STATS 790 - Homework 2

Yicen Li

February 13, 2023

1 Question 1

1.1 part a

Suppose $\mathbf{X}$ is a $n \times p$ matrix, $\mathbf{y}$ is a $n \times 1$ vector and $\boldsymbol{\beta}$ is a $p \times 1$ vector. For naive linear algebra, in order to minimize the ordinary least squares:

$$\begin{aligned} RSS &= (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \\ &= \mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathbf{X}\boldsymbol{\beta} - \boldsymbol{\beta}^T\mathbf{X}^T y + \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} \end{aligned}$$

then, we need to make

$$\begin{aligned} \frac{\partial RSS}{\partial \boldsymbol{\beta}} &= \frac{\partial(\mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathbf{X}\boldsymbol{\beta} - \boldsymbol{\beta}^T\mathbf{X}^T y + \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \\ &= \mathbf{0} - \mathbf{X}^T\mathbf{y} - \mathbf{X}^T\mathbf{y} + 2\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} \\ &= \mathbf{0}. \end{aligned}$$

Thus, we conclude:

$$\begin{aligned} 2\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} &= 2\mathbf{X}^T\mathbf{y} \\ \boldsymbol{\beta} &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}. \end{aligned} \tag{1}$$

Hence the equation 1 indicates how to compute linear regression coefficients by naive linear algebra.

By QR decomposition, any real square matrix $\mathbf{A}$ may be decomposed as $\mathbf{A} = \mathbf{QR}$, where $\mathbf{Q}$ is an orthogonal matrix and $\mathbf{R}$ is an upper triangular matrix. Therefore when $\mathbf{X}^T\mathbf{X}$ is ill-conditioned, we may try to rewrite equation 1 based on $(\mathbf{X}^T\mathbf{X}) = \mathbf{QR}$:

$$\begin{aligned} \boldsymbol{\beta} &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \\ &= \mathbf{R}^{-1}\mathbf{Q}^{-1}\mathbf{X}^T\mathbf{y}. \end{aligned}$$

In particular, when $\mathbf{X}$ is a square matrix, we have:

$$\begin{aligned} \boldsymbol{\beta} &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \\ &= ((\mathbf{QR})^T\mathbf{QR})^{-1}(\mathbf{QR})^T\mathbf{y} \\ &= (\mathbf{R}^T\mathbf{Q}^T\mathbf{QR})^{-1}\mathbf{R}^T\mathbf{Q}^T\mathbf{y} \\ &= \mathbf{R}^{-1}\mathbf{Q}^{-1}(\mathbf{Q}^T)^{-1}(\mathbf{R}^T)^{-1}\mathbf{R}^T\mathbf{Q}^T\mathbf{y} \\ &= \mathbf{R}^{-1}\mathbf{Q}^{-1}\mathbf{y} \\ &= \mathbf{R}^{-1}\mathbf{Q}^T\mathbf{y}. \end{aligned}$$

By SVD decomposition, a real $m \times n$ matrix $\mathbf{A}$ can be expressed by $\mathbf{A} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T$ where $\mathbf{U}$ and $\mathbf{V}$ are both orthogonal, and with the form of $m \times m$ and $n \times n$ respectively. $\boldsymbol{\Sigma}$ is a $m \times n$ rectangular diagonal matrix. By $\mathbf{X} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T$, we express equation 1 as:

$$
\begin{aligned}
\boldsymbol{\beta} &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \\
&= [(\mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T)^T\mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T]^{-1}(\mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T)^T\mathbf{y} \\
&= (\mathbf{V}\boldsymbol{\Sigma}^T\mathbf{U}^T\mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T)^{-1}(\mathbf{V}\boldsymbol{\Sigma}^T\mathbf{U}^T)\mathbf{y} \\
&= (\mathbf{V}^T)^{-1}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\Sigma}^T)^{-1}\mathbf{V}^{-1}\mathbf{V}\boldsymbol{\Sigma}^T\mathbf{U}^T\mathbf{y} \\
&= (\mathbf{V}^T)^{-1}\boldsymbol{\Sigma}^{-1}\mathbf{U}^T\mathbf{y} \\
&= (\mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T)^{-1}\mathbf{y}
\end{aligned}
$$

For Cholesky decomposition, in this case, we can expand the real symmetric matrix as $\mathbf{X}^T\mathbf{X} = \mathbf{L}\mathbf{L}^T$, where $\mathbf{L}$ is a lower triangular matrix. We obtain:

$$
\begin{aligned}
\boldsymbol{\beta} &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \\
&= (\mathbf{L}\mathbf{L}^T)^{-1}\mathbf{X}^T\mathbf{y} \\
&= (\mathbf{L}^{-1})^T\mathbf{L}^{-1}\mathbf{X}^T\mathbf{y}
\end{aligned}
$$

1.2 part b

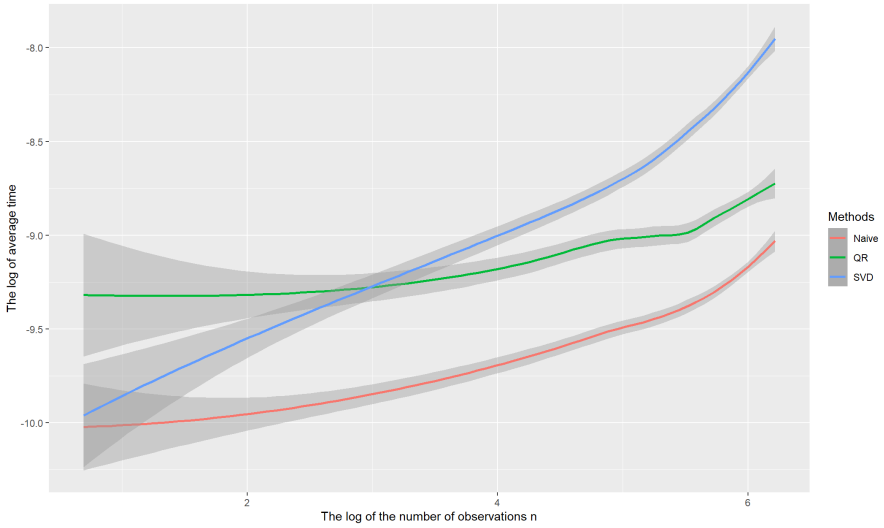

Figure 1: The log-log plot for the number of observations n, fitting in average time. Codes available on Github

In this comparison, we will evaluate three algorithms: the naive approach, QR decomposition, and Singular Value Decomposition (SVD). The design of the comparison involves creating a nested loop to generate a randomly generated design matrix $n \times p$ $\mathbf{X}$ with and corresponding response matrix of size $p \times 1$ $\mathbf{Y}$. We set a huge variance to avoid sparsity. We will then compute the coefficient matrix β using each of these methods and record the average computation time. In cases the $\mathbf{X}^T\mathbf{X}$ becomes ill-conditioned, we will use ginv() function from the MASS package in R to calculate the pseudo-inverse. Note that the running time for QR and SVD will include the decomposition steps, not just the calculation of the product of matrices.

From the log-log plot, we observe that compared to QR decomposition, SVD and naive approach seem to have a slower increasing trend as the growth of log n. However, the naive approach further has a less running time on average than the SVD method.

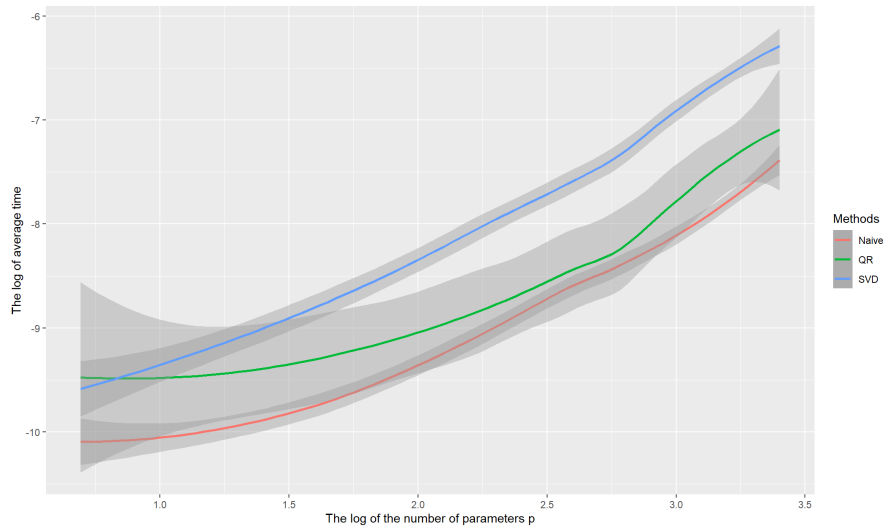

Figure 2: The log-log plot for the number of parameters p, fitting in average time. Codes available on Github

From the above figure, we find that all methods have a close slope as the increase of log p. Still, the naive approach kept a minimal computation time under a given number of parameters to fit the linear regression model.

According to the above initial comparison, we find that applying naive linear algebra to obtain coefficients has the best computation efficiency among the methods.

2 Question 2

In order to perform ridge regression via data augmentation, we apply:

$$\hat{X} = \begin{pmatrix} \mathbf{X} \\ \sqrt{\lambda}\boldsymbol{I}_p \end{pmatrix}$$

$$\hat{y} = \begin{pmatrix} \mathbf{y} \\ \mathbf{0} \end{pmatrix}$$

So that we have

$$\boldsymbol{\beta}_{\boldsymbol{\lambda}} = (\hat{X}^T\hat{X})^{-1}\hat{X}^T\hat{y} \tag{2}$$

Based on Equation 2, a naive function was developed to estimate the $\boldsymbol{\beta}_{\boldsymbol{\lambda}}$ coefficients and fit the ridge regression model on the prostate cancer data provided by [1]. A comparison was conducted between the proposed naive function and the conventional glmnet() function.

It was observed that the data augmentation model takes a computation time of 0.001392841 seconds, which was even slightly faster than the 0.001541138 seconds required by the glmnet() function. To assess the equality of the means of the coefficients, a two-sample T-test was performed. The resulting P-value of 0.9918 indicated extremely weak evidence against the null hypothesis of equal means. The plot 3 of the absolute differences for each coefficient $\beta_0, ..., \beta_9$ is presented in the accompanying figure. Additionally, the squared training loss for both methods was found to be very similar, with an error of approximately 0.4440.

In conclusion, on this dataset, the data augmentation method demonstrated comparable performance to the glmnet() function.

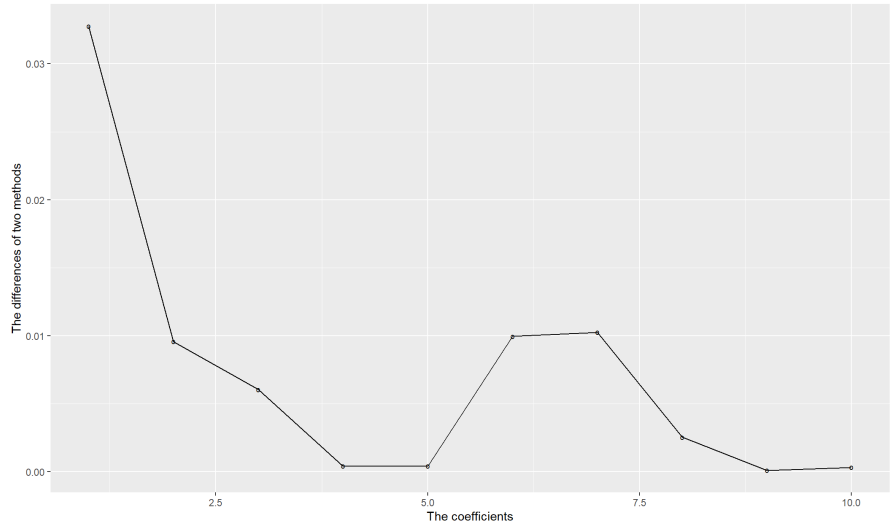

Figure 3: The differences between the coefficients of two ridge regression methods when $\lambda = 0.01$. Codes available on Github

3 Question ESL3.6

We know that $\beta \sim N(0, \tau\mathbf{I})$, which indicates

$$p(\beta) = (2\pi\tau\mathbf{I})^{-1/2} exp(-\frac{\beta^2}{2\tau})$$

and we have

$$p(y|\beta) = ((2\pi)^n \sigma^2\mathbf{I})^{-1/2} exp(-\frac{(y - x\beta)^2}{2\sigma^2})$$

We may try the scalar case to simplify the proof. By Bayes's therom, the posterior distribution of β

$$\begin{aligned} p(\beta|y) \; &\alpha \; p(y|\beta)p(\beta) \\ &\alpha \; exp(-\frac{(y - x\beta)^2}{2\sigma^2}) exp(-\frac{\beta^2}{2\tau}) \\ &\alpha \; exp(-\frac{(y - x\beta)^2 + \frac{\sigma^2}{\tau}\beta^2}{2\sigma^2}) \end{aligned}$$

Notice that it has the same form as the function $exp(-(\beta + b)^2 + a)$, which is symmetric and the mean of this function should be located at the center, and this center is also exactly the maximum point. For example, for N(0,1), the mean is 0 and also the point which has maximum density. Therefore to get mean, we compute

$$\arg\min_{\beta} \; (y - x\beta)^2 + \frac{\sigma^2}{\tau}\beta^2$$

if we treat $\lambda = \frac{\sigma^2}{\tau}$, we find that it is exactly the ridge regression estimate. Similarly, replace the above scalar with a vector, we obtain: $\boldsymbol{\beta}_\lambda = (\mathbf{X}^T\mathbf{X} + \boldsymbol{\lambda}\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}$.

4 Question ESL3.19

By ESL equation 3.47, which the author derived from performing SVD on $\mathbf{X}$. That is, $\mathbf{X} = \mathbf{U}\boldsymbol{D}\mathbf{V}^T$. It shows that

$$\mathbf{X}\boldsymbol{\beta}_\lambda = \mathbf{U}\mathbf{D}(\mathbf{D}^2 + \lambda\mathbf{I})^{-1}\mathbf{D}\mathbf{U}^T\mathbf{y}$$

therefore,

$$\begin{aligned} \boldsymbol{\beta}_\lambda &= \mathbf{X}^{-1}\mathbf{X}\boldsymbol{\beta}_\lambda \\ &= \mathbf{V}(\mathbf{D}^2 + \lambda\mathbf{I})^{-1}\mathbf{D}\mathbf{U}^T\mathbf{y}. \end{aligned}$$

Then we have:

$$\begin{aligned}\|\boldsymbol{\beta}_\lambda\|_2^2 &= \boldsymbol{\beta}_\lambda^T\boldsymbol{\beta}_\lambda \\ &= \mathbf{y}^T\mathbf{U}\mathbf{D}^T(\mathbf{D}^2+\Lambda\mathbf{I})^{-1}\mathbf{V}^T\mathbf{V}(\mathbf{D}^2+\lambda\mathbf{I})^{-1}\mathbf{D}\mathbf{U}^T\mathbf{y} \\ &= \mathbf{y}^T\mathbf{U}\mathbf{D}^T(\mathbf{D}^2+\Lambda\mathbf{I})^{-2}\mathbf{D})\mathbf{U}^T\mathbf{y} \\ &= (\mathbf{U}^T\mathbf{y})^T(\mathbf{D}^T(\mathbf{D}^2+\Lambda\mathbf{I})^{-2}\mathbf{D})(\mathbf{U}^T\mathbf{y})\end{aligned}$$

Since $\mathbf{D}$ is a diagonal matrix, we have:

$$\|\boldsymbol{\beta}_\lambda\|_2^2 = \sum_i^n \frac{d_i^2(U^Ty)_i}{(d_i^2+\lambda)^2}.$$

Therefore, since λ is under the fraction, it indicates that as λ goes 0, the $\|\boldsymbol{\beta}_\lambda\|_2^2$ tend to increase.

As for Lasso, it is still a convex optimization problem but without explicit solutions to β. So that we try to explore it from the loss function:

$$\arg\min_{\beta_{lasso}} \frac{1}{2}\|\mathbf{y}-\beta_0-\mathbf{X}\boldsymbol{\beta}_{lasso}\|^2+\lambda|\boldsymbol{\beta}_{lasso}|$$

When we achieve the minimal error, if λ turns to 0, then the $|\boldsymbol{\beta}_{lasso}|$ should also increase to maintain the minimal error. Based on this idea, the p-norm case still holds. Thus I think this behaviour holds for Lasso, ridge and other regression using p-norm as a penalty.

5 Question ESL3.28

Given predictors $\mathbf{X}$, response $\mathbf{y}$, and under the same value t, the solution provided by lasso is: $\mathbf{y}=\mathbf{X}\boldsymbol{\beta}$. Now, the "training set" turns to be $\hat{\mathbf{X}}=[\mathbf{X}^*,\mathbf{X}^*]$, where $\mathbf{X}^*=\mathbf{X}$ (It is not the $[\mathbf{X}^*,\mathbf{X}^*]^T$). We assume the coefficients for the new design matrix should be the $\hat{\boldsymbol{\beta}}=[\boldsymbol{\beta}_1^*,\boldsymbol{\beta}^*]_2^T$. Thus, the regression function becomes:

$$\begin{aligned}\mathbf{y} &= \hat{\mathbf{X}}\hat{\boldsymbol{\beta}} \\ &= [\mathbf{X}^*,\mathbf{X}^*][\boldsymbol{\beta}_1^*,\boldsymbol{\beta}^*]_2^T \\ &= \mathbf{X}^*\boldsymbol{\beta}_1^*+\mathbf{X}^*\boldsymbol{\beta}_2^* \\ &= \mathbf{X}\boldsymbol{\beta}_1^*+\mathbf{X}\boldsymbol{\beta}_2^* \\ &= \mathbf{X}(\boldsymbol{\beta}_1^*+\boldsymbol{\beta}_2^*).\end{aligned}$$

Recall that $\mathbf{y}=\mathbf{X}\boldsymbol{\beta}$, thus, we conclude:

$$\mathbf{X}(\boldsymbol{\beta}^*+\boldsymbol{\beta}^*)=\mathbf{X}\boldsymbol{\beta}$$

So that we must have $\boldsymbol{\beta}_1^*+\boldsymbol{\beta}_2^*=\boldsymbol{\beta}$. Thus, for an individual jth variable $(\beta_1)_j^*,(\beta_2)_j^*$, it should fulfill: $(\beta_1)_j^*+(\beta_2)_j^*=\beta_j=a$.

6 Question ESL3.30

Inspired by the data augmentation for ridge regression [1], we can try:

$$\hat{X}=\begin{pmatrix}\mathbf{X}\\ \tau\boldsymbol{I}_p\end{pmatrix}$$

$$\hat{y}=\begin{pmatrix}\mathbf{y}\\ \mathbf{0}\end{pmatrix}$$

Then the least squares problem becomes:

$$\arg\min_{\beta}\ \|\hat{y}-\hat{X}\boldsymbol{\beta}\|^2=\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|^2+\tau^2\|\boldsymbol{\beta}\|^2$$

If we apply Lasso on the $\hat{X}$ and $\hat{y}$, it should have this form:

$$\arg\min_{\beta} \ \|\hat{y} - \hat{X}\boldsymbol{\beta}\|^2 + \hat{\lambda}\|\boldsymbol{\beta}\|_1$$

where

$$\arg\min_{\beta} \ \|\hat{y} - \hat{X}\boldsymbol{\beta}\|^2 + \hat{\lambda}\|\boldsymbol{\beta}\|_1 = \arg\min_{\beta} \ \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \tau^2\|\boldsymbol{\beta}\|^2 + \hat{\lambda}\|\boldsymbol{\beta}\|_1$$

By comparison with the given elastic-net optimization expression (eq 3.91), we can take $\tau^2 = \lambda\alpha$ and our $\hat{\lambda} = \lambda(1-\alpha)$ to match. In conclusion, we take data augmentation as:

$$\hat{X} = \begin{pmatrix} \mathbf{X} \\ \sqrt{\lambda\alpha}\boldsymbol{I}_p \end{pmatrix}$$

$$\hat{y} = \begin{pmatrix} \mathbf{y} \\ \mathbf{0} \end{pmatrix}$$

Then we turn this into a lasso problem on $\hat{X}$ and $\hat{y}$ with $\hat{\lambda} = \lambda(1-\alpha)$.

References

[1] Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.