Assignment 1 for STAT7050

Songze Yang, u7192786

2022-08-04

The prostate cancer data set contains 9 variables. We are interested in the lpsa, which is the logarithm of prostate-specific antigen. We split the data into the training and testing data by the last colmun.

Question 1

Firstly, we apply OLS, ridge regression, the lasso, the naive elastic net and the elastic net to the data set.

```
OLS<-lm(lpsa~.,data = train)
OLS.pred<-predict(OLS,newdata = test)
mse.OLS<-mean((test$lpsa-OLS.pred)^2)
coef.OLS<-coef(OLS)
```

```
ridge <- glmnet (x, y, alpha = 0)
set.seed (1)
cv.out <- cv.glmnet (x, y, alpha = 0,type.measure = "mse")
bestlam.ridge <- cv.out$lambda.min
ridge.pred <- predict (ridge , s = bestlam.ridge ,
                       newx = x.test)
mse.ridge<-mean((test$lpsa-ridge.pred)^2)
coef.ridge<-coef(ridge, s = bestlam.ridge)
```

```
lasso <- glmnet (x, y, alpha = 1)
set.seed(1)
cv.out <- cv.glmnet (x, y, alpha = 1,type.measure = "mse")
bestlam.lasso <- cv.out$lambda.min
lasso.pred <- predict (lasso , s = bestlam.lasso,
                       newx = x.test)
mse.lasso<-mean((test$lpsa-lasso.pred)^2)
coef.lasso<-coef(lasso, s = bestlam.lasso)
```

```
grid.search<-seq(0,1,by = 0.001) #grid search algorithm
cv.result<-matrix(rep(0,length(grid.search)*2),length(grid.search),3)##matrix to store cv, lambda and alpha value.

for(i in 1:length(grid.search)){
  set.seed(1)
  cv.out <- cv.glmnet (x, y,alpha = grid.search[i])
  cv.result[i,1]<-cv.out$lambda.min
  cv.result[i,2]<-min(cv.out$cvm)
  cv.result[i,3]<-grid.search[i]
}

best.lam.elanet<-cv.result[,1][which.min(cv.result[,2])]
best.alpha.elanet<-cv.result[,3][which.min(cv.result[,2])]

elasticnet <- glmnet (x, y, alpha = best.alpha.elanet)## navie elastic net
elasticnet.pred <- predict (elasticnet , s = best.lam.elanet ,newx = x.test)
mse.elasticnet<-mean((test$lpsa-elasticnet.pred)^2)
coef.naive<-coef(elasticnet, s = best.lam.elanet)

lambda2<-best.lam.elanet*((1-best.alpha.elanet))##elastic net
coef.elanet<-(1+lambda2)*coef.naive

x.t.mat<-as.matrix(cbind(1,x.test))
els.pred<-x.t.mat%*%coef.elanet
mse.elsnet<-mean((test$lpsa-els.pred)^2)
```

The objective function in the glm package is shown as below.

$$argmin\frac{1}{2}|y-X\beta|^2+\lambda((1-\alpha)/2||\beta||_2^2+\alpha||\beta||_1)$$

So the coefficient of naive elastic net is calculated as following.

$$argmin|y-X\beta|^2+\lambda_2||\beta||_2^2+\lambda_1||\beta||_1)$$
$$where$$
$$\lambda_2=\lambda((1-\alpha)$$
$$and$$
$$\lambda_1=2\alpha$$

The coefficient estimate of elastic net is calculated as:

$$\theta_(elastic)=(1+\lambda_2)*\theta_(naive)$$

The result is summarized as below.

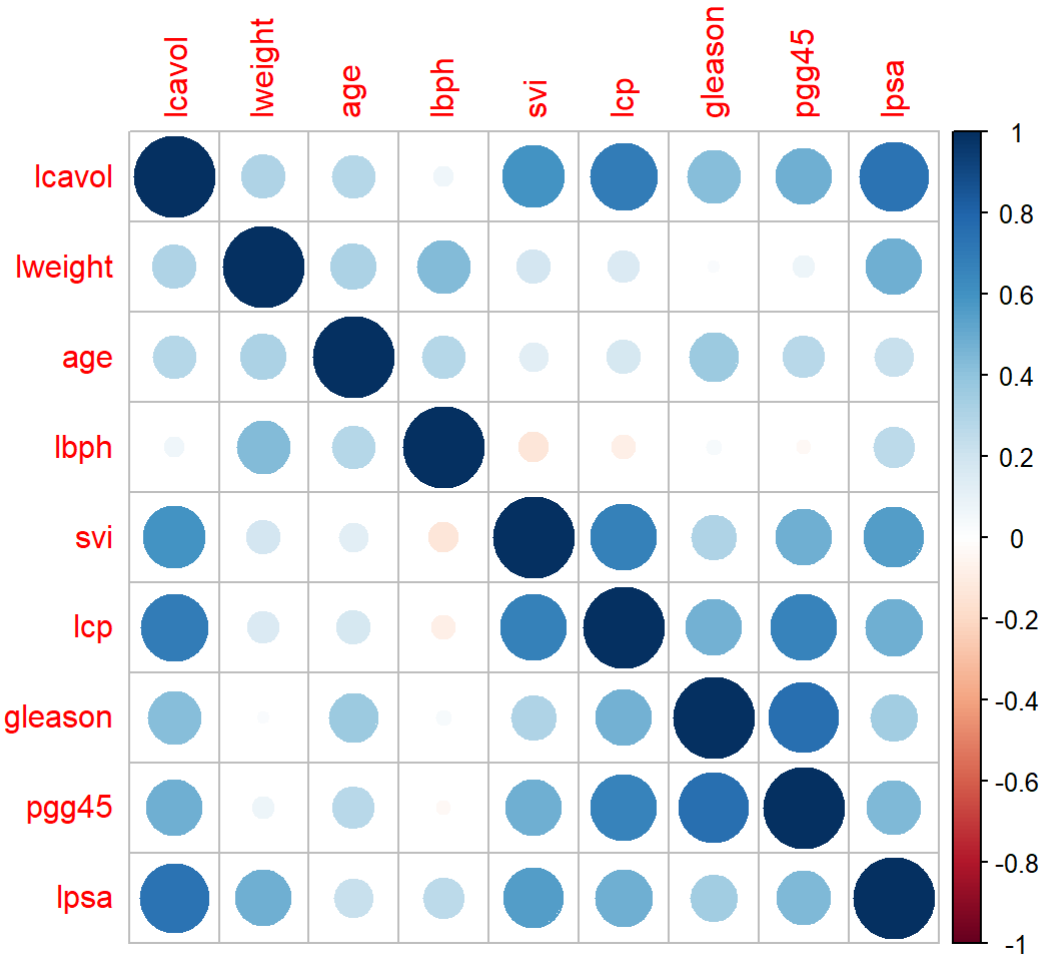
```
## Loading required package: rhandsontable
```

Method	Tuning.Parameters	Test.MSE	Variable.Selected
OLS		0.5212	ALL
Ridge Regression	lambda = 0.0878	0.4943	ALL
Lasso	lambda = 0.0076	0.5025	Exclude gleason
Naïve elastic net	lambda = 0.0580, alpha = 0.002	0.4997	ALL
Elastic net	lambda = 0.0580, alpha = 0.002	0.5091	ALL

Question 2

All of the models except the lasso include all the variables in the model. The lasso is the best model in terms of the sparsity. The lasso excludes the gleason, which has a high correlation to our pgg45 and low correlation to our response. The groups effect collects high correlated gleason and ppg45 variables and performs a good sparsity for lasso.

```
corrplot(cor(train))
```



In terms of the prediction accuracy, the best model is the ridge regression. It realizes a Test MSE of 0.4943, which is 3 percent above classical OLS. While the classical OLS is the worst predictive method, it gives us the unbiased efficient coefficients.

The ridge regression achieves lower Test MSE by performing a bias-variance trade-off, as followed.

$$Reducible\ error\ = x_i^T * Var(\theta) * x_i + x_i^T * Bias(\theta) * Bias(\theta)^T * x_i$$

The OLS method equals bias to 0 but remain the variance. So the ridge performs a selection between bias and variance in the model to achieve a lower MSE.

The naive elastic net combines the l-1 and l-2 penalty. It can be interpret as firstly a ridge type coefficient shrinkage and then a lasso type variable shrinkage. The ridge type shrinkage will only perform a bias and variance trade-off in the reducible error of the model. No variable will shrink to zero after. The lasso will also shrink the coefficient, however, the coefficients are allowed to minimize to zero under a l-1 penalty.

The two stage shrinkage for naive elastic net leads to the problem of over shrinkage. Therefore, a rescaled version of coefficient is formulated known as elastic net. The elastic net rescales the coefficient of its naive version by a factor of (1+).

However, in our case, every variable is very important in explaining our response. A lasso type shrinkage in elastic net deteriorates the predictive power. The naive elastic net over shrink our coefficient estimate, while the rescaled version remove this problem but also l-1 penalty in elastic net decay its prediction accuracy just like lasso.

Consequently, ridge achieves the best performance.

Question 3

Both naive elastic net and ridge in our case did not perform any variable selection. However, the coefficient for naive elastic net is slightly smaller than those of ridge. If we interpret the naive elastic net as method that firstly conducts the ridge type variable shrinkage and then apply a lasso type variable selection. Then the naive elastic net over shrinks the coefficient.

```
coef.naive;coef.ridge
```

```
## 10 x 1 sparse Matrix of class "dgCMatrix"
##                       s1
## (Intercept)  0.179484975
## (Intercept)  .
## lcavol       0.516497686
## lweight      0.606195008
## age         -0.016056303
## lbph         0.140243662
## svi          0.695924135
## lcp         -0.140665778
## gleason      0.005693304
## pgg45        0.007675482
```

```
## 10 x 1 sparse Matrix of class "dgCMatrix"
##                       s1
## (Intercept)  0.095549713
## (Intercept)  .
## lcavol       0.492656404
## lweight      0.601227708
## age         -0.014818243
## lbph         0.137965816
## svi          0.679288020
## lcp         -0.116652810
## gleason      0.017256035
## pgg45        0.007077847
```

The naive elastic net is like a stretchable fishing net that remain “all the big fish” (Zou and Hastie, 2005). Under a combination of l-1 and l-2 penalty, the naive elastic net keeps the gleason variable, which may be more important to include to our model. This is confirmed by the Training MSE by ridge(0.4473) and the naive elastic net(0.4434). The ridge model did a better job in extrapolation but naive elastic net achieve lower prediction accuracy in the training data.

However, in the data set that is large in number of covariates, the naive elastic net may perform better as it can present a variable selection to our data set. The ridge method may fail to select as under a l-2 penalty, no covariate can be shrinked to zero.

```
train.mse.ridge;train.mse.elasticnet
```

```
## [1] 0.4473332
```

```
## [1] 0.4434198
```

Question 4

While the classical OLS method gives us the unbiased efficient coefficient, the ridge regression achieves lower Test MSE by performing a bias-variance trade-off, as followed.

$$Reducible\ error\ = x_i^T * Var(\theta) * x_i + x_i^T * Bias(\theta) * Bias(\theta)^T * x_i$$

The OLS method equals bias to 0 but remain the variance. So the ridge performs a selection between bias and variance in the model to achieve a lower MSE. By liberating a certain of bais, a larger amount of variance can be decreased. Thus, a overall Test MSE will go down and the ridge overtakes the classical OLS.

Question 5

The lasso will also shrink the coefficient, however, the coefficients are allowed to minimize to zero under a l-1 penalty. The lasso model excludes the gleason. This may come from a grouping effect for gleason and pgg45. The correlation between the gleason and pgg45 is as high as 0.7570. By grouping effect, the lasso will selection only one variable from the two. The lasso losses information regarding the group of gleason and ppg45, where only one is picked.

However, the elastic net perform a trade off between excluding and including the variables under a l-1 after l-2 shrinkage. In our case, it is much finer to keep gleason variable than to exclude it. Therefore, the naive elastic net performs better in our data.

Further, under l-1 penalty, the estimate in lasso will bias towards to zero. It is better to refit the model using the selected variable to fit a lasso model, known as the relaxed lasso. This problem is mild in elastic net as the estimate is re-scaled by a factor.

Reference

Zou, H. and Hastie, T., 2005. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67(2), pp.301-320.