LASSO Regression and GAMs

James O'Reilly
james.oreilly@student.kuleuven.be

Introduction

For this assignment we use the prostate dataset. The dataset contains data about prostate cancer patients with information on the size of the prostate, the age of the patient, a blood marker (lpsa), and so on. The response variable is a score (Cscore) on the progression of the cancer after detailed study of the tumour pathology. It is difficult to find information on what the Cscore actually represents, but my guess is that it is some continuous variant of the Gleason score, which is used to help evaluate the prognosis of men with prostate cancer using samples from a prostate biopsy.

Question One

Study and describe the predictor variables. Do you see any issues that are relevant for making predictions?

The response variable 'Cscore' is a continuous variable. Each of the predictor variables and their meanings are given below.

Variable	Abbreviation
Log cancer volume	LCAVOL
Log prostate weight	LWEIGHT
Age in years	AGE
Log benign prostatic hyperplasia amount	LBPH
Seminal vesicle invasion (Yes = 1, No = 0)	SVI
Log prostate specific antigen	LPSA
Log capsular penetration	LCP

Seven of the eight predictor variables are continuous, while the svi variable needs to be treated as a categorical variable. A scatterplot matrix for the prostate dataset is given in Figure 1. From this, we can try to spot any non-linear relationships between the response and predictor variables.

Looking at Figure 1, both 'lcavol' and 'lpsa' predictor variables seem to vary non-linearly with the predictor variable. Each of these plots, along with regression lines, are presented in Figure 2. Looking at these plots, it is clear that there is a strong non-linear relationship between 'lpsa' and the response variable. A much weaker non-linearity exists between 'lcavol' and the response variable. This will be taken into account in the final question where we fit a model with non-linear effects.

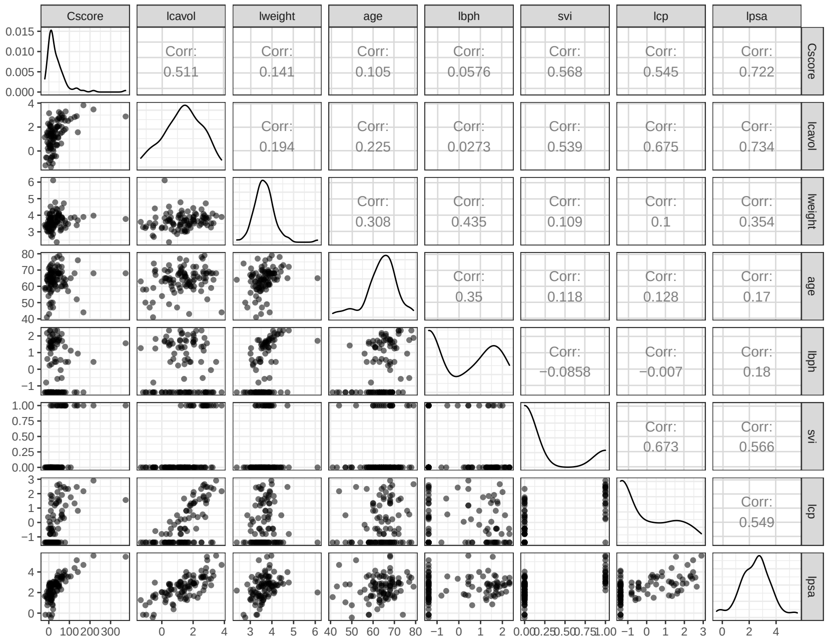

Figure 1: A scatterplot matrix for the `prostate` dataset

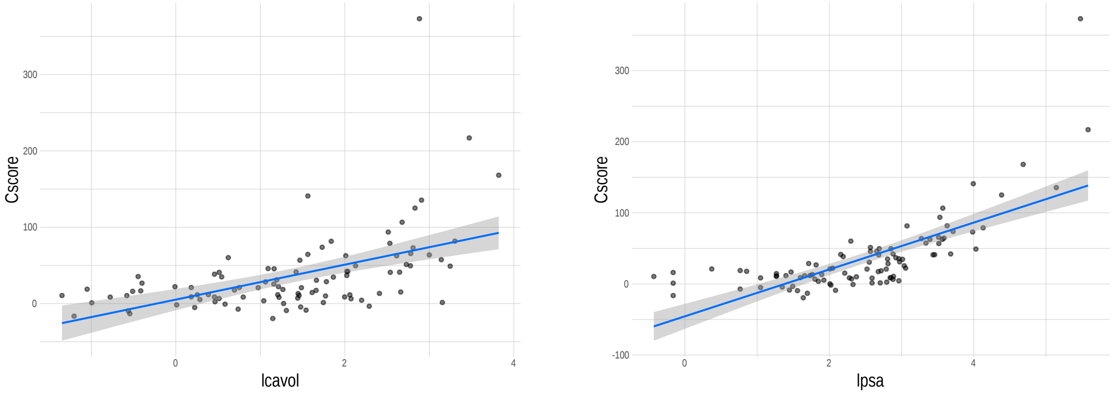

Figure 2: Scatter plots and regression lines for LCAVOL *(left)* and LPSA *(right)*. The 95% confidence intervals for the regression lines are also shown.

Question Two

Make an appropriate LASSO model, with the appropriate link and error function, and evaluate the prediction performance.

Choosing the Appropriate Link Function

Linear regression assumes that the response variable is normally distributed. Generalised linear models can have response variables with distributions other than the Normal distribution – poisson, gamma, multinomial, etc. The link function defines the relationship

$$f(\mu) = Xb$$

between the mean response μ and the linear combination Xb of the predictors. The link function chosen is dependent on the distribution of the response. In order to choose the correct link function, a histogram of the response variables along with a smoothed density estimate was plotted (see Figure 3). The distribution is evidently not perfectly normal, and seems to follow a gamma distribution. However, the `glmnet` function in R does not allow fitting to a gamma distribution, and the other options it provided were not more desirable than with a Gaussian distribution. Instead, an identity link function is used with the Gaussian distribution.

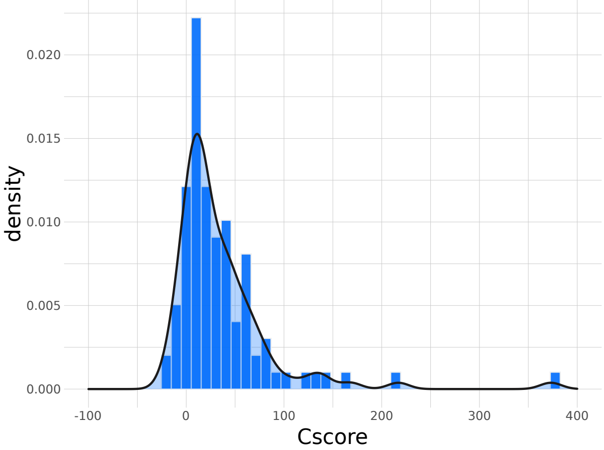

Figure 3: The density of the response variable. It is clear the distribution is right-skewed and non-normal.

Fitting the Model

The data was then split into a training and test set, and a LASSO model was fitted. Note that the data should be appropriately scaled before performing LASSO, but this is done automatically by the `glmnet` function. Cross-validation was used with a grid search to determine the optimal shrinkage parameter λ. Figures 4 and 5 show how both the coefficients and error on the test set vary with λ, with the optimal $\log(\lambda)$ given by the dotted vertical line in Figure 4. The best MSE on the test set was 725.94, with $\lambda = 1.28$.

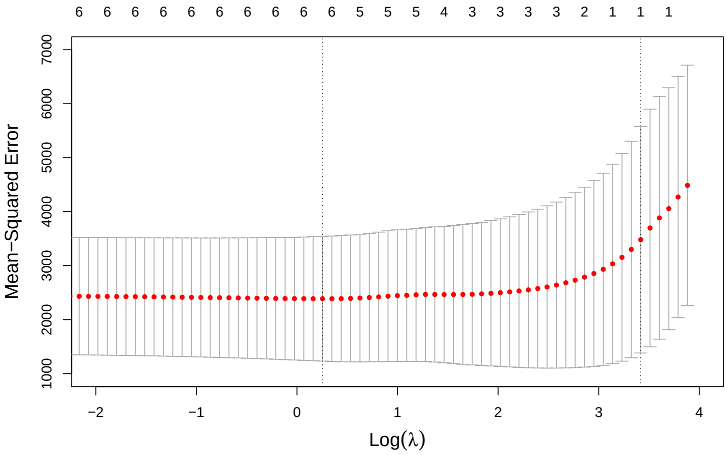

Figure 4: Mean-squared error vs $\log(\lambda)$. The first vertical line indicates the value of $\log(\lambda)$ for which the MSE is smallest. The numbers at the top of the graph indicate the number of variables left in the model.

The final model is given by

$$\text{CSCORE} \sim -5.68 - 3.61\text{LCAVOL} - 7.23\text{LWEIGHT} + 18.86\text{SVI} + 5.36\text{LCP} + 28.24\text{LPSA} \quad (1)$$

From the model above, you can see that LASSO set the coefficients for both AGE and LBPH to zero, effectively performing variable selection.

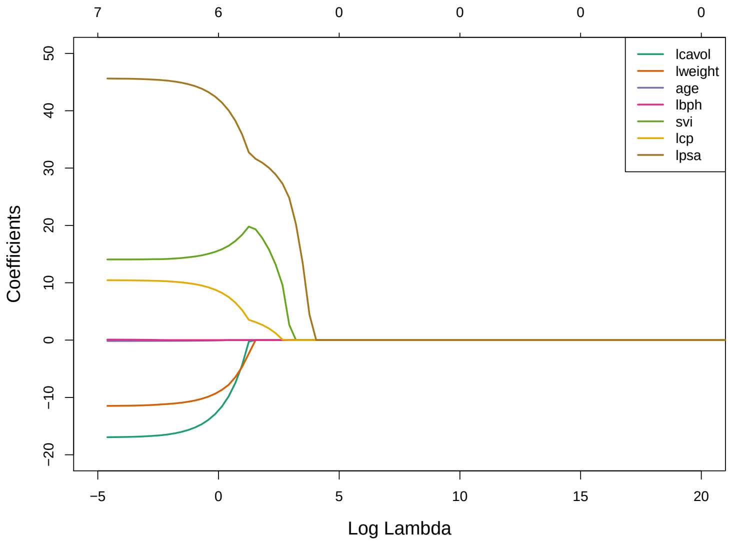

Figure 5: Trajectories of the coefficients vs $\log(\lambda)$. The numbers at the top of the graph indicate the number of variables left in the model.

Transforming the Response

In question one it was noted that the distribution of the response resembled a gamma distribution. The regression performed above neither normalised the response variable before fitting, nor used an appropriate link function to compensate for this. As the `glmnet` function does not provide the option to set `family = 'gamma'`, one cannot use the appropriate negative inverse link function. Instead, the response variable can be transformed using the `bestNormalize` function prior to fitting the model. This function uses the Lambert, Box Cox, or Yeo-Johnson transformations to transform normalise the response. The model can then be fitted using this normal response. The predicted response values given by this model can then be transformed back to the original non-normal distribution, giving properly scaled response values. The response distributions before and after normalisation are given in Figure 6.

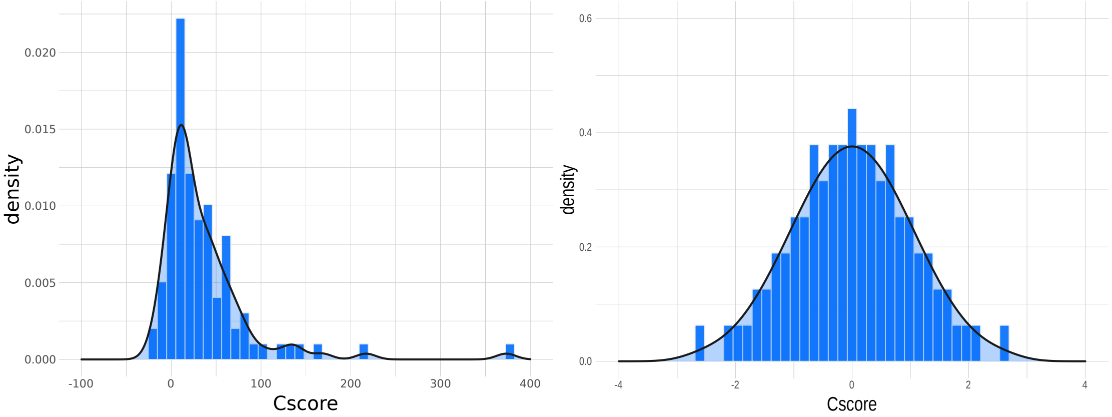

Figure 6: Original and normalised response distributions, along with density curves. Note the difference in axes.

LASSO was performed using the same cross-validation approach to determine the optimal λ. A GLM was fitted using this distribution and used to predict response variables, which were then transformed to the original response distribution. The MSE on the test set was 443.08, a significant improvement compared to the 725.94 from the previous model. The final model with

(SVI = 1) is given by

$$\begin{aligned}\text{CSCORE} \sim - &- 0.0184\text{LCAVOL} + 0.0072\text{AGE} + 0.0156\text{LBPH} + 0.0065\text{SVI}\\ &+ 0.1323\text{LCP} + 0.3081\text{LPSA}\end{aligned}$$

It is important to note that while the transformed model achieved a better performance on the test set, this does not mean it should be chosen over the previous model. Interpreting the coefficients given by the transformed model is difficult. This is a classic example of the tradeoff between model interpretability and model accuracy. Which model is chosen depends on the questions we wish to ask. If we are interested solely in predicting Cscore as accurately as possible, then the transformed model is better. If we instead want to understand how Cscore will vary with the predictors, and understand the specific impact of each predictor, then we would choose the original model.

Question Three

Look at the coefficient for 'lcavol' in your LASSO model. Does this coefficient correspond to how well it can predict Cscore? Explain your observation.

Assuming the response variable is normally distributed, the coefficient for the LCAVOL variable β_{lcavol} is -3.6. How do we interpret this? Firstly, we can give the same interpretation we would give with any linear model: a unit increase in LCAVOL yields a decrease of 3.6 in CSCORE, if all other variables are held constant.

Does this coefficient correspond to how well it can predict CSCORE? In short, no. It is tempting to claim that the relative magnitudes of the coefficients given by the LASSO regression describe to the relative importance of the corresponding variables in the model, as LASSO shrinks the coefficients of the non-important variables toward zero. While this appears to make sense, it is actually false. Given in Figure 7 is a plot of the trajectories of the coefficients as λ increases, focused on the range of values of λ at which the coefficients are set to zero. Looking at this figure, we can see that the coefficient trajectories for LCAVOL and LWEIGHT intersect. What this tells us is that for two different values of λ, either LCAVOL or LWEIGHT will have a coefficient with larger magnitude.

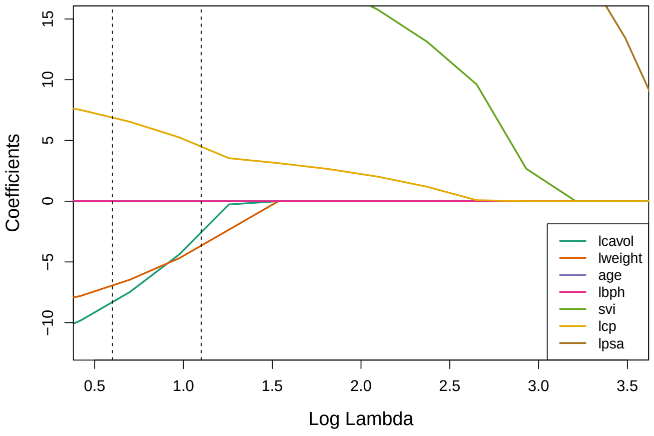

Figure 7: Coefficient trajectories. Focused on the points where the coefficients are set to zero. The vertical black lines show how the relative magnitudes of LCAVOL and LWEIGHT change at different values of λ. This illustrates why it false to make inferences on the relative importance of variables based on the magnitudes of their relative coefficients at a given point.

However, as the coefficient for LCAVOL reaches zero before the coefficient for LWEIGHT, we know that LCAVOL has less predictive capacity than LWEIGHT. If we were to make inferences about the predictive capacity of these variables based on their coefficients, there are values of λ at which we would have wrongly determined that LWEIGHT is less important than LCAVOL. This example shows that the relative magnitudes of the coefficients at a given value of λ do not tell us about the relative importance of variables to the model. Therefore the coefficient for LCAVOL does not correspond to how well it can predict CSCORE.

Ignoring the coefficients, however, and looking solely at Figure 7, we can say that LCAVOL has is less important (i.e. has less predictive capacity) than the other variables which were retained in the model, as it is the first of these coefficients to be set to zero.

Question Four

Fit a model with appropriate non-linear effects.

Figure 2 shows that two of the predictor variables, LPSA and LCAVOL, have a non-linear relationship with the response variable. In this section, we relax the linearity assumption and fit a generalised additive model with non-linear functions for the aforementioned variables. The non-linear variables which were not included in the LASSO model were also removed from the GAM. Smoothing splines were fitted to the non-linear predictors, with both the number of knots and the size of the smoothing parameter λ determined by cross-validation. The generalised additive model is therefore given by

$$\text{CSCORE} \sim \beta_0 + \beta_1\phi_1(\text{LCAVOL}) + \beta_2\phi_2(\text{LPSA}) + \beta_3\text{LWEIGHT} + \beta_4\text{SVI} + \beta_5\text{LCP}$$

where ϕ_1 and ϕ_2 are the smoothing splines fitted for LCAVOL and LPSA, respectively.

Before building the GAM, both the optimal number of knots and the optimal smoothing parameter λ must be calculated for the smoothing splines. This is done using cross-validation, which is performed automatically by the `smooth.splines` function. The splines fitted for LCAVOL and LPSA are shown in Figure 8.

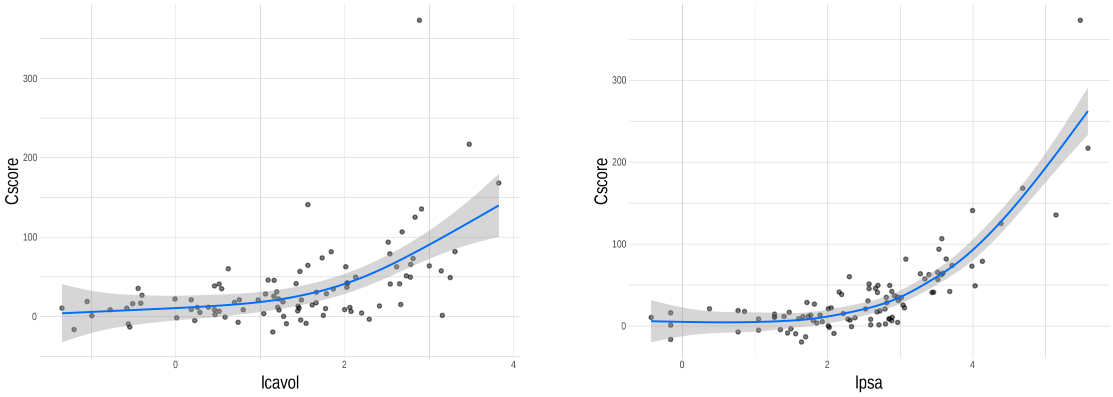

Figure 8: Fitted splines for LCAVOL and LPSA, with 3.8 and 4 degrees of freedom, respectively.

The prostate data was randomly divided into training and test data in the ratio of 4 : 1. The model was then fitted using the `gam` function on the training data and the MSE was calculated on the test data. The final model is given by

$$\text{CSCORE} \sim -52.85 - 8.64\phi_1(\text{LCAVOL}) + 40.19\phi_2(\text{LPSA}) + 0.38\text{LWEIGHT} - 1.45\text{SVI} + 8.14\text{LCP}$$

where ϕ_1 and ϕ_2 are the smoothing splines fitted for LCAVOL and LPSA, respectively.

Comparing Models

Report a comparison of performance to LASSO and explain what you find.

The GAM model and the LASSO regression model can be compared both quantitatively and qualitatively. A quantitative comparison includes assessing the performance of these models on test data. Usually the performance metric is MSE or R^2. The average test errors (MSEs) for the LASSO regression model and generalised additive model were 725.94 and 233.46, respectively. The GAM gives a significant improvement in prediction accuracy, even above the LASSO performed on a normalised response variable, which had an MSE of 443.08.

Qualitatively, the models can be compared in terms of their interpretability. While GAMs will give better performance than GLMs, they are often more difficult to interpret. More specifically, the coefficients for the non-linear variables cannot be interpreted in the same manner as for linear variables. For the GAM given in the equation above, it is difficult to interpret the coefficient of 40.19 for LPSA, as the LPSA variable has effectively been transformed through a smoothing spline. It is possible, however, to interpret the partial effects of the non-linear terms visually, though this is less rigourous. Lastly, note that as the model is additive, the coefficients for the linear variables in the GAM can be interpreted in the standard manner. For this reason, GAMs are often useful for accounting for a non-linear phenomenon that is not directly of interest, but needs to be accounted for when making inference about other variables.

To conclude, which model is best depends entirely on the questions being asked. If one is interested solely in predictive capacity, then the GAM is the better choice, as it can effectively account for non-linear relationships. When inferences about variables in the model, which model is best depends on the linearity or non-linearity of the variable. If the variable is linear and you wish to account for other non-linear variables, then the GAM is the better model. If instead you wish to make inferences about the effects non-linear variables, this can be difficult with a GAM, and a GLM might sometimes be the better choice.

References

[1] James, G., Witten, D., Hastie, T., and Tibshirani, R. *An introduction to statistical learning*, vol. 112. Springer, 2013.