Problem 4 (P, 20 Points)

1. (2P) Read and normalize the data: use read.table() to load the data; column 9 is the output lpsa for the regression and column 10 determines whether this data entry belongs to the training set. Column 1 is just an index and should not be used for prediction. Normalize each input feature to a mean of 0 and a variance of 1. Split up the data set into training and test set respectively. Useful functions: mean(), sd(), and the MASS library.

   Solution. Refer to section 4.1 in the code. ■

2. (4P) Compare LOOCV, 5-fold and 10-fold cross validation on the training data set to estimate the test error of using linear regression to predict lpsa from all other features. Use the full training data set to train a linear regression model and compute the test error. Compare your estimates obtained from cross validation to the error obtained from the test set and argue about your findings. Which of the methods is (theoretically) fastest?

   Solution. Refer to section 4.2 in the code.

   The test error for LOOCV is **0.583**, the corresponding error for 5-fold CV and 10-fold CV is **0.658** and **0.598** respectively. Finally, the full training set has a test error of **0.521**

   From the results, full train-test linear regression model gives the best result on the test set. Hence better model compared to the rest. The CV methods seem to have low accuracy since the amount of data samples are less. Therefore, CV methods would reduce the amount of data available to for training and hurt the model performance.

   Theoretically, the fastest would be the method with just train and test split without any cross validation. ■

3. (3P) Use the training set to fit ridge regression models and generate a plot showing the values of the coefficients in relation to the parameter λ (cf. Figure 6.4, p. 216, ISLR). What can you observe?

   Solution. Refer to section 4.3 in the code.

   Figure1 represents the plot of coefficients of ridge regression model and λ. From the plot we can infer that the when $\log(\lambda)$ is approximately equal to 5 all the coefficients are diminish to zero. Upon relaxing the λ, the coeffs increase in a smooth manner until close to zero, from where the coeffs explode in unregularized manner. ■

4. (3P) Perform 10-fold cross-validation on the training set to determine the optimal value for λ for the ridge regression model. Report train and test error measured in MSE for this λ.

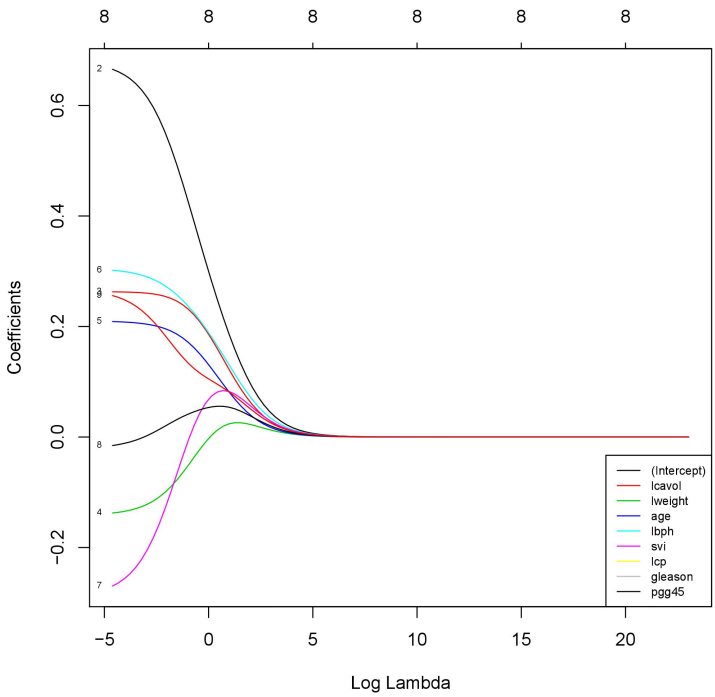

Figure 1: Ridge Regression Coeff vs λ

Solution. Refer to section 4.4 in the code.

The best λ for the ridge regression model is **0.087**. The corresponding train and test error are **0.447** and **0.494** respectively. ■

5. (3P) Use the training set to fit lasso models and generate a plot showing the values of the coefficients in relation to the parameter λ (cf. Figure 6.6, p. 220, ISLR). What can you observe in comparison to the plot generated in 3.?

   Solution. Refer to section 4.5 in the code.

   In the lasso regression model it is clear from the **Figure2**, the some of the coeffs (less significant) are made zero, while in the case of ridge regression, the are made smaller but not zero.

   In other words, ridge regression reduces the impact of irrelevant features but lasso regression is more strict by removing the irrelevant features. ■

6. (3P) Perform 10-fold cross-validation on the training set to determine the optimal value for λ in the lasso. Report train and test error measured in MSE for this λ. How many and which features are used? Compare this to the coefficients determined for ridge regression in 4.

   Solution. Refer to section 4.6 in the code.

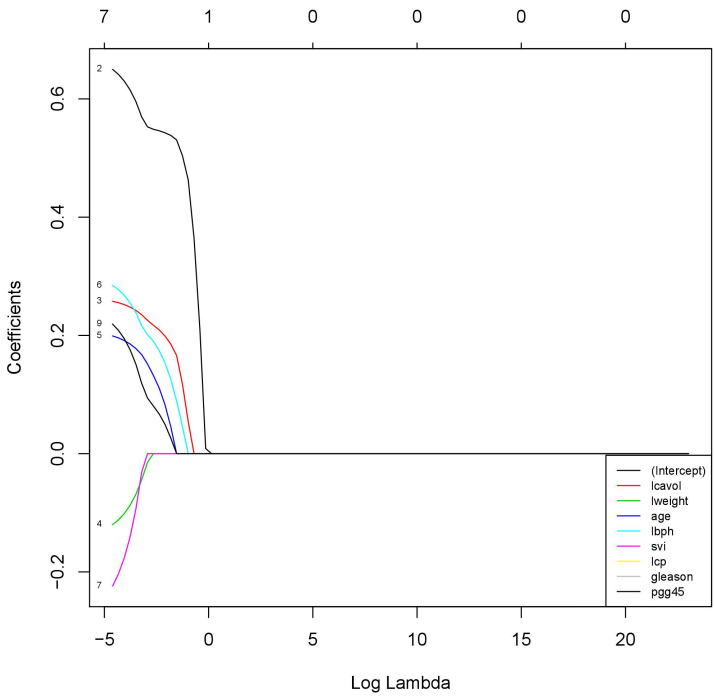

Figure 2: Lasso Regression Coeff vs λ

The best λ for the lasso regression model is **0.0092**. The corresponding train and test error are **0.440** and **0.499** respectively.

Totally seven features are used excluding the variable "gleason".

The **Table 1** show the comparison between ridge regression and lasso regression models. We can see that the values of the coeffs are almost equal except for the variable "gleason" which is very low in ridge regression and zero in lasso model, indicating that it is not a significant feature.

Lasso Regression Coeffs		Ridge Regression Coeffs	
Intercept	2.46	Intercept	2.46
lcavol	0.652	lcavol	0.58
lweight	0.258	lweight	0.257
age	-0.121	age	-0.110
lbph	0.199	lbph	0.2
svi	0.286	svi	0.281
lcp	-0.228	lcp	-0.163
gleason	0	gleason	0.012
pgg45	0.221	pgg45	0.199

Table 1: Coefficients of the Regression Models

■

7. . (2P) Compare the models generated in 4. and 6. to the model generated in 2. Which model would you choose and why? What alternative model could have been used?

   Solution. The Lasso and Ridge regression models have similar test error, but for the best lamda, the lasso model has less features hence a simpler model. Therefore the best model would be the Lasso model.

   Alternative model that could be used is the linear regression model with test-train split since it is simpler and has a very good accuracy for the given dataset. ■