CMSE 820 HW 1

Hao Lin

September 13, 2018

1 Question 1

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

2 Question 2

By definition,

$$\begin{aligned}
\hat{f}(x_0) &= \arg\min_f \mathbb{E}\big[(Y - f(X))^2\big] \\
&= \arg\min_f \int_{\Omega_X}\int_{\Omega_Y} (y - f(x))^2 f_{\mathrm{X,Y}}(x, y)\, dxdy \\
&= \arg\min_f \int_{\Omega_X}\int_{\Omega_Y} (y - f(x))^2 f_{\mathrm{Y|X}}(y|x) f_{\mathrm{X}}(x)\, dxdy
\end{aligned}$$

In order to minimize $\mathbb{E}\big[(Y - f(X))^2\big]$ with $\hat{f}$, we impose that

$$\forall x' \in \Omega_{\mathrm{X}}, \frac{\delta \mathbb{E}\big[(Y - f(X))^2\big]}{\delta \hat{f}(x')} = 0,$$

where Ω_{X} denotes the sample space of X.

$$\begin{aligned}
\forall x' \in \Omega_{\mathrm{X}}, \frac{\delta \mathbb{E}\big[(Y - f(X))^2\big]}{\delta \hat{f}(x')} &= -2\, f_{\mathrm{X}}(x') \int (y - \hat{f}(x')) f_{\mathrm{Y|X}}(y|x')\, dy \\
&= -2\, f_{\mathrm{X}}(x')(\mathbb{E}[y|x'] - \hat{f}(x')) = 0.
\end{aligned}$$

It follows immediately that $\hat{f}(X) = \mathbb{E}[Y|X]$.

3 Question 3

3.1 a.

Let $v \in \mathbb{R}^p$ be a vector such that $\mathbf{X}^T v = 0$.

$$\begin{aligned} \|\mathbf{y} - \mathbf{X}^T(\hat{\beta} + c \cdot v)\| &= \|\mathbf{y} - \mathbf{X}^T\hat{\beta} + c\mathbf{X}^T v\| \\ &= \|\mathbf{y} - \mathbf{X}^T\hat{\beta}\|. \end{aligned}$$

Thus, $\hat{\beta} + c \cdot v$ is also a minimizer.

3.2 b.

If $x_1, ..., x_p \in \mathbb{R}^p$ are linearly independent and $p \leq n$, then X is a full rank matrix, which implies that the kernel space $\ker(X) = \{\mathbf{0}\}$. So v can only be $\mathbf{0}$.

3.3 c.

$$\text{rank}(\text{col}(X)) = \text{rank}(X) \leq n < p,$$

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

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

Equivalently, $X^T v = \mathbf{0}$.

According to (a.), if $\hat{\beta}$ minimizes $\|\mathbf{y} - \mathbf{X}^T\hat{\beta}\|$, then $\forall c \in \mathbb{R}$, $\hat{\beta} + c \cdot v$ also minimizes it. So there are infinitely many linear regression estimates.

Since v is a nonzero vector, there exists a subscript i such that $v_i \neq 0$. Let $\hat{\beta}$ be one of the 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 coefficient of the estimates have different signs. The behavior of the response function with respect to the i-th degree of freedom is not monotonic and is dependent on other degrees of freedom, due to the fact that the i-th degree of freedom is a linear combination of the other degrees of freedom.

4 Question 4

$$\hat{\beta} = \arg\min_{\beta} \|\mathbf{y} - \beta_{[1]}\mathbb{1} - \tilde{X}^T\beta_{-[1]}\|^2 + \lambda\|\beta_{-[1]}\|^2.$$

Setting the derivative with respect to $\beta_{[1]}$ of the above expression to be 0 yields

$$\begin{aligned} & -2\,\Sigma_{i=1}^{n}\big(y_i - \beta_{[1]} - x_i^T\beta_{-[1]}\big) = 0 \\ \Rightarrow \beta_{[1]} &= \frac{1}{n}\Sigma_{i=1}^{n}\big(y_i - x_i^T\beta_{-[1]}\big) \\ &= \frac{1}{n}\Sigma_{i=1}^{n}y_i - \Big(\frac{1}{n}\Sigma_{i=1}^{n}x_i^T\Big)\beta_{-[1]} \\ &= \bar{y} - \mathbf{0}\beta_{-[1]} = \bar{y}. \end{aligned}$$

Note that $\frac{1}{n}\Sigma_{i=1}^{n}x_i^T = \mathbf{0}$ because $\tilde{X}^T$ is centered.

5 Question 5

5.1 a.

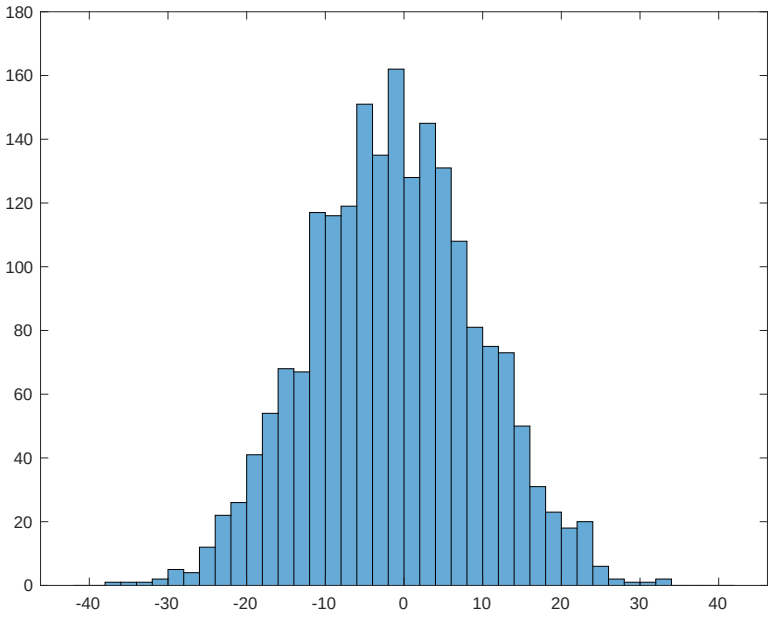

Figure 1: Histogram for $\hat{\beta}_{[1]}^{\text{ols}}$

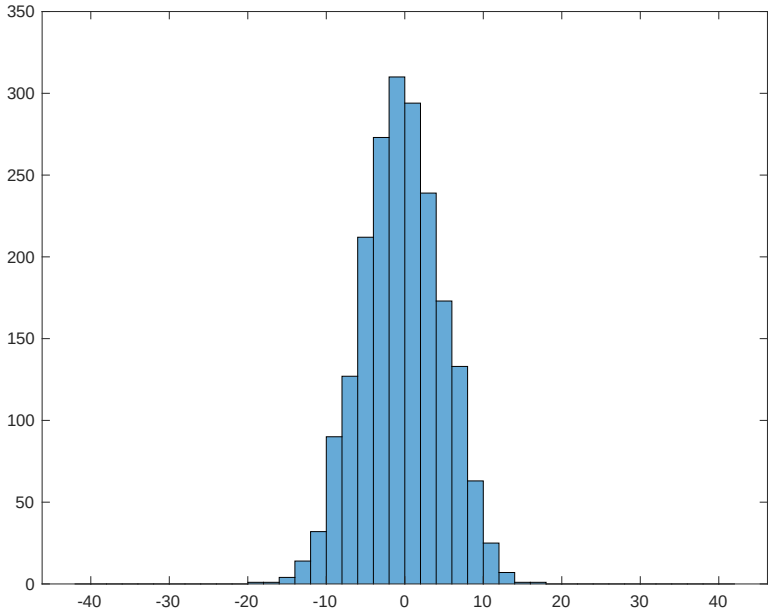

Figure 2: Histogram for $\hat{\beta}^{\text{ridge}}_{[1]}$

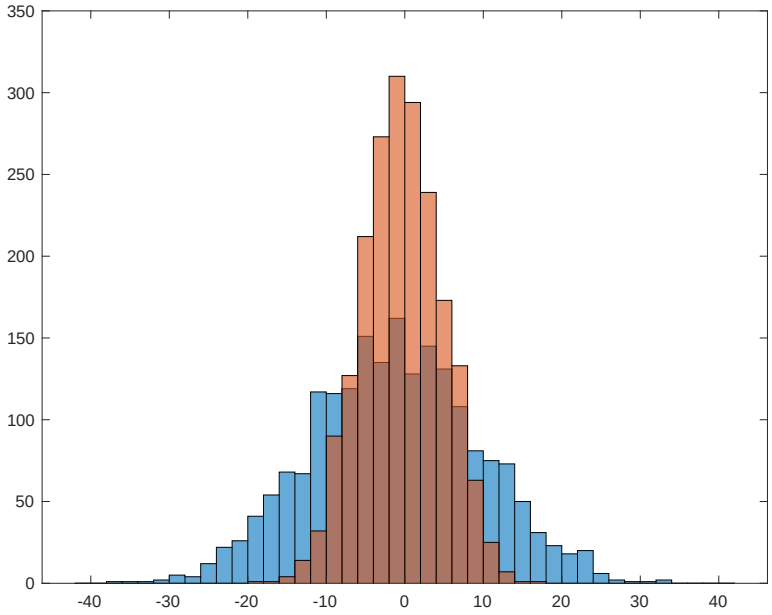

Figure 3: Histogram for $\hat{\beta}^{\text{ols}}_{[1]}$ and $\hat{\beta}^{\text{ridge}}_{[1]}$

Conclusions: 1. The predictions from OLS have a greater variance than those from Ridge. 2. The modes of both predictions fall in the $[-2, 0]$ bin, where the true $\beta^*_{[1]}$ resides as well.

5.2 b.

Among the 2000 runs, $|\beta^*_{[1]} - \hat{\beta}^{\text{ridge}}_{[1]}| < |\beta^*_{[1]} - \hat{\beta}^{\text{ols}}_{[1]}|$ for 1845 times. That is, about 92.25% of the time, Ridge regression yields a better estimate of $\beta^*_{[1]}$, when compared with ordinary least squares (OLS) regression.