Chapter 6 : Regularization

6.1 Ridge

Seong yeon Park

March 19, 2025

The Three Sisters of Newton
School of Mathematics, Statistics and Data Science
Sungshin Women's University

Outline

We formulate the least squares method for multiple regression with matrices.

$$L := \sum_{i=1}^{N} (y_i - \beta_0 - \beta_1 x_{i,1}, \cdots, -\beta_1 x_{i,p})^2,$$

$$L = \| y - X\beta \|^2 = (y - X\beta)^T (y - X\beta)$$
$$= y^T y - y^T X\beta - \beta^T X^T y - \beta^T X^T X\beta$$

- Partial differentiation with L

$$\nabla L := \frac{\partial L}{\partial \beta} = -X^T y - X^T y + 2X^T X\beta = -2X^T (y - X\beta)$$

- Set to zero to find the minimum value

$$-2X^T (y - X\beta) = 0$$

Multiple Regression Singular/Nonsingular Matrix

- When a matrix X^TX is nonsingular, we have

$$2X^TX\beta = 2X^Ty$$
$$\hat{\beta} = (X^TX)^{-1}X^Ty$$

- If matrix X^TX is singular, Determinant is too small, β becomes large and an inconvenient situation occurs.

Ridge Regression

- $\lambda \geq 0$ be a constant, we often use to minimize the square error plus by the squared norm of β multiplied by λ.
- Loss function of existing linear regression (sum of error squared)

$$L = \frac{1}{N} \parallel y - X\beta \parallel^2$$

- Loss function of ridge regression

$$L := \frac{1}{N} \parallel y - X\beta \parallel^2 + \lambda \parallel \beta \parallel_2^2$$

- $\lambda \parallel \beta \parallel_2^2$ is the regularization term of ridge regression.
- The larger the λ, the smaller the β size.

is the square of the L2 norm of β.

Differentiate Loss function of Ridge Regression

- Loss function of ridge regression

$$
\begin{aligned}
L &:= \frac{1}{N} \parallel y - X\beta \parallel^2 + \lambda \parallel \beta \parallel_2^2 \\
&= \frac{1}{N}(y - X\beta)^T(y - X\beta) + \lambda\beta^T\beta \\
&= \frac{1}{N}(y^Ty - y^TX\beta - \beta^TX^Ty + \beta^TX^TX\beta) + \lambda\beta^T\beta
\end{aligned}
$$

- Differentiate L by β,

$$
\begin{aligned}
\frac{\partial L}{\partial \beta} &= \frac{1}{N}(-2\beta^TX^Ty + 2X^TX\beta + 2\lambda\beta) \\
&= -\frac{2}{N}X^T(y - X\beta) + 2\lambda\beta = 0 \\
&= -\frac{1}{N}X^T(y - X\beta) + \lambda\beta = 0 \\
&= \frac{1}{N}X^T(y - X\beta) = \lambda\beta
\end{aligned}
$$

- This additional term serves to control the size of β.

Calculate the Weight $\hat{\beta}$

- If $X^T X + \lambda I$ is nonsingular,

$$
\begin{aligned}
\frac{1}{N} X^T (y - X\beta) &= \lambda\beta \\
X^T (y - X\beta) &= N\lambda\beta \\
X^T y - X^T X\beta &= N\lambda\beta \\
X^T y &= X^T X\beta + N\lambda\beta \\
X^T y &= (X^T X + N\lambda I)\beta
\end{aligned}
$$

- $\hat{\beta} = (X^T X + N\lambda I)^{-1} X^T y$

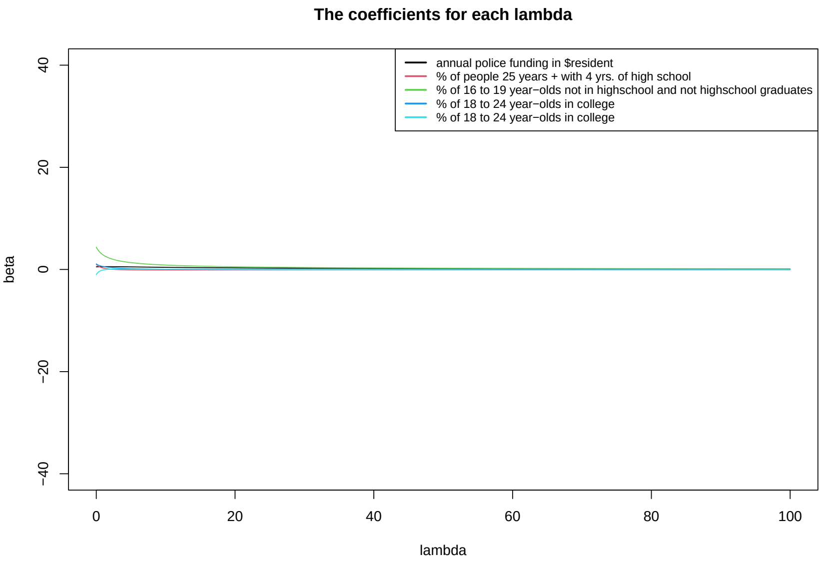
R Code for Ridge Regression

```r
ridge=function(X,y,lambda=0){
  X=as.matrix(X);p=ncol(X);n=length(y);X.bar=array(dim=p);s=array(dim=p)
  for (j in 1:p){X.bar[j]=mean(X[,j]);X[,j]=X[,j]-X.bar[j];};
  for (j in 1:p){s[j]=sd(X[,j]);X[,j]=X[,j]/s[j]};
  y.bar=mean(y);y=y-y.bar
  beta=drop(solve(t(x)%*%X+n*lambda*diag(p))%*%t(X)%*%y)
  for (j in 1:p)beta[j]=beta[j]/s[j]
  beta.0=y.bar-sum(X.bar*beta)
  return(list(beta=beta,beta.0=beta.0))
}
```

Example 48

```
df=read.table("crime.txt");x=df[,3:7];y=df[,1];p=ncol(x);
lambda.seq=seq(0,100,0.1);coef.seq=lambda.seq
plot(lambda.seq,coef.seq,xlim=c(0,100),ylim=c(-40,40),
     xlab="lambda",ylab="beta",main="The coefficients for each lambda",
     type="n",col="red")
for (j in 1:p){
  coef.seq=NULL;for(lambda in lambda.seq)coef.seq=c(coef.seq,
                                                        ridge(x,y,lambda)$beta[j])
  par(new=TRUE);lines(lambda.seq,coef.seq,col=j)
}
legend("topright",legend=
         c("annual police funding in $resident","% of people 25 years +
           with 4 yrs. of high school",
           "% of 16 to 19 year-olds not in highschool and not highschool
           graduates", "% of 18 to 24 year-olds in college",
           "% of 18 to 24 year-olds in college"),col=1:p,lwd=2,cex=.8)
```

Example 48


The coefficients for each lambda
annual police funding in $resident
% of people 25 years + with 4 yrs. of high school
% of 16 to 19 year−olds not in highschool and not highschool graduates
% of 18 to 24 year−olds in college
% of 18 to 24 year−olds in college
beta
lambda
40
20
0
−20
−40
0
20
40
60
80
100

Q & A

Thank You