem at hand. If one requires the model to be easily interpretable, a linear model may be a feasible option. However, if prediction is required, a model that is able to capture more complex relationships such as the GAM models may be more suited to the problem. Further research can go into transformation of the data using various link functions to achieve a lower MSE and other methods of comparison such as can be used.

# References

Gareth James, D. W. (2017). *An Introduction to Statistical Learning.*

https://stackoverflow.com/questions/30560689/adding-labels-on-curves-in-glmnet-plot-in-r