STA721 HW6

Zheng Yuan Evan Wyse

2019/10/30

Problem 1

(a)

First,the mle for $\hat{\beta}$ conditional on X is

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

The expected MSE for OLS under the full model can be written as

$$E[(\beta - \hat{\beta})^T (\beta - \hat{\beta})] = tr(IVar(\hat{\beta})) + (E(\beta - \hat{\beta}))^T E(\beta - \hat{\beta}) = tr(Var(\hat{\beta})) + (E(\beta - \hat{\beta}))^T E(\beta - \hat{\beta}),$$

and

$$E(\beta - \hat{\beta}) = \beta - E(\hat{\beta}) = \beta - (X^T X)^{-1} X^T X \beta = \beta - \beta = 0$$
$$Var(\hat{\beta}) = Var((X^T X)^{-1} X^T Y) = \sigma^2 (X^T X)^{-1} X^T X (X^T X)^{-1} = \sigma^2 (X^T X)^{-1}$$

so

$$E[(\beta - \hat{\beta})^T (\beta - \hat{\beta})] = tr(\sigma^2 (X^T X)^{-1}) = \sigma^2 \sum_{i=1}^{p} \frac{1}{\lambda_i},$$

here $\lambda_i' s$ are the eigenvalues of $X^T X$.

(b)

First, we calculate the average of observed MSEs,

```
## [1] 49.56834
```

In order to see whether the average of observed MSEs provides a good estimate of the expected MSE, we can also calculate the expected MSE analytically, that is

$$E[(\beta - \hat{\beta})^T (\beta - \hat{\beta})] = tr(\sigma^2 (X^T X)^{-1})$$

and compare with it.

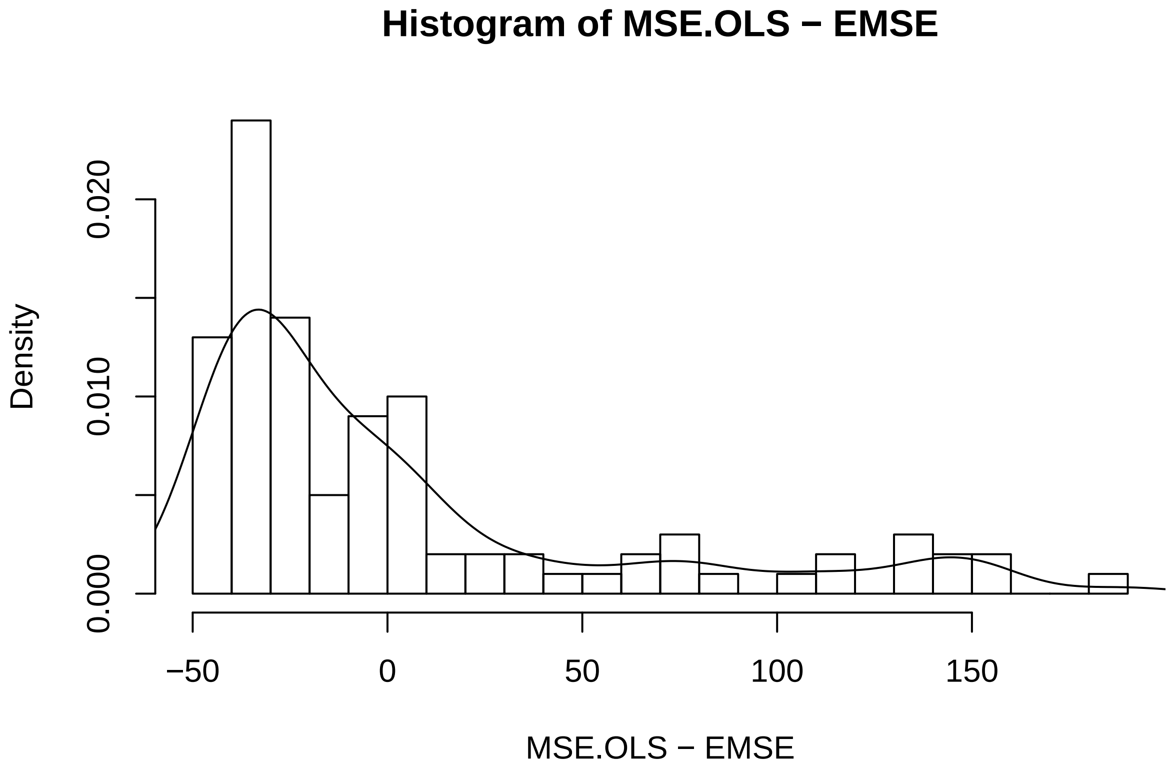
```
EMSE=(sigma^2)*sum(diag(solve(t(X)%*%X)))
EMSE
```

```
## [1] 45.42631
```

After comparison, there is still a unignorable difference between the average of observed MSEs and the expected MSE, so it does not provide a good estimate actually.

The distribution of observed MSE minus expected MSEs:

```r
hist(MSE.OLS-EMSE, freq = F, breaks=20);lines(density(MSE.OLS-EMSE))
```

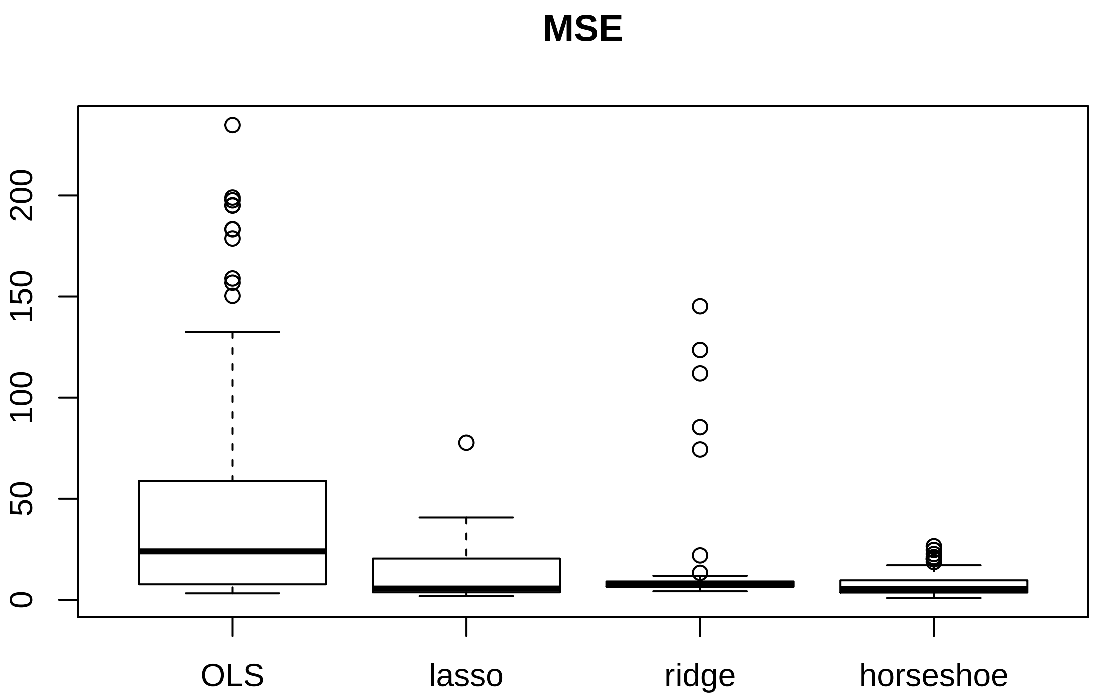


As we can see, the distribution of observed MSE minus expected MSEs looks like a normal distribution centering around its mean, which is around -35. In this case, I don't think increasing the number of simulated date sets would help, because the distribution of observed MSE minus expected MSEs does not center around 0, by the Central Limit Theorem, the average of observed MSEs must not converge to expected MSE in the end. Therefore, in fact, the more simulated date sets we have, the more difference between the average and the expected value there will be, until the difference is stable around a value.

(c)

```r
library(lars)
library(MASS)
library(matrixStats)
library(monomvn)

OLS.MSE = rep(0,100)
lasso.MSE = rep(0,100)
ridge.MSE = rep(0,100)
bhs.MSE = rep(0,100)

Coef.OLS = matrix(rep(0,2100),nrow=100) ## record OLS estimates of beta for each simulation
Coef.lasso = matrix(rep(0,2000),nrow=100)## record lasso estimates of beta for each simulation
Coef.ridge = matrix(rep(0,2100),nrow=100)## record ridge estimates of beta for each simulation
Coef.bhs = matrix(rep(0,2100),nrow=100)## record bhs estimates of beta for each simulation
```

```
for( i in 1:100) {
  rm(df)
  load(fname[i])
  X = as.matrix(df[,-1]); Y = df[,1]
  X.scale = scale(X); n = length(Y); X.mean = colMeans(X); X.sd = colSds(X)
  coeftrue = betatrue
  coeftrue[1] = betatrue[1]+sum(betatrue[-1]*X.mean)
  coeftrue[-1] = betatrue[-1]*X.sd


  ## ols
  nk.ols = lm(Y ~ X)
  coef.ols = coef(nk.ols)
  OLS.MSE[i] = sum((betatrue - coef.ols)^2)
  Coef.OLS[i,1:21]=coef.ols

  ## lasso
  nk.lasso = lars(X, Y, type="lasso")
  coef.lasso = coef(nk.lasso, s=which.min(nk.lasso$Cp))
  lasso.MSE[i] = sum((betatrue[-1] - coef.lasso)^2)
  Coef.lasso[i,1:20]=coef.lasso

  ## ridge
  seq.lambda = seq(0,10,0.01)
  nk.ridge = lm.ridge(Y ~ X, lambda = seq.lambda)
  lambda = seq.lambda[which.min(nk.ridge$GCV)]
  nk.ridge = lm.ridge(Y ~ X, lambda = lambda)
  coef.ridge = coef(nk.ridge)
  ridge.MSE[i] = sum((betatrue - coef.ridge)^2)
  Coef.ridge[i,1:21]=coef.ridge

  ## horseshoe
  nk.bhs = blasso(X, Y, case="hs", RJ=FALSE, normalize=F, verb=0)
  coef.bhs = c(mean(nk.bhs$mu), apply(nk.bhs$beta, 2, mean))
  bhs.MSE[i] = sum((betatrue - coef.bhs)^2)
  Coef.bhs[i,1:21]=coef.bhs

}
## Observed MSE for each simulation
boxplot(OLS.MSE, lasso.MSE, ridge.MSE, bhs.MSE, names=c("OLS", "lasso", "ridge", "horseshoe"), main="MS
```



```
## eatimation of squared bias and variance for each estimation
bias2.OLS=sum((colMeans(Coef.OLS)-betatrue)^2)
var.OLS=sum((Coef.OLS-colMeans(Coef.OLS))^2)/nsim

bias2.lasso=sum((colMeans(Coef.lasso)-betatrue[-1])^2)
var.lasso=sum((Coef.lasso-colMeans(Coef.lasso))^2)/nsim

bias2.ridge=sum((colMeans(Coef.ridge)-betatrue)^2)
var.ridge=sum((Coef.ridge-colMeans(Coef.ridge))^2)/nsim

bias2.bhs=sum((colMeans(Coef.bhs)-betatrue)^2)
var.bhs=sum((Coef.bhs-colMeans(Coef.bhs))^2)/nsim

bias2.OLS
```

```
## [1] 0.06188577
```

```
var.OLS
```

```
## [1] 126.7882
```

```
bias2.lasso
```

```
## [1] 1.450675
```

```
var.lasso
```

```
## [1] 43.83271
```

```
bias2.ridge
```

```
## [1] 2.441856
```

```
var.ridge
```

```
## [1] 67.97367
```

```
bias2.bhs
```

```
## [1] 2.693687
```

```
var.bhs
```

```
## [1] 62.22825
```

For OLS: $\hat{bias}^2 = 0.06$, $v\hat{a}r = 126.79$
For lasso: $\hat{bias}^2 = 1.45$, $v\hat{a}r = 43.82$
For Ridge: $\hat{bias}^2 = 2.44$, $v\hat{a}r = 67.97$
For horseshoe: $\hat{bias}^2 = 2.80$, $v\hat{a}r = 62.15$

Problem 2

Ridge prior

$p_\lambda(|\theta|) = \lambda|\theta|^2$

$p'_\lambda(|\theta|) = 2\lambda|\theta|$

1. Unbiasedness: No. $p'_\lambda(|\theta|) = 2\lambda|\theta|$ is not 0 when $|\theta|$ is large. So ridge is not unbiased.
2. Sparsity: No. min $|\theta| + p'_\lambda(|\theta|) = 0$. So it does not have sparsity.
3. Continuity: Yes. min $|\theta| + p'_\lambda(|\theta|)$ is attained at 0. So it have continuity.

Lasso prior

$p_\lambda(|\theta|) = \lambda|\theta|$

$p'_\lambda(|\theta|) = 2\lambda$

1. Unbiasedness: No. $p'_\lambda(|\theta|) = 2\lambda$ is not 0 when $|\theta|$ is large. So ridge is not unbiased.
2. Sparsity: Yes. min $|\theta| + p'_\lambda(|\theta|) = |\theta| + 2\lambda > 0$. So it has sparsity.
3. Continuity: Yes. the minimum is attained at 0. So it have continuity.

Cauchy prior

1. Unbiasedness: Yes.
2. Sparsity: Yes, handling Sparsity.
3. Continuity: No, does not hold.

The horseshoe penalty function $p_\lambda(\theta_i) = -\log\log(1 + \frac{2\tau^2}{\theta_i^2})$.

$p'_\lambda(\theta_i) = \frac{4\tau^2/|\theta_i|^3}{(1+2\tau^2/|\theta_i|^2)\log(1+2\tau^2/|\theta_i|^2)}$. Then unbiasedness follows.

For $\theta \neq 0$, $|\theta_i| + p'_\lambda(\theta_i) = |\theta_i| + \frac{4\tau^2/|\theta_i|^3}{(1+2\tau^2/|\theta_i|^2)\log(1+2\tau^2/|\theta_i|^2)}$ is strictly larger than 0. So it hanles Sparsity.

Since $|\theta_i| + p'_\lambda(\theta_i) \to \infty$, as $|\theta| \to 0$. So continuity does not hold.

Generalized Double Pareto prior

The penalty function is $p_\lambda(|\theta|) = (\alpha + 1)\log(|\theta| + \sigma\eta)$.

$p'_\lambda(|\theta|) = (\alpha + 1)\frac{1}{|\theta|}$ goes to 0 as $|\theta| \to \infty$. Then unbiasedness holds.

1. Unbiasedness: Yes.
2. Sparsity: Holds when $\eta < 2\sqrt{1 + \alpha}$
3. Continuity: Yes when $\eta <= \sqrt{1 + \alpha}$