BIOSTAT 274 Spring 2021 Homework 1

Due 11:59 PM 04/21/2020 (Submit to CCLE)

Zian ZHUANG

Theoretical Part

1.

(a) Write down the optimization problem of general linear model, ridge regression and LASSO in estimating β respectively.

Answer:

- As for general linear model, the optimization problem:

$$\hat{\beta} = \arg\min ||y - \beta||^2$$

- As for Ridge Regression, the optimization problem:

$$\hat{\beta}_\lambda^R = \arg\min ||y - \beta||^2 + \lambda||\beta||^2$$

In other words, we need to find

$$\begin{aligned} &\text{minimize}_\beta \sum_{i=1}^{n} (y_i - \beta)^2 \\ &\text{subject to } \beta^2 \leqslant s \end{aligned}$$

As for Lasso Regression, the optimization problem:

$$\hat{\beta}_\lambda^L = \arg\min ||y - \beta||^2 + \lambda||\beta||$$

In other words, we need to find

$$\begin{aligned} &\text{minimize}_\beta \sum_{i=1}^{n} (y_i - \beta)^2 \\ &\text{subject to } \beta \leqslant s \end{aligned}$$

(b) For fixed tuning parameter λ, solve for $\hat{\beta}$ (general linear model), $\hat{\beta}_\lambda^R$ (ridge regression) and $\hat{\beta}_\lambda^L$ (LASSO) respectively.

Answer: - As for general linear model, we took derivative the loss function with respect to β and set to zero, then we have

$$\begin{aligned}
0 =&(\sum_{i=1}^{n}(y_i - \hat{\beta})^2)^{'} \\
0 =&\sum_{i=1}^{n} -2 * y_i + 2\hat{\beta} \\
0 =&\sum_{i=1}^{n} -y_i + \hat{\beta} \\
n * \hat{\beta} =&\sum_{i=1}^{n} y_i \\
\hat{\beta} =&\frac{1}{n}\sum_{i=1}^{n} y_i = \bar{y_i}
\end{aligned}$$

- As for Ridge Regression, we took derivative the loss function with respect to β and set to zero (fixed tuning parameter λ), then we have

$$\begin{aligned}
0 =&((y - X * \hat{\beta}^R)^T(y - X * \hat{\beta}^R) + \lambda||\hat{\beta}^R||^2)^{'}, \text{(note that X is intercept 1)} \\
0 =&-2\sum_{i=1}^{n} y_i + 2n\hat{\beta}^R + 2n\lambda\hat{\beta}^R \\
(1+\lambda)n * \hat{\beta}^R =&\bar{y_i} \\
\hat{\beta}^R =&\frac{1}{(1+1\lambda)}\bar{y_i}
\end{aligned}$$

- As for Lasso Regression, firstly we have,

$$\begin{aligned}
(y - \hat{\beta}_\lambda^L)^2 =&(y - X\hat{\beta}_{ols} + X\hat{\beta}_{ols} - X\hat{\beta}_\lambda^L)^2, \text{(note that X is intercept 1)} \\
=&(y - X\hat{\beta}_{ols})^2 + 2 * (y - X\hat{\beta}_{ols})^T * (X\hat{\beta}_{ols} - X\hat{\beta}_\lambda^L) + (\hat{\beta}_{ols} - X\hat{\beta}_\lambda^L)^2 \\
=&(y - X\hat{\beta}_{ols})^2 + (X\hat{\beta}_{ols} - X\hat{\beta}_\lambda^L)^2
\end{aligned}$$

Since $(y - \hat{\beta}_{ols})^2$ is not a function of $\hat{\beta}_\lambda^L$, we only need to minimize $(\hat{\beta}_{ols} - \hat{\beta}_\lambda^L)^2 + \lambda||\hat{\beta}||$. We took derivative the loss function with respect to $\hat{\beta}_\lambda^L$ and set to zero (fixed tuning parameter λ) and plug in $\hat{\beta}_{ols} = \bar{y_i}$, then we have

$$\begin{aligned}
0 =&((X\hat{\beta}_{ols} - X\hat{\beta}_\lambda^L)^2 + \lambda||\hat{\beta}_\lambda^L||)^{'} \\
0 =&((\hat{\beta}_{ols} - \hat{\beta}_\lambda^L)^T(\hat{\beta}_{ols} - \hat{\beta}_\lambda^L) + \lambda||\hat{\beta}_\lambda^L||)^{'} \\
0 =&((\hat{\beta}_{ols} - \hat{\beta}_\lambda^L)^2 + \lambda||\hat{\beta}_\lambda^L||)^{'} \\
0 =&-2\hat{\beta}_{ols} + 2\hat{\beta}_\lambda^L + \lambda * \text{sign}(\hat{\beta}_\lambda^L) \\
2(\hat{\beta}_{ols} - \hat{\beta}_\lambda^L) =&\lambda * \text{sign}(\hat{\beta}_\lambda^L) \\
\hat{\beta}_\lambda^L =&\hat{\beta}_{ols} - \frac{\text{sign}(\hat{\beta}_\lambda^L)\lambda}{2}
\end{aligned}$$

Then we can have,

$$\hat{\beta}_\lambda^L = \begin{cases} \hat{\beta}_{ols} - \frac{\lambda}{2}, \hat{\beta}_{ols} \geqslant \frac{\lambda}{2} \\ 0, \hat{\beta}_{ols} \in (-\frac{\lambda}{2}, \frac{\lambda}{2}) \\ \hat{\beta}_{ols} + \frac{\lambda}{2}, \hat{\beta}_{ols} \leqslant -\frac{\lambda}{2} \end{cases} \Longrightarrow \begin{cases} \bar{y_i} - \frac{\lambda}{2}, \bar{y_i} \geqslant \frac{\lambda}{2} \\ 0, \bar{y_i} \in (-\frac{\lambda}{2}, \frac{\lambda}{2}) \\ \bar{y_i} + \frac{\lambda}{2}, \bar{y_i} \leqslant -\frac{\lambda}{2} \end{cases}$$

(c) Represent $\hat{\beta}_\lambda^R$ and $\hat{\beta}_\lambda^L$ by $\hat{\beta}$ and create plots of them separately for $\lambda = 1, 5, 10$. What can you tell?

Answer: we have

$$\hat{\beta}^R = \frac{1}{(1+1\lambda)}\hat{\beta}$$

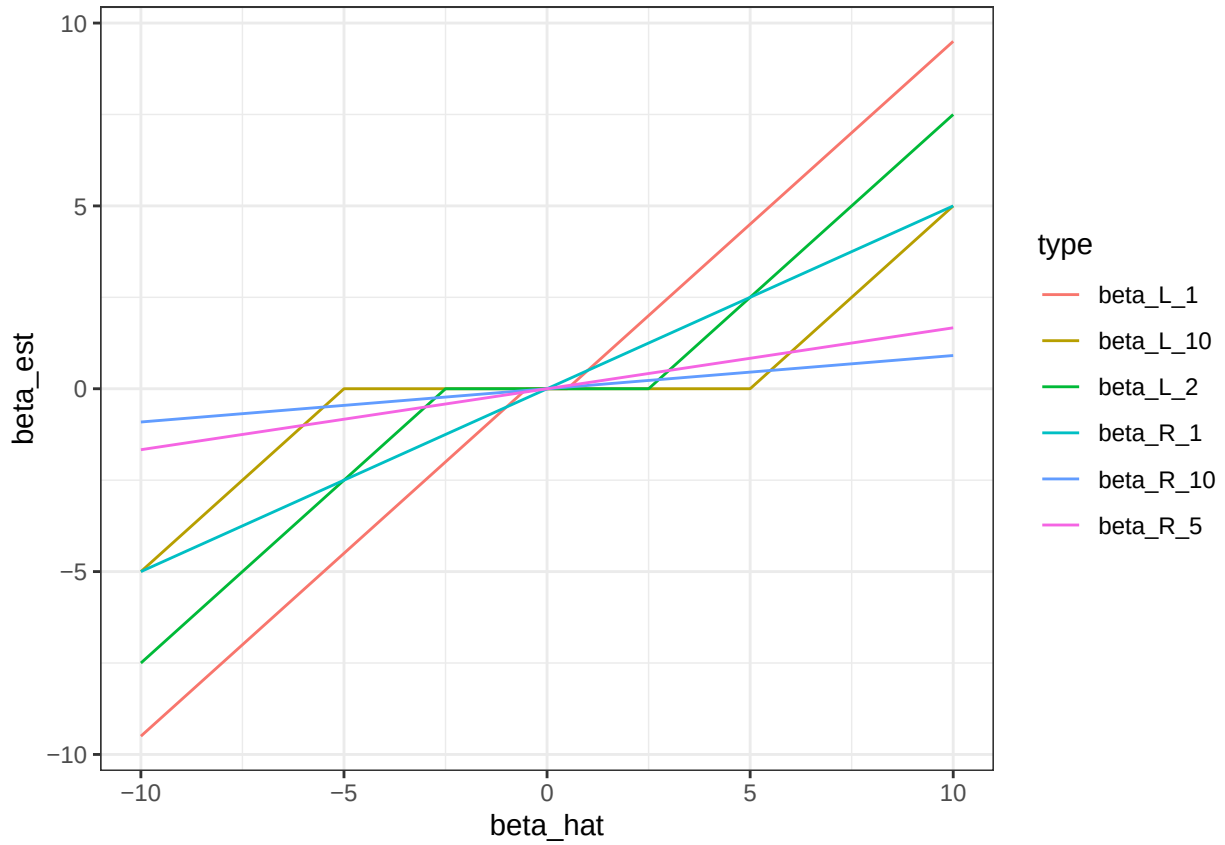
$$\hat{\beta}_\lambda^L = \begin{cases} \hat{\beta} - \frac{\lambda}{2}, \hat{\beta} \geqslant \frac{\lambda}{2} \\ \hat{\beta} + \frac{\lambda}{2}, \hat{\beta} < \frac{\lambda}{2} \end{cases}$$

```
# generate beta
set.seed(199609)
beta <- seq(-10,10,0.01)
beta_R_1 <- 1/(1+1)*beta
beta_R_5 <- 1/(1+5)*beta
beta_R_10 <- 1/(1+10)*beta

.trans <- function(x,lam){
  if(x>=lam/2){
  x-lam/2
}else if(x<=-lam/2){
  x+lam/2
}else{0}
}
beta_L_1 <- apply(beta %>% as.matrix,1,.trans,lam=1)
beta_L_2 <- apply(beta %>% as.matrix,1,.trans,lam=5)
beta_L_10 <- apply(beta %>% as.matrix,1,.trans,lam=10)

plots <- rbind(data.frame(beta_hat=beta, beta_est=beta_R_1, type="beta_R_1"),
      data.frame(beta_hat=beta, beta_est=beta_R_5, type="beta_R_5"),
      data.frame(beta_hat=beta, beta_est=beta_R_10, type="beta_R_10"),
      data.frame(beta_hat=beta, beta_est=beta_L_1, type="beta_L_1"),
      data.frame(beta_hat=beta, beta_est=beta_L_2, type="beta_L_2"),
      data.frame(beta_hat=beta, beta_est=beta_L_10, type="beta_L_10"))

ggplot(plots)+
  geom_line(aes(x=beta_hat,y=beta_est, group=type, color=type))+
  theme_bw()
```



We can tell from the plot that the a larger λ will shrink the estimated coefficient β closer to zero in both of Ridge and lasso regression. The difference is that the Ridge regression shrinks the coefficients towards 0, but lasso can shrink the coefficients to exact 0.

2.

(a) Write out the ridge regression optimization problem in this setting.

Answer:

Ridge regression optimization problem is to minimize,

$$(y_1 - \hat{\beta}_1 x_1 - \hat{\beta}_2 x_1)^2 + (y_2 - \hat{\beta}_1 x_2 - \hat{\beta}_2 x_2)^2 + \lambda(\hat{\beta}_1^2 + \hat{\beta}_2^2)$$

(b) Argue that the ridge coefficient estimates satisfy $\hat{\beta}^R_{\lambda 1} = \hat{\beta}^R_{\lambda 2}$

Answer: we took derivative the loss function with respect to $\hat{\beta}^R_{\lambda 1}$ and set to zero (fixed tuning parameter λ), then we have

$$\begin{aligned}
0 =& \frac{\partial}{\partial \hat{\beta}^R_{\lambda 1}}((y_1 - \hat{\beta}^R_{\lambda 1} x_1 - \hat{\beta}^R_{\lambda 2} x_1)^2 + (y_2 - \hat{\beta}^R_{\lambda 1} x_2 - \hat{\beta}^R_{\lambda 2} x_2)^2 + \lambda(\hat{\beta}^{R2}_{\lambda 1} + \hat{\beta}^{R2}_{\lambda 2})) \\
0 =& (2\hat{\beta}^R_{\lambda 1} x_1^2 - 2x_1 y_1 + 2\hat{\beta}^R_{\lambda 2} x_1^2) + (2\hat{\beta}^R_{\lambda 1} x_2^2 - 2x_2 y_2 + 2\hat{\beta}^R_{\lambda 2} x_2^2) + 2\lambda\hat{\beta}^R_{\lambda 1} \\
0 =& (\hat{\beta}^R_{\lambda 1} x_1^2 - x_1 y_1 + \hat{\beta}^R_{\lambda 2} x_1^2) + (\hat{\beta}^R_{\lambda 2} x_2^2 - x_2 y_2 + \hat{\beta}^R_{\lambda 2} x_2^2) + \lambda\hat{\beta}^R_{\lambda 1} \\
x_1 y_1 + x_2 y_2 =& \hat{\beta}^R_{\lambda 1}(x_1^2 + x_2^2) + \hat{\beta}^R_{\lambda 2}(x_1^2 + x_2^2) + \lambda\hat{\beta}^R_{\lambda 1} \\
x_1 y_1 + x_2 y_2 =& (\hat{\beta}^R_{\lambda 1} + \hat{\beta}^R_{\lambda 2})(x_1^2 + x_2^2) + \lambda\hat{\beta}^R_{\lambda 1}
\end{aligned}$$

Then we add $2x_1x_2\hat{\beta}^R_{\lambda 1}$ and $2x_1x_2\hat{\beta}^R_{\lambda 2}$ at both sides of the equation,

$$\begin{aligned}
x_1y_1 + x_2y_2 + 2\hat{\beta}^R_{\lambda 1}x_1x_2 + 2\hat{\beta}^R_{\lambda 2}x_1x_2 =& (\hat{\beta}^R_{\lambda 1} + \hat{\beta}^R_{\lambda 2})(x_1^2 + x_2^2 + 2x_1x_2) + \lambda\hat{\beta}^R_{\lambda 1} \\
x_1y_1 + x_2y_2 + 2\hat{\beta}^R_{\lambda 1}x_1x_2 + 2\hat{\beta}^R_{\lambda 2}x_1x_2 =& (\hat{\beta}^R_{\lambda 1} + \hat{\beta}^R_{\lambda 2})(x_1 + x_2)^2 + \lambda\hat{\beta}^R_{\lambda 1} \\
x_1y_1 + x_2y_2 + 2\hat{\beta}^R_{\lambda 1}x_1x_2 + 2\hat{\beta}^R_{\lambda 2}x_1x_2 =& \lambda\hat{\beta}^R_{\lambda 1}, (\text{Given that } x_1 + x_2 = 0)
\end{aligned}$$

Similarly, we took derivative the function with respect to $\hat{\beta}^R_{\lambda 2}$, then we have

$$\lambda\hat{\beta}^R_{\lambda 2} = x_1y_1 + x_2y_2 + 2\hat{\beta}^R_{\lambda 1}x_1x_2 + 2\hat{\beta}^R_{\lambda 2}2x_1x_2$$

Thus we know that,

$$\begin{aligned}
\lambda\hat{\beta}^R_{\lambda 1} =& x_1y_1 + x_2y_2 + 2\hat{\beta}^R_{\lambda 1}x_1x_2 + 2\hat{\beta}^R_{\lambda 2}x_1x_2 \\
=& \lambda\hat{\beta}^R_{\lambda 2} \\
\implies \hat{\beta}^R_{\lambda 1} =& \hat{\beta}^R_{\lambda 2} = \frac{x_1y_1 + x_2y_2}{\lambda - 4x_1x_2}
\end{aligned}$$

(c) Write out the LASSO optimization problem in this setting.

Answer:

LASSO regression optimization problem is to minimize,

$$(y_1 - \hat{\beta}^L_{\lambda 1}x_1 - \hat{\beta}^L_{\lambda 2}x_1)^2 + (y_2 - \hat{\beta}^L_{\lambda 1}x_2 - \hat{\beta}^L_{\lambda 2}x_2)^2 + \lambda(|\hat{\beta}^L_{\lambda 1}| + |\hat{\beta}^L_{\lambda 2}|)$$

(d) (Optional) Argue that in this setting, the LASSO coefficients $\hat{\beta}^L_{\lambda 1}$ and $\hat{\beta}^L_{\lambda 2}$ are not unique. Describe these solutions.

Answer:

Firstly, we took derivative the loss function with respect to $\hat{\beta}^L_{\lambda 1}$ and set to zero (fixed tuning parameter λ), then we have

$$\begin{aligned}
0 =& \frac{\partial}{\partial\hat{\beta}^L_{\lambda 1}}((y_1 - \hat{\beta}^L_{\lambda 1}x_1 - \hat{\beta}^L_{\lambda 2}x_1)^2 + (y_2 - \hat{\beta}^L_{\lambda 1}x_2 - \hat{\beta}^L_{\lambda 2}x_2)^2 \\
0 =& (2\hat{\beta}^L_{\lambda 1}x_1^2 - 2x_1y_1 + 2\hat{\beta}^L_{\lambda 2}x_1^2) + (2\hat{\beta}^L_{\lambda 1}x_2^2 - 2x_2y_2 + 2\hat{\beta}^L_{\lambda 2}x_2^2) \\
0 =& (\hat{\beta}^L_{\lambda 1}x_1^2 - x_1y_1 + \hat{\beta}^L_{\lambda 2}x_1^2) + (\hat{\beta}^L_{\lambda 2}x_2^2 - x_2y_2 + \hat{\beta}^L_{\lambda 2}x_2^2) \\
x_1y_1 + x_2y_2 =& \hat{\beta}^L_{\lambda 1}(x_1^2 + x_2^2) + \hat{\beta}^L_{\lambda 2}(x_1^2 + x_2^2) \\
x_1y_1 + x_2y_2 =& (\hat{\beta}^L_{\lambda 1} + \hat{\beta}^L_{\lambda 2})(x_1^2 + x_2^2) \\
x_1y_1 + x_2y_2 + 2\hat{\beta}^L_{\lambda 1}x_1x_2 + 2\hat{\beta}^L_{\lambda 2}x_1x_2 =& (\hat{\beta}^L_{\lambda 1} + \hat{\beta}^L_{\lambda 2})(x_1^2 + x_2^2 + 2x_1x_2) \\
x_1y_1 + x_2y_2 + 2\hat{\beta}^L_{\lambda 1}x_1x_2 + 2\hat{\beta}^L_{\lambda 2}x_1x_2 =& (\hat{\beta}^L_{\lambda 1} + \hat{\beta}^L_{\lambda 2})(x_1 + x_2)^2 \\
x_1y_1 + x_2y_2 + 2\hat{\beta}^L_{\lambda 1}x_1x_2 + 2\hat{\beta}^L_{\lambda 2}x_1x_2 =& 0, (\text{Given that } x_1 + x_2 = 0) \\
2(\hat{\beta}^L_{\lambda 1} + \hat{\beta}^L_{\lambda 2})x_1x_2 =& -(x_1y_1 + x_2y_2) \\
\hat{\beta}^L_{\lambda 1} + \hat{\beta}^L_{\lambda 2} =& -\frac{x_1y_1 + x_2y_2}{x_1x_2 * 2}
\end{aligned}$$

Since we know that $x_1 + x_2 = y_1 + y_2 = 0$, we have,

$$\begin{aligned}\hat{\beta}^L_{\lambda 1} + \hat{\beta}^L_{\lambda 2} &= -\frac{x_1 y_1 + x_2 y_2}{x_1 x_2 * 2} \\ &= -\frac{2x_1 y_1}{2x_1 x_2} \\ &= -\frac{y_1}{x_2} \\ &= \frac{y_1}{x_1}\end{aligned}$$

We need to find the point at which contour of the objective function $\frac{y_1}{x_1}$ touch the the constraint set. Since we know that the Lasso constraints for $\hat{\beta}^L_{\lambda 1}$ and $\hat{\beta}^L_{\lambda 2}$ is the area $|\hat{\beta}^L_{\lambda 1}| + |\hat{\beta}^L_{\lambda 2}| \leqslant s$ (s is a constant) and objective function is $\frac{y_1}{x_1}$, then we found that the objective function is parallel with the edge of the Lasso constraints area $|\hat{\beta}^L_{\lambda 1}| + |\hat{\beta}^L_{\lambda 2}| \leqslant s$. Thus, the intersections between the the objective function and the Lasso constraints area are more than one, i.e., all points on one edge of Lasso constraints area are able to optimize the objective function $\frac{y_1}{x_1}$.

Thus,the LASSO coefficients $\hat{\beta}^L_{\lambda 1}$ and $\hat{\beta}^L_{\lambda 2}$ are not unique. The possible values of $\hat{\beta}^L_{\lambda 1}$ and $\hat{\beta}^L_{\lambda 2}$ meet:

$$\hat{\beta}^L_{\lambda} = \begin{cases} \hat{\beta}^L_{\lambda 1} + \hat{\beta}^L_{\lambda 2} = s, \hat{\beta}^L_{\lambda 2} \geqslant 0, \hat{\beta}^L_{\lambda 2} \geqslant 0 \\ \hat{\beta}^L_{\lambda 1} + \hat{\beta}^L_{\lambda 2} = -s, \hat{\beta}^L_{\lambda 2} < 0, \hat{\beta}^L_{\lambda 2} < 0 \end{cases}$$