MATH5470 Assigment 1

SUN Yuchang, 20034768R

1 Q1: ESL, Q3.6

The ridge regression estimate is:

$$\hat{\beta}^{\text{ridge}} = \arg\min_{\beta}\{\sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2\} \tag{1}$$

We can rewrite x with the intercept and y accordingly, then we have

$$\hat{\beta}^{\text{ridge}} = \arg\min_{\beta} \|y - X\beta\|^2 + \lambda \|\beta\|^2 \tag{2}$$

$$= \arg\min_{\beta}(y - X\beta)^T(y - X\beta) + \lambda \|\beta\|^2 \tag{3}$$

According to Bayesian Principle where

$$\text{Posterior probability} \propto \text{Prior} \times \text{Likelihood},$$

since we have the Gaussian sampling model $y \sim \mathcal{N}(X\beta, \sigma^2 I)$ and Gaussian prior $\beta \sim \mathcal{N}(0, \tau I)$, we can get

$$\begin{aligned} p(\beta|y, X) &= p(\beta)p(y|x, \beta) && (4)\\ &\propto p(\beta)p(y|x, \beta) && (5) \end{aligned}$$

$$\propto exp[-\frac{1}{2}(\beta - 0)\frac{1}{\tau}(\beta - 0)]exp[-\frac{1}{2}(y - \beta X)\frac{1}{\sigma}^2(y - \beta X)] \tag{6}$$

$$= exp[-\frac{1}{2\sigma^2}(y - \beta X)^T(y - \beta X) - \frac{1}{2\tau}||\beta||_2^2] \tag{7}$$

Using MAP estimate, we have

$$\hat{\beta} = \arg\max_{\beta} p(\beta|y, X) \tag{8}$$

$$= \arg\max_{\beta} exp[-\frac{1}{2\sigma^2}(y - \beta X)^T(y - \beta X) - \frac{1}{2\tau}||\beta||_2^2] \tag{9}$$

$$= \arg\min_{\beta}[\frac{1}{\sigma^2}(y - \beta X)^T(y - \beta X) + \frac{1}{\tau}||\beta||_2^2] \tag{10}$$

$$= \arg\min_{\beta}[(y - \beta X)^T(y - \beta X) + \frac{\sigma^2}{\tau}||\beta||_2^2] \tag{11}$$

By comparing equation (11) with equation (3), we show that the ridge regression estimate is the mean of the the posterior distribution; λ is $\frac{\sigma}{2\tau}$.

2 Q2: ESL, Q3.30

The elastic net estimate is

$$\hat{\beta} = \arg\min_{\beta} \|y - X\beta\|^2 + \lambda[\alpha \|\beta\|_2^2 + (1 - \alpha) \|\beta\|_1] \tag{12}$$

$$= \arg\min_{\beta} (y - X\beta)^T(y - X\beta) + \lambda[\alpha \|\beta\|_2^2 + (1 - \alpha) \|\beta\|_1] \tag{13}$$

$$= \arg\min_{\beta} y^T y - y^T X\beta - (X\beta)^T y + \beta^T X^T X\beta + \lambda\alpha\beta^T\beta + \lambda(1 - \alpha) \|\beta\|_1 \tag{14}$$

$$= \arg\min_{\beta} y^T y - y^T X\beta - (X\beta)^T y + \beta^T(X^T X + \lambda\alpha I)\beta + \lambda(1 - \alpha) \|\beta\|_1 \tag{15}$$

Let $\tilde{X} = [\frac{X}{\sqrt{\lambda\alpha I}}]$, then $\tilde{X}^T\tilde{X} = X^T X + \lambda\alpha I$; let $\tilde{y} = [\substack{y \\ 0}]$, then $\tilde{y}^T\tilde{y} = y^T y$. Thus $\left\|\tilde{y} - \tilde{X}\beta\right\|_2^2 =$ $y^T y - y^T X\beta - (X\beta)^T y + \beta^T(X^T X + \lambda\alpha I)\beta$.

Finally, we have

$$\hat{\beta} = \arg\min_{\beta}[\left\|\tilde{y} - \tilde{X}\beta\right\|_2^2 + \tilde{\lambda} \|\beta\|_1], \tag{16}$$

where $\tilde{X} = [\frac{X}{\sqrt{\lambda\alpha I}}]$, $\tilde{y} = [\substack{y \\ 0}]$, and $\tilde{\lambda} = \alpha(1 - \lambda)$.

3 Q3: ESL, Q3.17

I have applied multiple subset and shrinkage methods to analysis the spam data, including LS(least squares), best subset selection, ridge regression, Lasso regression, PCR(principle components regression) and PLS(partial least squares). For best subset selection, due to the large number of predictor variables, I only tried "k" feature selection, where k varies from 1 to 10. And the best case is 8 variables.
Table 1 summaries the test error and standard error of different methods.

Method	LS	Best Subset	Ridge	Lasso	PCR	PLS
Test Error	0.0049	0.0065	0.0048	0.0112	0.0100	0.0057
Std Error	0.0057	0.0102	0.0048	0.0031	0.0034	0.0111

Table 1: Results for different methods applied to the spam data.

The results are affected by the implement details. In my results, Ridge regression performs best with a test error of 0.0048 and a stand error of 0.0048.

4 Q4: ESL, Reproduce Prostate cancer example.

The relationships between the variables are shown as Fig.1. According to it, lcavol is more correlated

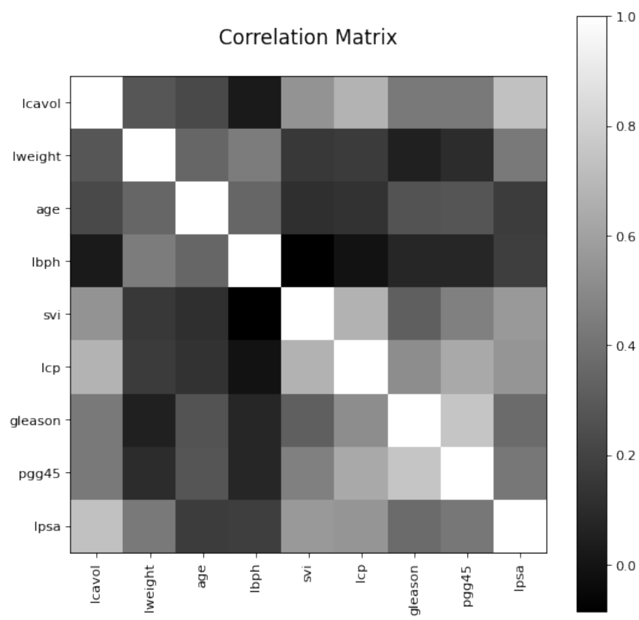

Figure 1: The correlation of variables

to lpsa.
I have applied multiple methods to analysis the prostate data, including LS, best subset selection, ridge regression($\alpha = 3.97$), Lasso regression, PCR and PLS.

Table 2 summaries the test error and standard error of different methods.

Method	LS	Best Subset	Ridge	Lasso	PCR	PLS
Intercept	2.4523	2.4523	2.4523	2.4683	2.4523	2.45234
lcavol	0.7164	0.7799	0.6273	0.5358	0.5706	0.4364
lweight	0.2926	0.3519	0.2878	0.1875	0.3233	0.3605
age	-0.1425		-0.1161		-0.1537	-0.0214
lbph	0.2120		0.2038		0.2160	0.2433
svi	0.3096		0.2891	0.0852	0.3221	0.2594
lcp	-0.2890		-0.1806		-0.0504	0.0858
gleason	-0.0209		0.0083		0.2286	0.0062
pgg45	0.2773		0.2162	0.0060	-0.0636	0.0843
Test Error	0.5213	0.4925	0.4969	0.4790	0.4483	0.5364
Std Error	0.1787	0.1431	0.1653	0.1645	0.1044	0.1493

Table 2: Results for different methods applied to the prostate data.

In the results, Lasso regression and PCR perform well. But PCR has a smaller standard error.