**Assignment Part Two** Chun-Ting Wu r0915592

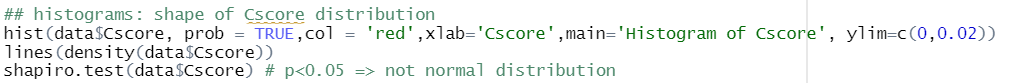
**Statistical Methods for Bioinformatics**  Yu-Jie Qiu r0823712

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

|  |  |  |
| --- | --- | --- |
| **Attributes** | **Description** | **Mean±SD** |
| Cscore | progression of the cancer score | 36.152±52.72 |
| lcavol | log of cancer volume (𝑐𝑐) | 1.350±1.179 |
| lweight | log of prostate weight (𝑔) | 3.629±0.428 |
| age | age of a patient (years) | 68.866±7.445 |
| lbph | log of the amount of benign prostatic hyperplasia (BPH) (𝑐𝑚2). A noncancerous enlargement of the prostate gland, as an area in a digitized image | 0.100±1.451 |
| svi | Seminal vesicle invasion; 1=Yes, 0=No.  Indicator of whether prostate cancer cells have invaded the seminal vesicle. | {'0'=76, '1'=21} |
| lcp | log of capsular penetration (𝑐𝑚).  Represents the level of extension of cancer into the capsule (the fibrous tissue which acts as an outer lining of the prostate gland) | -0.179±1.398 |
| lpsa | log Prostate specific antigen (PSA) (𝑛𝑔/𝑚𝐿) | 2.478±1.154 |

The data contains 8 variables and 97 clinical measurements which is a relatively small sample size making it difficult to make accurate predictions (lead to overfitting). Table 1 shows the descriptions and brief statistical summary of the variables, where the svi (seminal vesicle invasion) is dummy variables (no=0/yes=1).

Table 1. Description and statistics summary of all eight attributes for prostate cancer dataset.

The distribution of response variables, Cscore, are shown in Figure 1, ranging from -19.473 to 373 which is a very high range. The response variables are not normal distribution by identifying from the shapiro test (p<0.05).

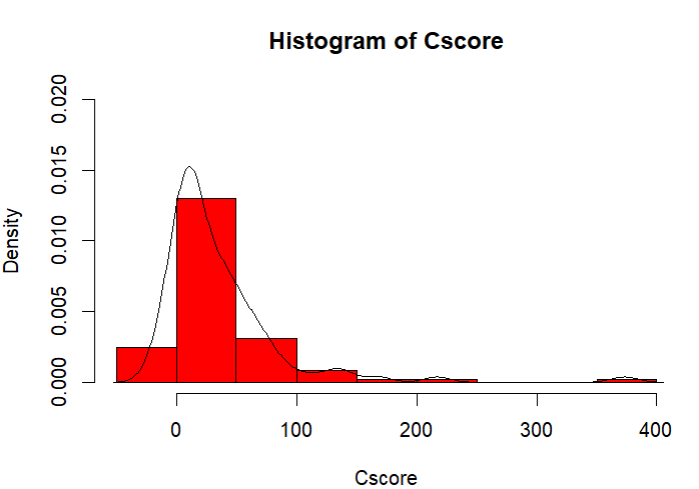


Figure 1. The distribution of Cscore.

一張含有 文字 的圖片

自動產生的描述Comparing Cscore between group 0 and group 1 can be seen in Figure 2 and Table 2. Because of the non-normal distribution of Cscore, we use Mann Whitney U test to analyze whether the score is different between group 0 and group 1. The result indicates that the Cscore in group 0 is significantly lower than group 1 (p<0.001). However, in the data, we can find that there is a large difference in the numbers between group 0 and group 1. The number of group 1 is only 21 but there are 76 patients of group 0 in the data. The highly unequal sample size would cause heterogeneity of variance (F test to compare two variances: p<0.05) and further dramatically affect type I error rate (p-value) as well as loss of statistical power (type II error), which means the accuracy of prediction decreased.

Table 2. The summary of Cscore in two svi groups.

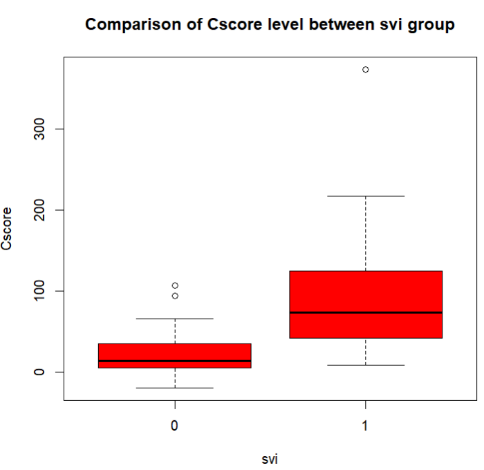


Figure 2. Comparison of Cscore in two svi groups.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Cscore | | | |
| group | **N** | **mean** | **std. error** | **Variance** |
| 0 | 76 | 20.5 | 24.1 | 579 |
| 1 | 21 | 92.8 | 82.9 | 6870 |

1. **Generate your best linear regression model using only linear effects. Are there any indications that assumptions underlying inferences with the model are violated? Evaluate the effect of any influential point, or outlier.**

We also use a scatterplot to validate the effect of correlations between variables (Figure 3). The result points out that the correlation of lcavol&lpsa is higher than 0.7 which can be considered highly correlated.

一張含有 文字 的圖片

自動產生的描述In order to generate best linear regression model, first step is checking whether there is multicollinearity. High correlation of predictor variables when performing multiple regression is called multicollinearity. It is a problem because it increases the standard errors of the regression coefficients, leading to noisy estimates. The variance inflation factor (VIF) quantifies the effect of collinearity on the variance of our regression estimates (VIF = 1/1-R-squared). It is common to say that any VIF greater than 5 considered that there is a multicollinearity issue. The result show that all the VIF is lower than 3, so there is no multicollinearity issue in our dataset, and we don’t need to deal with it.

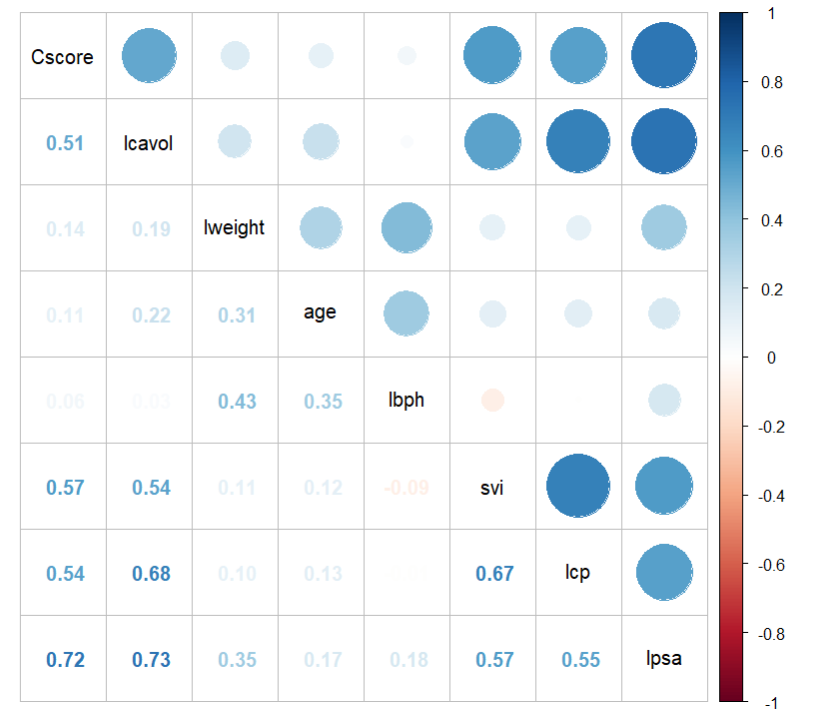
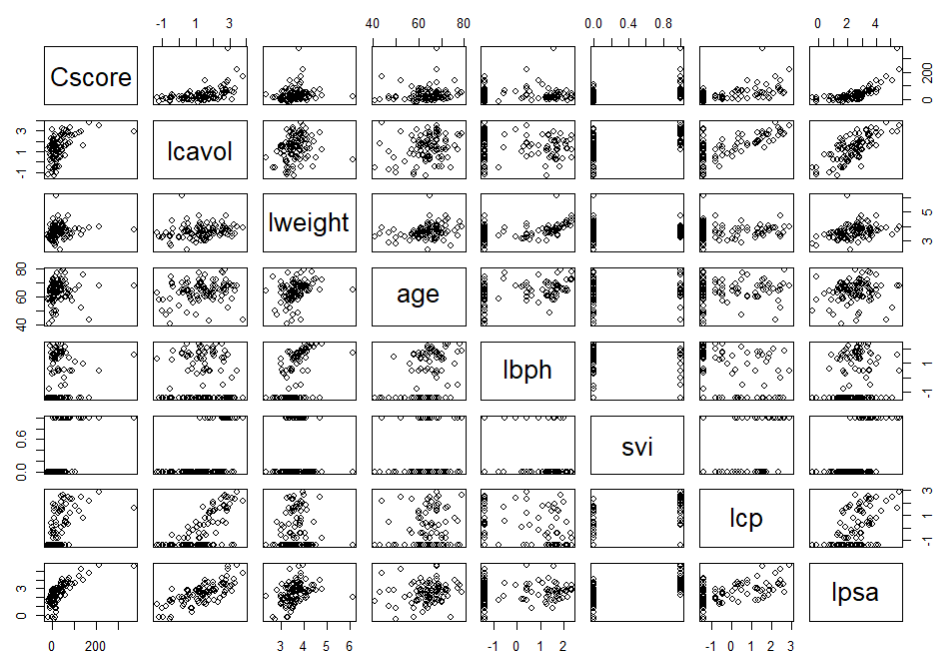


Figure 3. Result of correlation between variables.

Interestingly, we have found the outlier issue in our dataset. The residual in regression model fits a dataset well should fall randomly around zero and follow a normal distribution to obtain the unbiased fit. The residual plots and quantile-quantile (Q-Q) plot (Figure. 4) below point out that the index 96 patient is outlier. The right bottom of Figure. 4 also confirm that the Cook’s Distance of index 96 is larger than 0.5, which is an influential point. Outlier increases the variance of estimated coefficients, so the dataset used later is the dataset with the entire row of the outlier removed. The residual standard error and variance both decreased after remove the outlier (residual error: 35->24, variance: 2779 -> 1601).

一張含有 文字 的圖片

自動產生的描述It is noteworthy that if the point is high leverage point with high residual, it would impact the slope of the regression line. However, in this dataset, the index 32 is high leverage point with low residual. This point would not impact the coefficient or accuracy of the regression model, so we keep this point.

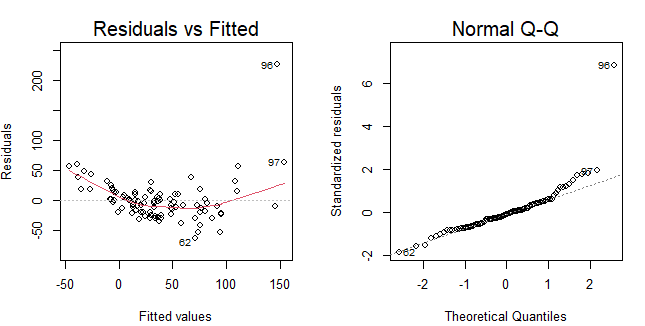
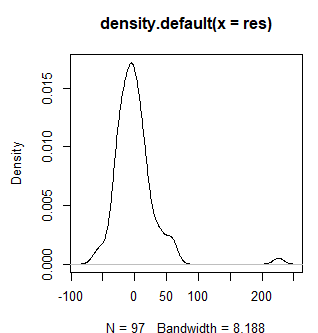
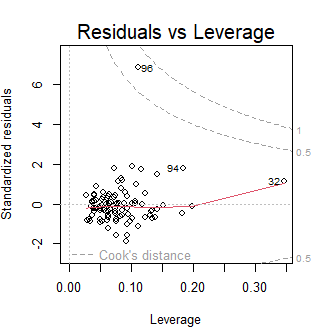
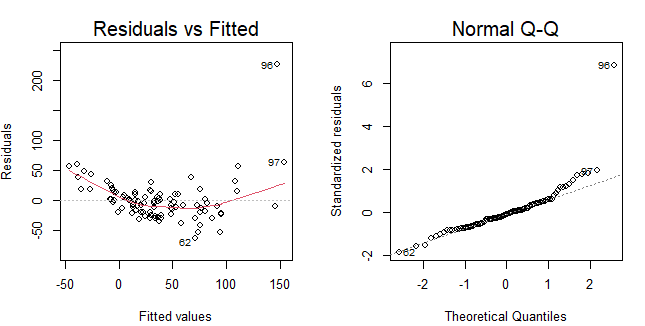


Figure 4. Residual and Q-Q plot detect outlier.

The normality of residuals assumption can be violated. This could happen if the relationship between dependent variable and independent variable isn’t linear or there exist outliers. the histogram of residuals is starting to look skewed, and the Q-Q plot isn’t looking as straight.



一張含有 文字 的圖片

自動產生的描述Next, we use the Best Subset Selection, Forward Selection and Backward Selection to select which model is “true” model by evaluating the Bayesian information criterion (BIC) score. The BIC penalizes the model more for its complexity. The more complex models would have a larger score, meaning the worse model. Therefore, using BIC make us follow the Occam's razor "The simplest explanation is usually the best one.”. Importantly, the BIC under the Bayesian probability framework means that we would have a higher probability which BIC will select the true model based on the prior probability of candidate model includes the true model.

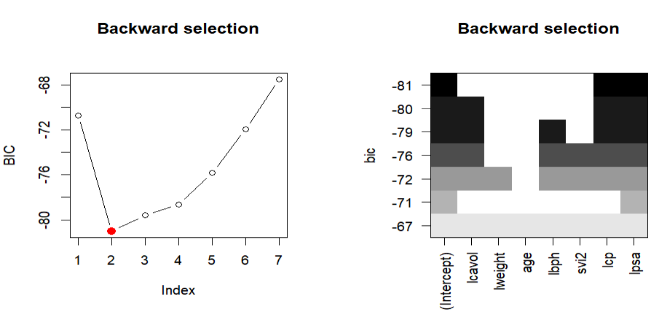
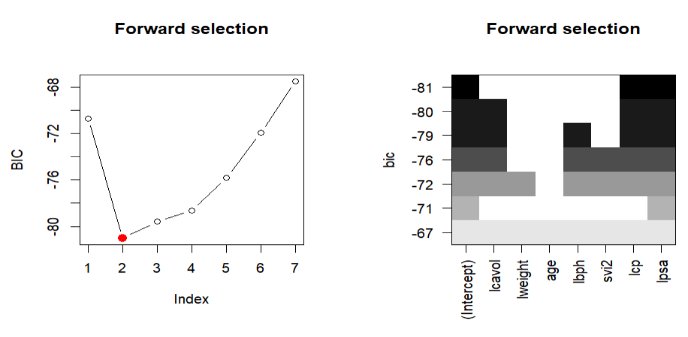
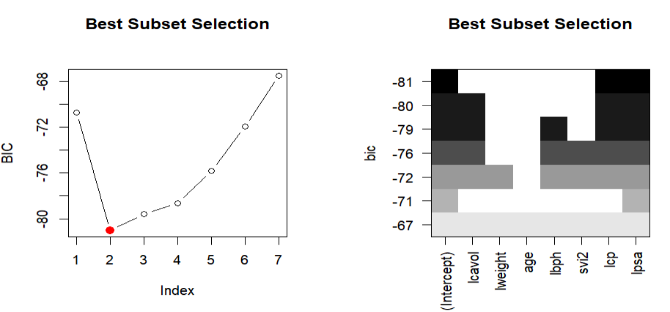
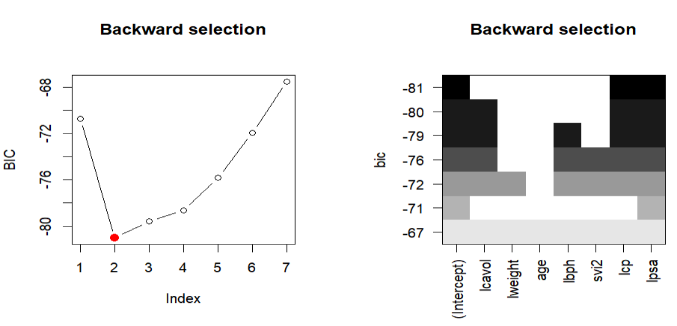
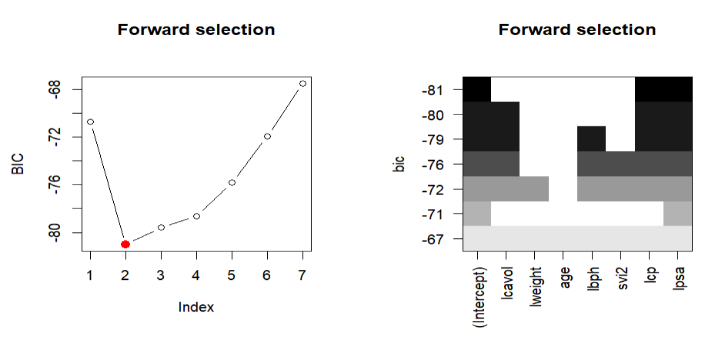
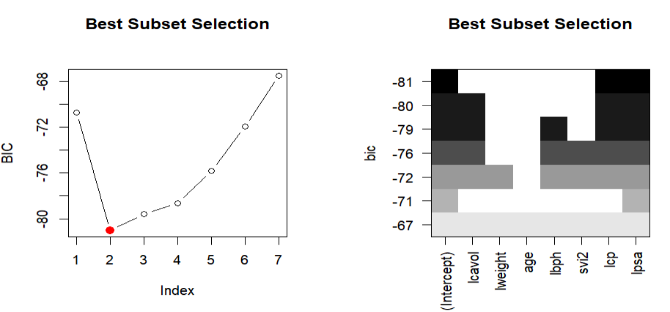


Figure 5. Best Subset Selection, Forward Selection and Backward Selection.

一張含有 文字 的圖片

自動產生的描述

The three methods all choose variables lcp and lpsa as the final model. Therefore, the formula of the final model is . This result is as we expected. As the level of invasion or migration of cancer cell into the capsule is getting higher, so does Cscore (the latter stage of the progression = more severe). PSA is a protein which is produced by normal as well as malignant cells of the prostate gland. Most men without prostate cancer have lower PSA levels of blood. When prostate cancer develops, the PSA level often goes up. As a result, increasing one unit of lcp would increase 8.5 of mean of the response variable changes while holding other variables constant; Furthermore, increasing one level of lpsa would increase 21.2 of mean of response variable while holding other variables constant.

1. **Make an appropriate LASSO model, with the appropriate link and error function, and evaluate the prediction performance. Do you see evidence that over-learning is an issue?**

一張含有 文字, 信 的圖片

自動產生的描述Lasso regression can be used for feature selection by adding a more penalty (controlled by tuning the parameter 𝜆) to the model with less important features (high variance).

The right plot (Figure. 6). above we can see when shrinkage penalty λ = 0.7 results in the smallest cross-validation error 750. Larger λ would instead introduce bias and increase the mean squared error (MSE). Lasso regression can avoid overfitting by shrinking the beta coefficients to zero (reduce flexibility) thereby introducing bias and reducing variance. The evidence of no overfitting issue can be seen in test MSE (in the code). We use 2/3 of the data as training data, and the remaining as test data and utilize 10-fold cross validation (cv.glmnet() code default) to choose the best λ which minimize the training MSE. Next is to evaluate the test MSE of prediction and compare to the test MSE of linear regression. The linear regression also uses training data to fit model and calculate test MES of prediction. The result suggests that lasso regression is more robust, and the variance of lasso model is smaller than linear model as we expected, because λ of lasso regression reduces the variance.

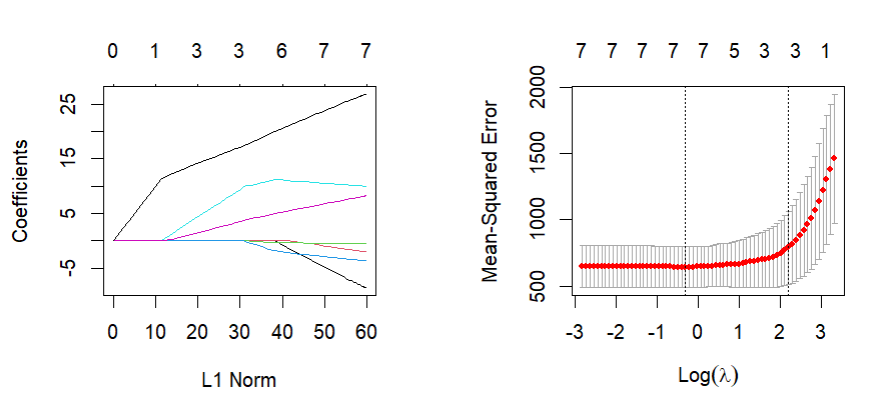
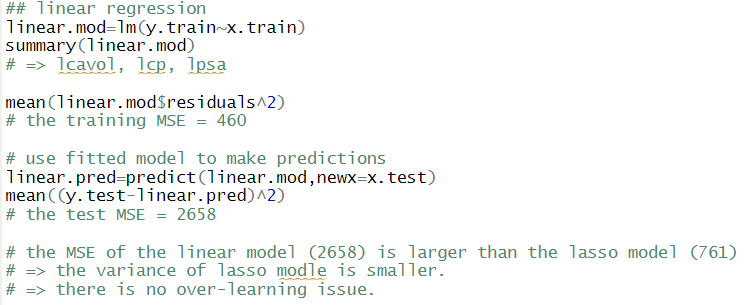
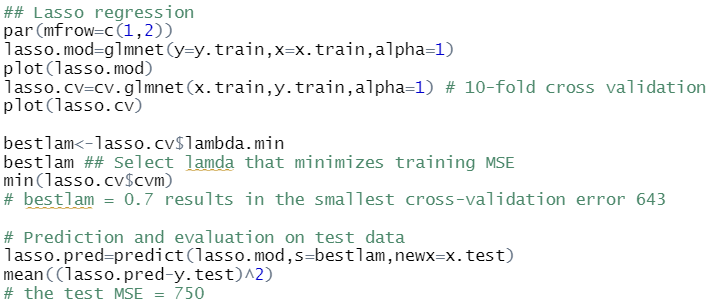
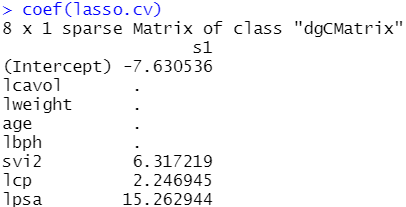


Figure 6. coefficient and MSE of lasso regression.

Left plot shows the change of coefficient of various parameters against different λ value. The lasso regression shrinks the beta coefficients of less important parameters exactly to zero as λ increased. Right plot shows training MSE against log (λ). Vertical dashed lines mean the minimum MSE and 1 SE from the minimum MSE which is based on 1 standard error rule (within one standard error of the minimum value), that λ1SE is a value that is a bit less prone to overfitting.



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



The coefficient of “lcavol” is exact zero. The strategy of lasso is using regularization to keep some important features of the model and exclude silly parameters (such as highly correlation independent variables) from the model. In Figure. 3, we know the lcavol and lpsa are highly correlated with each other, so that lcavol and lpsa can substitute each other in predicting Y. More to say that if there is a group of variables which correlations are very high, then the lasso tends to select only one variable from the group at random. Therefore, the lasso is not a good way to deal with highly correlated independent variables if they are all “true” parameters. Ridge and lasso are both regularization methods, but they are designed for different purposes. Ridge regression shrink coefficient close to zero which can still keep these “true” parameters.

1. **Fit your best model with *appropriate* non-linear effects. Report a comparison of performance to LASSO and your model reported under question 2. Explain what you find and indicate relevant issues or limitations of your analysis.**