STAT675 – Homework 2
Due: October 7

1. Derive the lasso, ridge, and least squares solutions under orthogonal design and contrast them.

   For each of the following we assume that orthogonal design: $\hat{\Sigma} = \mathbb{X}^\top \mathbb{X}/n$ has 1's on diagonal, and thus $\mathbb{X}^\top \mathbb{X} = nI$.

   least squares

$$\begin{aligned}
&\|y - Xb\|_2 = (y - Xb)^T (y - Xb) \\
&= (y^T - (Xb)^T)(y - Xb) \\
&= (y^T - b^T X^T)(y - Xb) \\
&= y^T y - y^T Xb - b^T X^T y + b^T X^T Xb \\
&= y^T y - 2(Xb)^T y + b^T bn
\end{aligned}$$

   and to minimize this expression above, we simply take the derivative with respect to b and set it equal to zero:

$$-2X^T y + 2bn = 0$$

   and this gives us that $\hat{b} = \frac{1}{n} X^T y$.

   ridge regression

   We can similarly find the ridge regression estimator under orthogonal design:

$$\begin{aligned}
&\|y - Xb\|_2 + \lambda \|b\| = (y - Xb)^T (y - Xb) + \lambda b^T b \\
&= y^T y - 2(Xb)^T y + b^T b(\lambda + n)
\end{aligned}$$

   and -once again- we simply take the derivative of this expression above with respect to b and set it equal to zero:

$$-2X^T y + 2b(\lambda + n) = 0$$

   and this gives us that $\hat{b} = \frac{1}{(\lambda+n)} X^T y$.

   lasso regression

   From the notes we want to minimize the following, component-wise, function:

$$L(\beta) = \beta^2/2 - \beta \hat{\beta}_{LS} + \lambda |\beta|$$

However, to solve this optimization problem (because of the absolute value) we will need to use the notion of *subderivative.* Keep in mind that the subderivative is equal to the derivative where a derivative exists, and where a derivative does not exist we appeal to the definition of the subderivative:

Definition 1. *We call c a subderivative of f at x_0 provided that*

$$f(x) - f(x_0) \geq c(x - x_0).$$

Thus -by definition- we find that this function has the following subdifferential (set of subderivatives):

$$\partial L_\beta = \begin{cases} \{\beta - \hat{\beta}_{LS} - \lambda\} & \beta < 0 \\ \{\beta - \hat{\beta}_{LS} + \lambda z : -1 \leq z \leq 1\} & \beta = 0 \\ \{\beta - \hat{\beta}_{LS} + \lambda\} & \beta > 0 \end{cases}$$

and we know that $\hat{\beta}_\lambda$ minimizes $L(\beta)$ iff $0 \in \partial L_\beta$. Using this condition we can easily find the following estimator:

$$\hat{\beta}_\lambda = \begin{cases} \hat{\beta}_{LS} + \lambda & \hat{\beta}_{LS} < -\lambda \\ 0 & -\lambda \leq \hat{\beta}_{LS} \leq \lambda \\ \hat{\beta}_{LS} - \lambda & \hat{\beta}_{LS} > \lambda \end{cases}$$

which can also be rewritten as $\hat{\beta}_\lambda = sign(\hat{\beta}_{LS})(|\hat{\beta}_{LS}| - \lambda)_+$

comparison of our estimators

By simply glancing at our three estimators, it is rather easy to see -under the orthogonal design- that the lasso and ridge estimators shrink our original least squares coefficients:

$$\hat{\beta}_{LS} = \frac{1}{n} X^T y$$

$$\hat{\beta}_{ridge,\lambda} = \frac{1}{(\lambda + n)} X^T y = \frac{n}{(\lambda + n)} \hat{\beta}_{LS}$$

$$\hat{\beta}_{lasso,\lambda} = \begin{cases} \hat{\beta}_{LS} + \lambda & \hat{\beta}_{LS} < -\lambda \\ 0 & -\lambda \leq \hat{\beta}_{LS} \leq \lambda \\ \hat{\beta}_{LS} - \lambda & \hat{\beta}_{LS} > \lambda \end{cases}$$

It is also rather easy to see that the only estimator which sends coefficients to zero is the lasso estimator, while the ridge coefficients will merely be shrunk closer to zero, yet never being completely zeroed out.

2. Let

$$\mathcal{T} = \left\{ \max_{1\le j\le p} 2|\epsilon^\top x_j|/n \le \lambda_0 \right\}$$

and assume $\mathbb{X}$ is standardized (that is, $\hat{\Sigma} = \mathbb{X}^\top \mathbb{X}/n$ has 1's on diagonal). Show that if $\lambda_0 = 2\sigma\sqrt{(t^2 + 2\log p)/n}$ and $\epsilon \sim N(0, \sigma^2 I)$, then

$$\mathbb{P}(\mathcal{T}) \ge 1 - 2e^{-t^2/2}$$

and hence that $\mathcal{T}$ is 'large'

$\because \hat{\Sigma} = \mathbb{X}^\top \mathbb{X}/n$ has 1's on diagonal.
$\therefore V_j = \epsilon^\top x_j / \left(\sqrt{n}\sigma\right)$ is $N(0,1)$ distributed, and

$$\begin{aligned}
&P\left(\max_{1\le j\le p} 2|\epsilon^\top x_j|/n > 2\sigma\sqrt{(t^2+2\log p)/n} \right) \\
&= P\left(\max_{1\le j\le p} |\epsilon^\top x_j| / \left(\sqrt{n}\sigma\right) > \sqrt{t^2 + 2\log p} \right) \\
&\le \sum_{j=1}^{p} P\left(|\epsilon^\top x_j / \left(\sqrt{n}\sigma\right)| > \sqrt{t^2+2\log p} \right) \\
&= 2pe^{-\frac{t^2+2\log p}{2}} \\
&= 2e^{-t^2/2}
\end{aligned}$$

$\therefore \mathbb{P}(\mathcal{T}) = 1 - \mathbb{P}(\mathcal{T}^c) \ge 1 - 2e^{-t^2/2}$.

3. Show that on $\mathcal{T}$ with $\lambda \ge 2\lambda_0$ and $S^* = \{j : \beta_j^* \ne 0\}$

$$2\left|\left|\mathbb{X}(\hat{\beta} - \beta^*)\right|\right|_2^2/n + \lambda \left|\left|\hat{\beta}_{S_*^c}\right|\right|_1 \le 3\lambda \left|\left|\hat{\beta}_{S_*} - \beta_{S_*}^*\right|\right|_1$$

By the reverse triangle inequality,

$$\begin{aligned}
&\left|\left|\hat{\beta}_{S_*} - \beta_{S_*}^*\right|\right|_1 \ge |\left|\left|\hat{\beta}_{S_*}\right|\right|_1 - ||\beta_{S_*}^*||_1| = ||\beta_{S_*}^*||_1 - \left|\left|\hat{\beta}_{S_*}\right|\right|_1 \\
&\Rightarrow \left|\left|\hat{\beta}_{S_*}\right|\right|_1 \ge ||\beta_{S_*}^*||_1 - \left|\left|\hat{\beta}_{S_*} - \beta_{S_*}^*\right|\right|_1 \\
&\Rightarrow \left|\left|\hat{\beta}\right|\right|_1 = \left|\left|\hat{\beta}_{S_*}\right|\right|_1 + \left|\left|\hat{\beta}_{S_*^c}\right|\right|_1 \ge ||\beta_{S_*}^*||_1 - \left|\left|\hat{\beta}_{S_*} - \beta_{S_*}^*\right|\right|_1 + \left|\left|\hat{\beta}_{S_*^c}\right|\right|_1 \\
&\Rightarrow 2\lambda \left|\left|\hat{\beta}_{S_*^c}\right|\right|_1 - 2\lambda \left|\left|\hat{\beta}\right|\right|_1 \le 2\lambda \left|\left|\hat{\beta}_{S_*} - \beta_{S_*}^*\right|\right|_1 - 2\lambda \, ||\beta_{S_*}^*||_1
\end{aligned}$$

We have the basic inequality

$$\left|\left|\mathbb{X}(\hat{\beta} - \beta^*)\right|\right|_2^2/n + \lambda \left|\left|\hat{\beta}\right|\right|_1 \le 2\epsilon^T \mathbb{X}(\hat{\beta} - \beta^*)/n + \lambda \, ||\beta^*||_1$$

On $\mathcal{T}$ with $\lambda \ge 2\lambda_0$,

$$2\epsilon^T \mathbb{X}(\hat{\beta} - \beta^*)/n \le \max_{1\le j\le p} 2|\epsilon^\top x_j|/n \left|\left|\hat{\beta} - \beta^*\right|\right|_1 \le \frac{\lambda}{2} \left|\left|\hat{\beta} - \beta^*\right|\right|_1$$

Therefore

$$\begin{aligned}
2\left|\left|\mathbb{X}(\hat{\beta}-\beta^*)\right|\right|_2^2/n+2\lambda\left|\left|\hat{\beta}\right|\right|_1 &\leq \lambda\left|\left|\hat{\beta}-\beta^*\right|\right|_1+2\lambda\left|\left|\beta^*\right|\right|_1 \\
\Rightarrow 2\left|\left|\mathbb{X}(\hat{\beta}-\beta^*)\right|\right|_2^2/n+2\lambda\left|\left|\hat{\beta}_{S_*^c}\right|\right|_1 &\leq \lambda\left|\left|\hat{\beta}-\beta^*\right|\right|_1+2\lambda(\left|\left|\beta^*\right|\right|_1-\left|\left|\beta^*_{S_*}\right|\right|_1)+2\lambda\left|\left|\hat{\beta}_{S_*}-\beta^*_{S_*}\right|\right|_1 \\
\Rightarrow 2\left|\left|\mathbb{X}(\hat{\beta}-\beta^*)\right|\right|_2^2/n+\lambda\left|\left|\hat{\beta}_{S_*^c}\right|\right|_1 &\leq \lambda\left|\left|\hat{\beta}-\beta^*\right|\right|_1+2\lambda\left|\left|\hat{\beta}_{S_*}-\beta^*_{S_*}\right|\right|_1-\lambda\left|\left|\hat{\beta}_{S_*^c}\right|\right|_1 \\
&= \lambda(\left|\left|\hat{\beta}-\beta^*\right|\right|_1-\left|\left|\hat{\beta}_{S_*^c}\right|\right|_1)+2\lambda\left|\left|\hat{\beta}_{S_*}-\beta^*_{S_*}\right|\right|_1 \\
&= 3\lambda\left|\left|\hat{\beta}_{S_*}-\beta^*_{S_*}\right|\right|_1
\end{aligned}$$

4. Suppose that the compatibility condition holds for S_* with constant ϕ_*. Then on $\mathcal{T}$ and for $\lambda \geq 2\lambda_0$

   $$n^{-1}\left|\left|\mathbb{X}(\hat{\beta}-\beta^*)\right|\right|_2^2+\lambda\left|\left|\hat{\beta}-\beta^*\right|\right|_1 \leq 4\lambda^2\frac{|S_*|}{\phi_*}$$

   Since the compatibility condition holds, and from 3, we have $\left|\left|\hat{\beta}_{S_*^c}\right|\right|_1 = \left|\left|\hat{\beta}_{S_*^c}-\beta^*_{S_*^c}\right|\right|_1 \leq 3\left|\left|\hat{\beta}_{S_*}-\beta^*_{S_*}\right|\right|_1$ (with large probability), then $\left|\left|\hat{\beta}_{S_*}-\beta^*_{S_*}\right|\right|_1^2 \leq \left((\hat{\beta}-\beta^*)^\top\hat{\Sigma}(\hat{\beta}-\beta^*)\right)\frac{|S_*|}{\phi_*} = \left|\left|\mathbb{X}(\hat{\beta}-\beta^*)\right|\right|_2^2\frac{|S_*|}{n\phi_*}$.

   $$\begin{aligned}
   &2\left|\left|\mathbb{X}\left(\hat{\beta}-\beta^*\right)\right|\right|_2^2/n+\lambda\left|\left|\hat{\beta}-\beta^*\right|\right|_1 \\
   &\leq 2\left|\left|\mathbb{X}\left(\hat{\beta}-\beta^*\right)\right|\right|_2^2/n+\lambda\left|\left|\hat{\beta}-\beta^*_{S_*}\right|\right|_1+\lambda\left|\left|\hat{\beta}_{S_*^c}\right|\right|_1 \\
   &\leq 4\lambda\left|\left|\hat{\beta}_{S_*}-\beta^*_{S_*}\right|\right|_1 && \because 3 \\
   &\leq 4\lambda\left|\left|\mathbb{X}\left(\hat{\beta}-\beta^*\right)\right|\right|_2\sqrt{|S_*|}/\left(\sqrt{n\phi_*}\right) && \because \text{compatibility condition} \\
   &\leq \left|\left|\mathbb{X}\left(\hat{\beta}-\beta^*\right)\right|\right|_2^2/n+4\lambda^2|S_*|/\phi_* && \because 4ab \leq a^2+4b^2
   \end{aligned}$$

   $\therefore \left|\left|\mathbb{X}\left(\hat{\beta}-\beta^*\right)\right|\right|_2^2/n+\lambda\left|\left|\hat{\beta}-\beta^*\right|\right|_1 \leq 4\lambda^2|S_*|/\phi_*$

5. Use HARNESS to infer relevant predictors in some regression data set.

   We used the prostate cancer data set from The Elements of Statistical Learning. The response variable was the log of prostate-specific antigen (lpsa). The possible predictors included log cancer volume (lcavol), log prostate weight (lweight), age, log of the amount of benign prostatic hyperplasia (lbph), seminal vesicle invasion (svi), log of capsular penetration (lcp), Gleason score (gleason), and percent of Gleason scores 4 or 5 (pgg45). All of these predictors were standardized before analysis.

   In the first step of HARNESS, we randomly split the data into a training set D_1 with 49 observations and a test set D_2 with 48 observations. We then used the Lasso to estimate the regression coefficients $\hat{\beta}$. These estimates are given below. The predictors with nonzero coefficients using Lasso turned out to be lcavol, lweight, lbph, svi and pgg45.

```
set.seed(31)
rand.i = sample(1:97)
D1 = my.data[rand.i[1:49],]
D2 = my.data[rand.i[50:97],]

x.train = as.matrix(D1[,-9])
y.train = D1[,9]
cvfit = cv.glmnet(x.train,y.train)
coef(cvfit, s="lambda.min")

## 9 x 1 sparse Matrix of class "dgCMatrix"
##                   1
## (Intercept) 2.50461
## lcavol      0.60689
## lweight     0.26878
## age         .
## lbph        0.11823
## svi         0.29684
## lcp         .
## gleason     .
## pgg45       0.07433
```

   The second step of HARNESS is to make inference using D_2. We estimated the predictive risk $R = E[(Y - \mathbb{X}^T\hat{\beta})^2|D_1]$ using $\hat{R} = \frac{1}{m}\sum_{i=1}^{m}\delta_i$ where the sum was over D_2 and $\delta_i = (Y_i - \mathbb{X}_i^T\hat{\beta})^2$. The 95 percent confidence interval for R was (0.23, 0.61).

```
x.test = as.matrix(D2[,-9])
y.test = D2[,9]
beta.hat = as.numeric(coef(cvfit, s="lambda.min"))
y.hat = cbind(rep(1,48),x.test)%*%beta.hat
delta = (y.test-y.hat)^2
(R.hat = mean(delta))

## [1] 0.4195
```

```
se.R = sd(delta)/sqrt(length(delta))
R.hat-qnorm(0.975)*se.R; R.hat+qnorm(0.975)*se.R

## [1] 0.2333
## [1] 0.6057
```

We also estimated the risk inflation by exluding predictor X_j:

$$R_j = E((Y - \mathbb{X}^T\hat{\beta}_{(j)})^2|D_1) - R$$

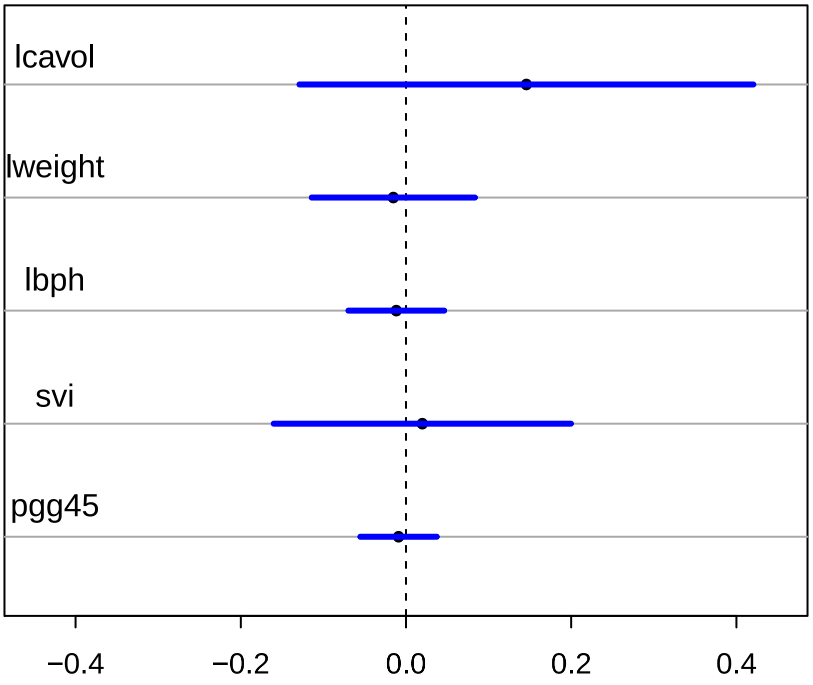
where $\hat{\beta}_{(j)}$ is equal to $\hat{\beta}$ except that $\hat{\beta}_j$ is set to zero.

```
S = which(beta.hat!=0)[-1]
e.list = list()

for(i in S){
  beta.j = beta.hat
  beta.j[i] = 0
  y.hat.j = cbind(rep(1,48),x.test)%*%beta.j
  e = list((y.test-y.hat.j)^2-delta)
  e.list=c(e.list,e)
}

Rj.hat = lapply(e.list, mean)
se.Rj = lapply(e.list, function(m){sd(m)/sqrt(length(m))})
```

The following plot shows the (Bonferroni-corrected) 95 percent confidence intervals for R_j.



Lastly, we used D_2 to estimate β^*, the least squares estimator over the set of predictors chosen by our model selection procedure from D_1.

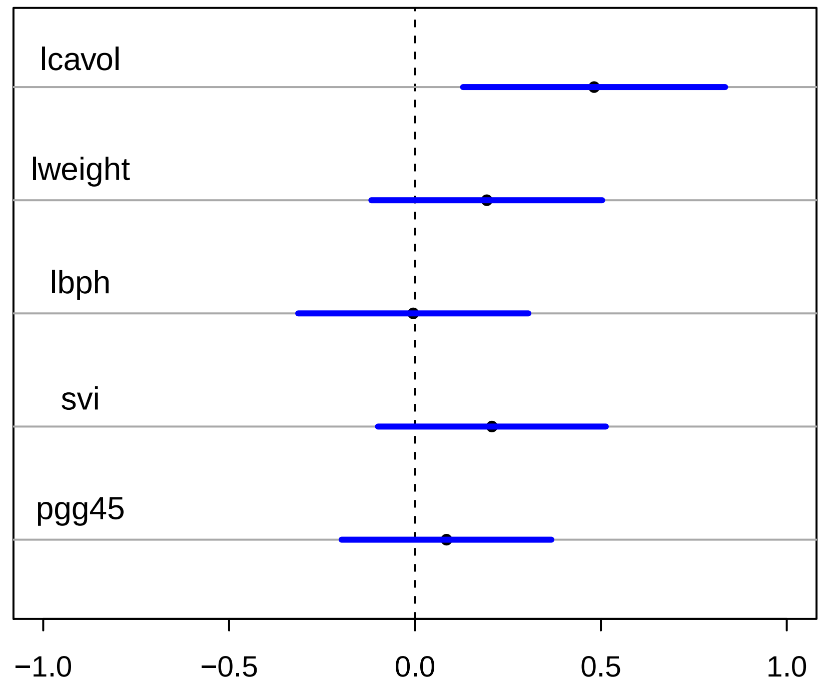
```
D2.S = D2[,c((S-1),9)]

mod.star = lm(lpsa˜., data=D2.S)
summary(mod.star)$coef[,c(1,2)]

##                Estimate Std. Error
## (Intercept)  2.531367     0.1014
## lcavol       0.481761     0.1369
## lweight      0.193082     0.1205
## lbph        -0.004577     0.1202
## svi          0.206807     0.1190
## pgg45        0.084727     0.1095

beta.star = summary(mod.star)$coef[,1][2:6]
se.beta.star = summary(mod.star)$coef[,2][2:6]
```

The following plot shows the (Bonferroni-corrected) 95 percent confidence intervals for β^*.


lcavol
lweight
lbph
svi
pgg45