Homework #2

Yongjin Shin
20090488
Industrial Management Engineering, POSTECH
dreimer@postech.ac.kr

March 21, 2018

Problem 1. Provide detailed description of ridge regression, including algorithm derivation. Implement your ridge regression to obtain the regularization path that was discussed in class.

Solution If the weights w_is are not constrained, they can explode and hence are susceptible to very high variance. To control variance, we might regularize the coefficients to impose the ridge constraint:

$$
\begin{aligned}
& minimize \quad \Sigma_{i=1}^{n}(y_i - w^T z_i)^2 \quad s.t. \quad \Sigma_{j=1}^{D} w_j^2 \leq B \\
\Leftrightarrow \quad & minimize \quad (y - Zw)^T(y - Zw) \quad s.t. \quad \Sigma_{j=1}^{D} w_j^2 \leq B
\end{aligned}
$$

where Z is assumed to be standardized (mean 0, unit variance) and y is assumed to be centered.
Then, the bounded constrained form is equivalent to loss function:

$$
\begin{aligned}
L(w)_{l_2} &= minmize \quad \frac{1}{2}||y - Zw||_2^2 + \frac{\lambda}{2}||w||_2^2 \\
&= \quad \frac{1}{2}(y - Zw)^T(y - Zw) + \frac{\lambda}{2}w^T w \\
&= \quad \frac{1}{2}(y^T y - y^T Zw - w^T Z^T y + w^T Z^T Zw) + \frac{\lambda}{2}w^T w
\end{aligned} \tag{1}
$$

Since the equation 1 is convex, and hence has a unique solution. Take derivatives, we can obtain:

$$
\begin{aligned}
\frac{\partial L(w)_{l_2}}{\partial w} &= \frac{\partial}{\partial w}\Big[\frac{1}{2}||y - Zw||_2^2 + \frac{\lambda}{2}||w||_2^2\Big] \\
&= \frac{\partial}{\partial w}\Big[\frac{1}{2}(y^T y - y^T Zw - w^T Z^T y + w^T Z^T Zw) + \frac{\lambda}{2}w^T w] \\
&= -Z^T y + Z^T Zw + \lambda w \\
&= -Z^T y + (Z^T Z + \lambda I)w
\end{aligned} \tag{2}
$$

Therefore, the solution is now seen to be:

$$\hat{w}_{\lambda}^{ridge} = (Z^T Z + \lambda I)^{-1} Z^T y \tag{3}$$

λ is the shrinkage parameter which controls the size of the coefficients. As λ goes to 0, we obtain the least squares solutions. On the other hand, as λ goes to $infinity$, we have $\hat{w}_{\lambda=\infty}^{ridge} = 0$ Notice that even if $Z^T Z$ is not invertible, inclusion of λ makes problem non-singular. Ridge regression can be implemented as following algorithm and the full script **Appendix A**:

Algorithm 1: Ridge Regression

input **:** Standardized Z (mean 0, unit variance), Centered y, λ
output: $\hat{w}_{lambda}^{ridge} = [\hat{w}_1, ..., \hat{w}_D]^T$, df_λ

1. $Compute \quad tmp = (Z^T Z + \lambda I)^{-1} Z^T;$
2. $\hat{w}_{\lambda}^{ridge} = tmp * y;$
3. $df_\lambda = trace(Z * tmp);$

A smoother matrix S is a linear operator satisfying: $\hat{y} = Sy$.The effective degrees of freedom (or effective number of parameters) is defined as:

$$df(S) = tr(S) \tag{4}$$

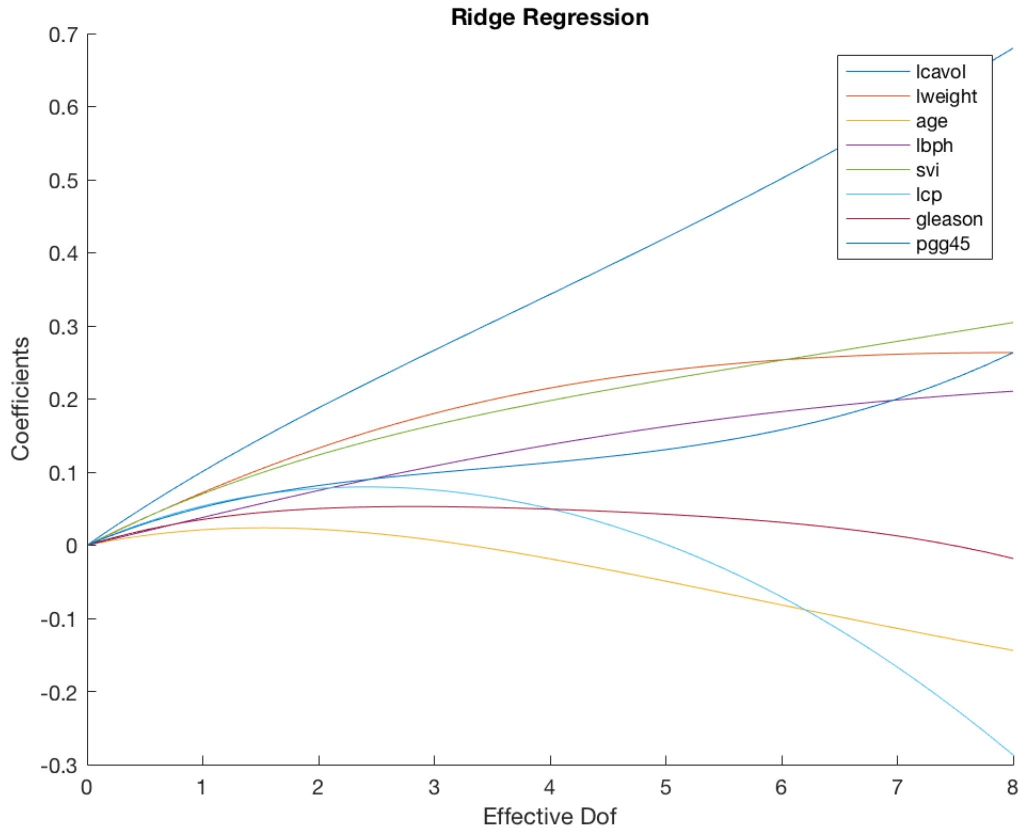
In ridge regression, the fits are given by:

$$\hat{y} = Z(Z^T Z + \lambda I)^{-1} Z^T y \tag{5}$$

So the effective degrees of freedom (effective number of parameters) in ridge regression are given by:

$$df(\lambda) = tr(S_\lambda) = tr\Big[Z(Z^T Z + \lambda I)^{-1} Z^T\Big] = \Sigma_{i=1}^{D} \frac{d_i^2}{d_i^2 + \lambda} \tag{6}$$

The result of ridge regression path can be given as the below.



Problem 2. Provide detailed description of LASSO regression, including the derivation of the coordinate descent algorithm. Implement your LASSO regression to obtain the regularization path that was discussed in class.

Solution LASSO coefficients are the solutions to the l_1 optimization problem:

$$minimize \quad (y - Zw)^T(y - Zw) \quad s.t. \quad \Sigma_{i=1}^{D}|w_i| \leq B \tag{7}$$

This is equivalent to loss function:

$$\begin{aligned} L(w)_{l_1} &= \Sigma_{i=1}^{n}(y_i - w^T z_i)^2 + \lambda \Sigma_{i=1}^{D}|w_i| \\ &= (y - Zw)^T(y - Zw) + \lambda ||w||_1 \end{aligned} \tag{8}$$

The coordinate descent will be used, which is to optimize the parameter one bye one tactic. Hence, Zw can be separated with two terms; one with dth column and the other without dth column.

$$\begin{aligned} L(w)_{l_1} &= \frac{1}{2}\Sigma_{i=1}^{n}(y_i - w^T z_i)^2 + \lambda \Sigma_{i=1}^{D}|w_i| \\ &= \frac{1}{2}\Sigma_{i=1}^{n}(y_i - w_d^T z_{i,d} - w_{-d}^T z_{i,-d})^2 + \lambda \Sigma_{j=1}^{D}|w_j| \end{aligned} \tag{9}$$

Now we only need to calculate the derivative of loss function with respect to weight w.

$$\begin{aligned}
\frac{\partial L(w)_{l_1}}{\partial w_d} &= \Sigma_{i=1}^{n}(y_i - w_d^T z_{i,d} - w_{-d}^T z_{i,-d})(-z_{i,d}) + \lambda \frac{\partial |w_d|}{\partial w_d} \\
&= w_d^T \left(\Sigma_{i=1}^{n} z_{i,d}^2\right) - \Sigma_{i=1}^{n}(y_i - w_{-d}^T z_{i,-d}) z_{i,d} + \lambda \frac{\partial |w_d|}{\partial w_d} \\
&= w_d^T \alpha_d + \beta_d + \lambda \frac{\partial |w_d|}{\partial w_d}
\end{aligned} \tag{10}$$

Computing $\frac{\partial ||w_d||_1}{\partial w_d}$ can be solved by sub-differentials.

$$\frac{\partial f}{\partial x} = \begin{cases} -1 & if \quad x < 0 \\ [-1, 1] & if \quad x = 0 \\ 1 & if \quad x > 0 \end{cases} \tag{11}$$

Thus, the estimate of w_d given the other parameters is calculated as:

$$\hat{w}_d^{lasso} = \begin{cases} \frac{\beta_d + \lambda}{\alpha_d} & if \quad \beta_d < -\lambda \\ 0 & if \quad \beta_d \in [-\lambda, \lambda] \\ \frac{\beta_d - \lambda}{\alpha_d} & if \quad \beta_d > \lambda \end{cases} \tag{12}$$

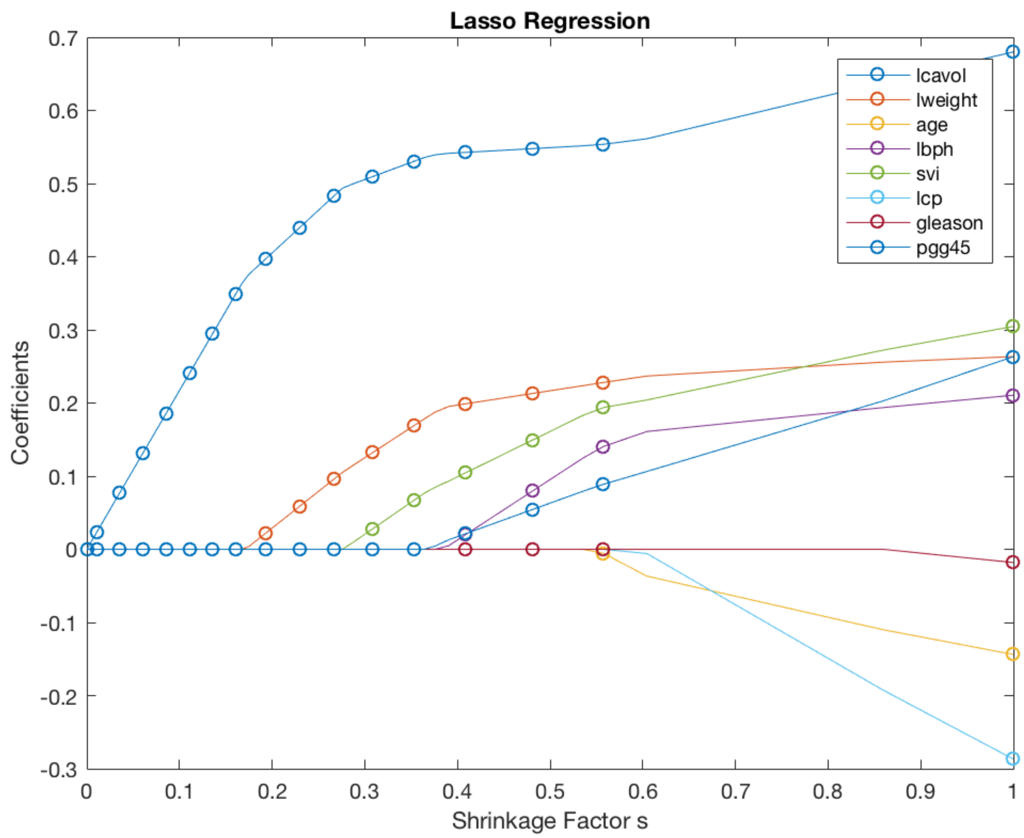
LASSO can be implemented as following algorithm and the full script **Appendix B**:

Algorithm 2: The Cooridnate descent algorithm for LASSO

input : Standardized Z (mean 0, unit variance), Centered y, λ
output: $\hat{w}^{lasso} = [\hat{w}_1, ..., \hat{w}_D]^T$

$\hat{w}_{\lambda}^{ridge} = ridge(Z, y, \lambda)$; `/* initialize parameters` w `*/`
while *Not Converged* **do**
  $w_{old} = w$; `/* Save the previous parameters into` w_{old} `*/`
  $Compute \quad \alpha_d = \Sigma_{i=1}^{n} z_{i,d}^2$;
  $Compute \quad \beta_d = \Sigma_{i=1}^{n} (y_i - w_{-d}^T z_{i,-d}) z_{i,d}$;
  if $w_d < -\lambda$ **then**
    $w_d = \frac{\beta_d + \lambda}{\alpha_d}$;
  else if $w_d > \lambda$ **then**
    $w_d = \frac{\beta_d - \lambda}{\alpha_d}$;
  else
    $w_d = 0$;
  end
  `/* Check the convergence condition */`
  if $max(abs(w_{old} - w)) < threshold$ **then**
    return $\hat{w}^{lasso}$
  else
    $Continue \quad looping$
  end
end

If $t_0 = \Sigma_{i=1}^{D} |\hat{w}_i^{OLS}|$ (equivalently, $\lambda = 0$), we obtain no shrinkage. The path of solution is indexed by a fraction of shrinkage factor of t_0. The result of LASSO path can be given as the below.



Lasso Regression
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
-0.1
-0.2
-0.3
Coefficients
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Shrinkage Factor s
lcavol
lweight
age
lbph
svi
lcp
gleason
pgg45