CMSE 820: Homework #1

Due on September 13, 2019 at 11:59pm

Professor Yuying Xie

Boyao Zhu

Problem 1

Assume $Y = X^T\beta + \epsilon$, where $X \in \mathbb{R}^P$ is not random and $\epsilon \sim N(0,1)$. Given i.i.d. data $\{(x_1, y_1),...(x_n, y_n)\}$, we would like to estimate $\beta \in \mathbb{R}^P$ through maximum likelihood framework. Write down the joint log likelihood and compare it with least square method.

Solution

Let us assume that the disturbances ϵ_t, which are the elements of the vector $\epsilon = [\epsilon_1, \epsilon_2, ..., \epsilon_t]'$, are distributed independently and identically according a normal distribution

$$N(\epsilon_t, 0, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{\frac{-1}{2\sigma^2}(y_t - x_t^T\beta)^2\right\}$$

Then, if the vectors x_t are taken as data, the observations y_t; $t = 1, ..., T$ have density functions $N(y_t; x_t^T\beta, \sigma^2)$ which are of the same form as those of the disturbances, and the likelihood function of β and σ^2, based on the sample, is

$$L = \prod_{T=1}^{T} N(y_t; x_t^T\beta, \sigma^2) = (2\pi\sigma^2)^{-T/2} \exp\left\{\frac{-1}{2\sigma^2}(y - X^T\beta)^T(y - X^T\beta)\right\}$$

The logarithm of this function

$$L^*(\beta, \sigma) = -\frac{T}{2}\ln(2\pi) - \frac{T}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}(y - X^T\beta)^T(y - X^T\beta)$$

To find the maximum-likelihood estimator of β, we set the derivative of equation with respect to β to zero

$$\frac{\partial L^*}{\partial\beta} = \frac{1}{\sigma^2}(y - X^T\beta)^T X^T = 0$$

The solution of the equation is the estimator

$$\tilde{\beta} = (XX^T)^{-1}Xy$$

which is equivalent to least square method.

$\boxed{\text{QED}}$

Problem 2

Show that we can decompose the expected prediction error, $\mathbb{E}[(Y_0 - \hat{f}(x_0))^2]$ at an input point $X = x_0$ for a general model $Y = f(X) + \epsilon$ with $\mathbb{E}(\epsilon) = 0$ and $\text{Var}(\epsilon) = \sigma^2$:

$$\mathbb{E}[(Y_0 - \hat{f}(x_0))^2] = \sigma^2 + \text{Var}(\hat{f}(x_0)) + \text{Bias}^2$$

Solution

$$\begin{aligned}
\mathbb{E}[(Y_0 - \hat{f}(x_0))^2] &= \mathbb{E}[(f(x_0) + \epsilon - \hat{f}(x_0))^2] \\
&= \mathbb{E}[(f(x_0) - \hat{f}(x_0))^2 + \epsilon^2 + 2(f(x_0) - \hat{f}(x_0))\epsilon] \\
&= \mathbb{E}[(f(x_0) - \hat{f}(x_0))^2] + \mathbb{E}[\epsilon^2] + 2\mathbb{E}[(f(x_0) - \hat{f}(x_0))\epsilon] \\
&= \mathbb{E}[(f(x_0) - \mathbb{E}[\hat{f}(x_0)] + \mathbb{E}[\hat{f}(x_0)] - \hat{f}(x_0))^2] + \mathbb{E}[\epsilon^2] + 2\mathbb{E}[(f(x_0) - \hat{f}(x_0))\epsilon] \\
&= \text{Var}(\hat{f}(x_0)) + \text{Bias}^2(\hat{f}(x_0)) + \sigma^2 + 2\mathbb{E}[f(x_0)\epsilon - \hat{f}(x_0)\epsilon] \\
&= \text{Var}(\hat{f}(x_0)) + \text{Bias}^2(\hat{f}(x_0)) + \sigma^2 - 2\hat{f}(x_0)\mathbb{E}(\epsilon) + 2\mathbb{E}[f(x_0)\epsilon] \\
&= \text{Var}(\hat{f}(x_0)) + \text{Bias}^2(\hat{f}(x_0)) + \sigma^2 - 2\hat{f}(x_0)0 + 2\mathbb{E}[f(x_0)]\mathbb{E}[\epsilon] \\
&= \text{Var}(\hat{f}(x_0)) + \text{Bias}^2(\hat{f}(x_0)) + \sigma^2
\end{aligned}$$

Problem 2 continued on next page. . .

where $\mathbb{E}[(f(x_0)-\hat{f}(x_0))^2] = \mathrm{Var}(\hat{f}(x_0))+\mathrm{Bias}^2(\hat{f}(x_0))$, as expected from MSE. And $\mathbb{E}[\epsilon^2] = \sigma^2+\mathbb{E}[\epsilon]^2 = \sigma^2$.
$\boxed{\text{QED}}$

Problem 3

Consider the usual linear regression setup, with response vector $\mathbf{y} \in \mathbb{R}^n$ and predictor matrix $\mathbf{X} \in \mathbb{R}^{p\times n}$. Let $x_1, ..., x_p$ be the rows of $\mathbf{X}$. Suppose that $\hat{\beta} \in \mathbb{R}^p$ is a minimizer of the least squares criterion

$$\|\mathbf{y} - \mathbf{X}^T\beta\|^2$$

1. Show that if $v \in \mathbb{R}^p$ is a vector such that $\mathbf{X}^T v = 0$, then $\hat{\beta}+c\cdot v$ is also a minimizer of the least squares criterion, for any $c \in \mathbb{R}$.
2. If $x_1, ..., x_p \in \mathbb{R}^p$ are linearly independent, then what vectors $v \in \mathbb{R}^p$ satisfy $(X)^T v = 0$? We assume $p \leq n$.
3. Suppose that $p > n$. Show that there exists a vector $v \neq 0$ such that $\mathbf{X}^T v = 0$. Argue, based on part (a), that there are infinitely many linear regression estimates. Further argue that there is a variable $i \in 1, ..., p$ such that the regression coefficient of variable $\beta_{[i]}$ can have different signs, depending on which estimate we choose, Comment on this.

Solution
1.
Let $L = \|\mathbf{y} - \mathbf{X}^T\beta\|^2$, $\hat{\beta} \in \mathbb{R}^p$ is a minimizer of the least squares if and only if $\frac{\partial L}{\partial \beta} = 0$.

$$\begin{aligned}\frac{\partial L}{\partial \beta} &= (\mathbf{y} - \mathbf{X}^T\beta)^T(\mathbf{y} - \mathbf{X}^T\beta)\\ &= \frac{\partial}{\partial \beta}(\|\mathbf{y}\|^2 - \mathbf{y}^T\mathbf{X}^T\beta - \beta^T\mathbf{X}\mathbf{y} + \beta^T\mathbf{X}\mathbf{X}^T\beta)\\ &= 2\mathbf{X}\mathbf{X}^T\beta - 2\mathbf{X}\mathbf{y}\\ &\equiv 0\end{aligned}$$

This indicates

$$\hat{\beta} = (\mathbf{X}\mathbf{X}^T)^{-1}\mathbf{X}y$$

$\boxed{\text{QED}}$
which minimizes L. Then $\beta' = \hat{\beta} + c\cdot v$ is also a minimizer of the least squares given $\mathbf{X}^T v = 0$ since

$$\begin{aligned}L' &= \|\mathbf{y} - \mathbf{X}^T\beta'\|^2\\ &= \|\mathbf{y} - \mathbf{X}^T(\hat{\beta} + c\cdot v)\|^2\\ &= \|\mathbf{y} - \mathbf{X}^T\hat{\beta} + c\cdot \mathbf{X}v\|^2\\ &= \|\mathbf{y} - \mathbf{X}^T\hat{\beta}\|^2\end{aligned}$$

$\boxed{\text{QED}}$
2.
Given $p \leq n$, $x_1, ..., x_p \in \mathbb{R}^n$ are linearly independent means X is a full rank matrix, which implies that the kernel space $ker(X) = \mathbf{0}$. So v can only be $\mathbf{0}$.
$\boxed{\text{QED}}$
3.
Suppose $p > n$, then X is rank deficient, which implies that $x_1, ..., x_p$ are not linearly independent, so there exists a **non-zero** vector $v \in \mathbb{R}^p$ such that

$$\Sigma_{i=1}^p v_i x_i = 0$$

Problem 3 continued on next page...

Equivalently, $X^T v = \mathbf{0}$
According to (a), if $\hat{\beta}$ minimizes $\|\mathbf{y} - \mathbf{X}^T \hat{\beta}\|$, then $\hat{\beta} + c \cdot v$ is also a minimizer of the least squares criterion, for any $c \in \mathbb{R}$. So there are infinitely many linear regression estimates. Suppose v is a nonzero vector, there is a variable $i \in 1, ..., p$ and let $\hat{\beta}$ be one of its linear regression estimates. Without loss of generality, assume $\hat{\beta}_{[i]} > 0$, let c be chosen as follows

$$c = \begin{cases} -\hat{\beta}_{[i]}/v_i - 1, & v_i > 0, \\ -\hat{\beta}_{[i]}/v_i + 1, & v_i < 0 \end{cases}$$

Define another linear regression estimate $\hat{\beta}' = \hat{\beta} + c \cdot v$. One can easily check that $\hat{\beta}_{[i]}\hat{\beta}'_{[i]} < 0$. In this case, the i-th element of the estimates have different signs. The behavior of the response with respect to the i-th degree of freedom is dependent on other degrees of freedom, which means it is a linear combination of the other degrees of freedom.

Problem 4

Implement the following model (you can use any language)

$$Y = X^T \beta + \epsilon$$

where $\epsilon \sim N(0, 1)$, $X \sim N(0, I_{p \times p})$ and $\beta \in \mathbb{R}^p$ with $\beta_{[1]} = 1$, $\beta_{[2]} = -2$ and the rest of $\beta_{[j]} = 0$. For $p = 5$, simulate $(x_1, ..., x_100)$ and the corresponding Ys. Based on this data, calculate $\hat{\beta}^{ols}$ and store it. then do the followings

1. Based on the β and $(x_1, ..., x_100)$ we first simulate the corresponding $Y's$ and calculate the $\hat{\beta}^{ols}$.
2. Use the same $(x_1, ..., x_{100})$, we then simulate another set of $\tilde{Y} = (y_1, ..., y_{100})$ and calculate the in-sample prediction error using $\hat{\beta}^{ols}$ calculated in (1). This is one realization of $\mathbf{PE}_{in}$.
3. Repeat (1) - (2) 5000 times and take average of those 5000 calculated $\mathbf{PE}_{in}$. You have an approximate $\mathbf{PE}_{in}$.
4. Repeat the same procedure for p = 10, 40, 80. What is the trend for the $\mathbf{PE}_{in}$? Comment your findings.

Solutions
1.
$\hat{\beta}$ = [0.99866, -1.96044, -0.04670, -0.01330, 0.01773]
2.
$\mathbf{PE}_{in}$ = 92.99847
3.
averaging over 5000 simulation, $\mathbf{PE}_{in}$ = 102.32915
4.
p = 5, $\mathbb{E}[\mathbf{PE}_{in}]$ = 102.32915
p = 10, $\mathbb{E}[\mathbf{PE}_{in}]$ = 106.15731
p = 20, $\mathbb{E}[\mathbf{PE}_{in}]$ = 118.64808
p = 40, $\mathbb{E}[\mathbf{PE}_{in}]$ = 132.09621
p = 80, $\mathbb{E}[\mathbf{PE}_{in}]$ = 173.45571
We saw that $\mathbb{E}[\mathbf{PE}_{in}]$ has a trend increase as the complexity of model increase, which corresponds to $\mathbf{PE}_{in}$ grows linearly as p increases.

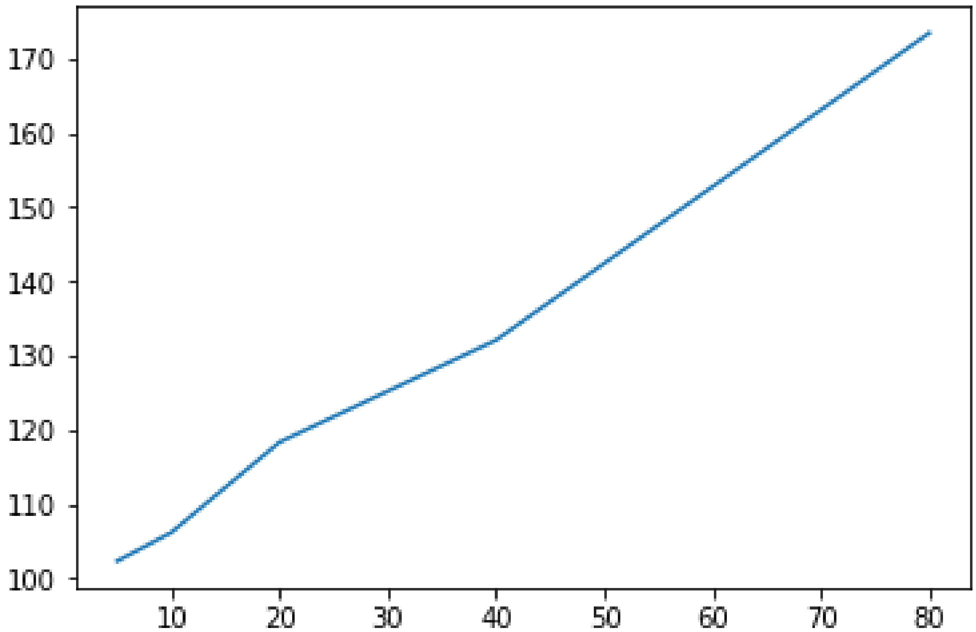

Figure 1: Brute plot of in-sample prediction error

Problem 5

Implement the following model (you can use any language)

$$y_i = \beta^*_{[1]} x_{i[1]} + \beta^*_{[2]} x_{i[2]} + \epsilon_i$$

where $\mathbb{E}(\epsilon_i) = 0$, $\text{Var}(\epsilon_i) = 1$, $\text{Cov}(x_i, x_j) = 0$ and $\beta = (-1, 2)^T$. We also assume $x_i \sim N(0, \Sigma_x)$ with

$$\Sigma_x = \text{Cov}(x_i) = \begin{pmatrix} 1 & 0.9999 \\ 0.9999 & 1 \end{pmatrix}$$

We repeated the following 2000 times:

- Generate $\mathbf{y} = (y_1, ..., y_{50})^T$ and $\mathbf{X} = (x_1, ..., x_{50})$.
- compute and record $\hat{\beta}^{ols}$ and $\hat{\beta}^{ridge}$. What conclusion can you make from these histograms?

Then report the followings:

a. The histograms for $\hat{\beta}^{ols}_{[1]}$ and $\hat{\beta}^{ridge}_{[1]}$. What conclusion can you make from these histograms?

b. For each replicate of the 2000 repeats, compare $|\hat{\beta}^*_{[1]} - \hat{\beta}^{ols}_{[1]}|$ with $|\beta^*_{[1]} - \beta^{ridge}_{[1]}|$. How many times does ridge regression return a better estimate of $\beta^*_{[1]}$?

Solutions
a. The predictions from OLS have a greater variance than those from Ridge, as shown in Figure 2.
b. Among the 2000 runs, $|\beta^*_{[1]} - \beta^{ridge}_{[1]}| < |\beta^*_{[1]} - \beta^{ols}_{[1]}|$ for 1829 times. That is, about 91.45% of the times, ridge regression yields a better estimate compared to ordinary least square (OLS) regression.

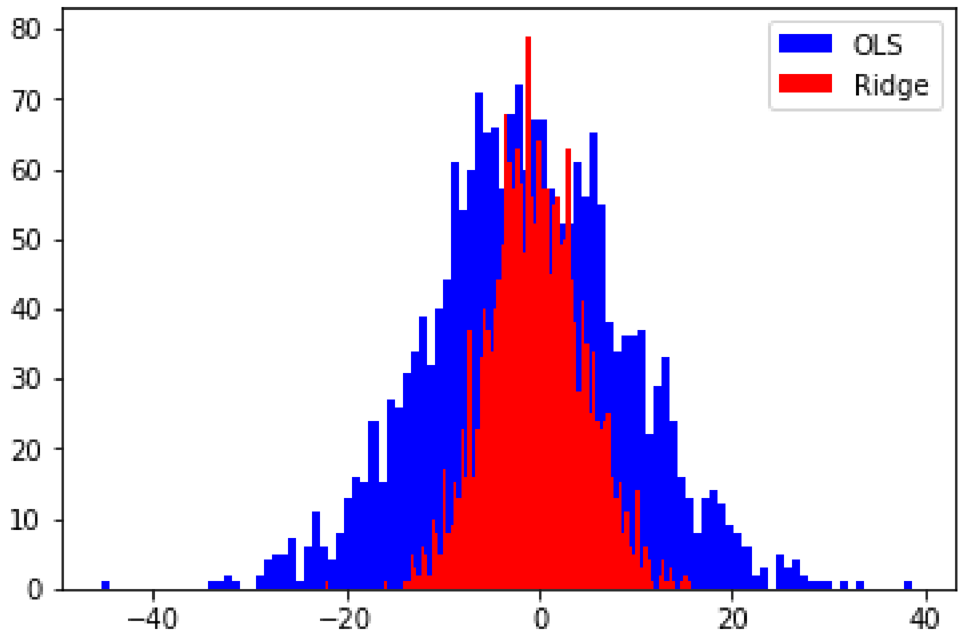

Figure 2: Histogram for $\hat{\beta}^{ols}_{[1]}$ and $\hat{\beta}^{ridge}_{[1]}$