THEORETICAL QUESTIONS - ANSWERS

WILLIAM LIAW

Ordinary Least Squares

Before demonstrating that the Ordinary Least Squares (OLS) estimator has the smallest variance among all linear unbiased estimators, we first analyze the OLS estimator:

$$
\begin{aligned}
y &= X\beta + \varepsilon \\
y - \varepsilon &= X\beta \\
X^T(y - \varepsilon) &= X^T X\beta \\
(X^T X)^{-1} X^T (y - \varepsilon) &= \beta \\
(X^T X)^{-1} X^T y - (X^T X)^{-1} X^T \varepsilon &= \beta \\
(X^T X)^{-1} X^T y &= \beta + (X^T X)^{-1} X^T \varepsilon \\
\beta^* &= \beta + (X^T X)^{-1} X^T \varepsilon
\end{aligned}
$$

We can observe that the OLS is unbiased $\beta^* = \beta$ only if X is deterministic and $\mathbb{E}(\varepsilon) = 0$. In this case, assuming $\mathbb{V}(\varepsilon) = \sigma^2 I$, we can calculate the variance of the OLS estimator:

$$
\begin{aligned}
\mathbb{V}(\beta^*) &= \mathbb{E}((\beta^* - \mathbb{E}(\beta^*))(\beta^* - \mathbb{E}(\beta^*))^T) \\
&= \mathbb{E}((\beta^* - \beta)(\beta^* - \beta)^T) \\
&= \mathbb{E}((\beta + (X^T X)^{-1} X^T \varepsilon - \beta)(\beta + (X^T X)^{-1} X^T \varepsilon - \beta)^T) \\
&= \mathbb{E}(((X^T X)^{-1} X^T \varepsilon)((X^T X)^{-1} X^T \varepsilon)^T) \\
&= \mathbb{E}(((X^T X)^{-1} X^T \varepsilon)(\varepsilon^T X (X^T X)^{-1})) \\
&= (X^T X)^{-1} X^T \mathbb{E}(\varepsilon \varepsilon^T) X (X^T X)^{-1} \\
&= \sigma^2 (X^T X)^{-1}
\end{aligned}
$$

Then, we compute the expected value and variance of the alternative estimator $\tilde{\beta}$.

Expected value

First, we compute the expected value of $\tilde{\beta}$:

$$
\begin{aligned}
\mathbb{E}(\tilde{\beta}) &= \mathbb{E}(Cy) \\
&= C\mathbb{E}(y) \\
&= (H + D)\mathbb{E}(y) \\
&= ((X^T X)^{-1} X^T + D)\mathbb{E}(y) \\
&= ((X^T X)^{-1} X^T + D)\mathbb{E}(X\beta + \varepsilon) \\
&= ((X^T X)^{-1} X^T + D)X\beta \\
&= (I + DX)\beta
\end{aligned}
$$

Thus, for $\tilde{\beta}$ to be unbiased $(I + DX)\beta = I\beta$, it is necessary that $DX = 0$.

Variance

Consequently, we compute the variance of $\tilde{\beta}$:

$$
\begin{aligned}
\mathbb{V}(\tilde{\beta}) &= \mathbb{V}(Cy) \\
&= C\mathbb{V}(y)C^T \\
&= C\mathbb{E}((y - \mathbb{E}(y))(y - \mathbb{E}(y))^T)C^T \\
&= C\mathbb{E}((X\beta + \varepsilon - \mathbb{E}(X\beta + \varepsilon))(X\beta + \varepsilon - \mathbb{E}(X\beta + \varepsilon))^T)C^T \\
&= C\mathbb{E}((\varepsilon - \mathbb{E}(\varepsilon))(\varepsilon - \mathbb{E}(\varepsilon))^T)C^T \\
&= C\mathbb{V}(\varepsilon)C^T \\
&= \sigma^2 CC^T \\
&= \sigma^2 (H + D)(H + D)^T \\
&= \sigma^2 (HH^T + HD^T + DH^T + DD^T) \\
&= \sigma^2 (HH^T + (XX^T)^{-1} X^T D^T + DX(XX^T)^{-1} + DD^T) \\
&= \sigma^2 (HH^T + DD^T) \\
&= \sigma^2 HH^T + \sigma^2 DD^T \\
&= \mathbb{V}(\beta^*) + \sigma^2 DD^T
\end{aligned}
$$

Conclusion

From the last expression, it's evident that $\mathbb{V}(\tilde{\beta})$ is always greater than or equal to $\mathbb{V}(\beta^*)$ since D is non-zero, hence DD^T introduces additional variance.

The assumption of OLS that we need to use here is that X is deterministic and $\mathbb{E}(\varepsilon) = 0$

Ridge regression

Biased estimator

To show that the estimator of ridge regression, denoted as β^*_{ridge}, is biased, we need to demonstrate that its expected value, $E(\beta^*_{\text{ridge}})$, does not equal the true parameter vector β.

Analyzing the ridge estimator $\beta^*_{\text{ridge}} = \text{argmin}_\beta \parallel y - X\beta \parallel_2^2 + \lambda \parallel \beta \parallel_2^2$ and denoting $f(\beta) = \parallel y - X\beta \parallel_2^2 + \lambda \parallel \beta \parallel_2^2$, we have:

$$
\begin{aligned}
f(\beta) &= \parallel y - X\beta \parallel_2^2 + \lambda \parallel \beta \parallel_2^2 \\
&= (y - X\beta)^T(y - X\beta) + \lambda\beta^T\beta \\
&= (y^T - \beta^T X^T)(y - X\beta) + \lambda\beta^T\beta \\
&= (y^T y - y^T X\beta - \beta^T X^T y + \beta^T X^T X\beta) + \lambda\beta^T\beta \\
\therefore f'(\beta) &= 2X^T X\beta - 2X^T y + 2\lambda\beta \\
\Rightarrow \beta^*_{\text{ridge}} &= (X^T X + \lambda I)^{-1} X^T y
\end{aligned}
$$

Now, let's calculate the expected value of the ridge regression estimator:

$$
\begin{aligned}
\mathbb{E}(\beta^*_{\text{ridge}}) &= \mathbb{E}((X^T X + \lambda I)^{-1} X^T y) \\
&= \mathbb{E}((X^T X + \lambda I)^{-1} X^T (X\beta + \varepsilon)) \\
&= \mathbb{E}((X^T X + \lambda I)^{-1} X^T X\beta) \\
&= (X^T X + \lambda I)^{-1} X^T X\beta
\end{aligned}
$$

Therefore, the expected value of the ridge estimator is generally different than β and, consequently, biased. It can be equal to β and unbiased, only if $\lambda = 0$, in which case the ridge estimator becomes exactly the OLS estimator.

SVD decomposition

Given the Singular Value Decomposition (SVD) $X = UDV^T$, we can rewrite the expression for β^*_{ridge}:

$$
\begin{aligned}
\beta^*_{\text{ridge}} &= (X^T X + \lambda I)^{-1} X^T y \\
&= ((UDV^T)^T UDV^T + \lambda I)^{-1} (UDV^T)^T y \\
&= ((VDU^T) UDV^T + \lambda I)^{-1} (VDU^T) y \\
&= (VD^2V^T + \lambda I)^{-1} (VDU^T) y \\
&= V(D^2 + \lambda I)^{-1} V^T (VDU^T) y \\
&= V(D^2 + \lambda I)^{-1} DU^T y
\end{aligned}
$$

This solution is particularly useful when the matrix X is ill-conditioned or nearly singular,. In this case, the SVD decomposition provides a numerically stable way to solve the ridge regression problem without directly inverting a potentially singular matrix. Additionally, SVD can be more computationally efficient for large datasets compared to directly computing the inverse of $X^T X + \lambda I$, which is no longer necessary through this method.

Comparison between OLS variance and Ridge variance

First we calculate the variance of the Ridge estimator:

$$
\begin{aligned}
\mathbb{V}(\beta^*_{\text{ridge}}) &= \mathbb{V}((X^T X + \lambda I)^{-1} X^T y) \\
&= (X^T X + \lambda I)^{-1} X^T \mathbb{V}(y) ((X^T X + \lambda I)^{-1} X^T)^T \\
&= (X^T X + \lambda I)^{-1} X^T \mathbb{V}(y) X (X^T X + \lambda I)^{-1} \\
&= (X^T X + \lambda I)^{-1} X^T \mathbb{V}(\varepsilon) X (X^T X + \lambda I)^{-1} \\
&= \sigma^2 (X^T X + \lambda I)^{-1} X^T X (X^T X + \lambda I)^{-1} \\
&= \sigma^2 ((UDV^T)^T UDV^T + \lambda I)^{-1} (UDV^T)^T UDV^T ((UDV^T)^T UDV^T + \lambda I)^{-1} \\
&= \sigma^2 ((VDU^T) UDV^T + \lambda I)^{-1} (VDU^T) UDV^T ((VDU^T) UDV^T + \lambda I)^{-1} \\
&= \sigma^2 V (D^2 + \lambda I)^{-1} D^2 (D^2 + \lambda I)^{-1} V^T \\
&= \sum_{i=1}^{\text{rank}(X)} \frac{d_i^2 \sigma^2}{(d_i^2 + \lambda)^2} v_i v_i^T
\end{aligned}
$$

We know that the OLS estimator is the Ridge estimator for $\lambda = 0$, thus:

$$
\mathbb{V}(\beta^*_{\text{OLS}}) = \sum_{i=1}^{\text{rank}(X)} \frac{\sigma^2}{d_i^2} v_i v_i^T
$$

Hence, it is apparent that for $\lambda \geq 0$, the variance of the Ridge estimator is smaller than or equal to the variance of the OLS estimator, that is $\mathbb{V}(\beta^*_{\text{ridge}}) \leq \mathbb{V}(\beta^*_{\text{OLS}})$.

Effect of the regularization parameter

Recalling the expression for the expected value of the Ridge estimator:

$$\begin{aligned}
\mathbb{E}(\beta^*_{\text{ridge}}) &= (X^TX + \lambda I)^{-1}X^TX\beta \\
&= ((UDV^T)^TUDV^T + \lambda I)^{-1}(UDV^T)^TUDV^T\beta \\
&= ((VDU^T)UDV^T + \lambda I)^{-1}(VDU^T)UDV^T\beta \\
&= (VD^2V^T + \lambda I)^{-1}VD^2V^T\beta \\
&= V(D^2 + \lambda I)^{-1}D^2V^T\beta \\
&= \sum_{i=1}^{\text{rank}(X)} \frac{d_i^2}{d_i^2 + \lambda} v_i v_i^T \beta
\end{aligned}$$

One can write the expression for the bias and variance of the ridge estimator:

$$\begin{aligned}
\text{b}(\beta^*_{\text{ridge}}, \beta) &= \mathbb{E}(\beta^*_{\text{ridge}}) - \beta \\
&= \sum_{i=1}^{\text{rank}(X)} \frac{d_i^2}{d_i^2 + \lambda} v_i v_i^T \beta - \beta
\end{aligned}$$

$$\mathbb{V}(\beta^*_{\text{ridge}}) = \sum_{i=1}^{\text{rank}(X)} \frac{d_i^2\sigma^2}{(d_i^2 + \lambda)^2} v_i v_i^T$$

Inference suggests that for small λ values, the ridge estimator closely resembles the OLS estimator, exhibiting low bias but high variance. Conversely, as λ increases, the bias of the ridge estimator amplifies in magnitude while its variance diminishes ($\lambda \to +\infty \Rightarrow \mathbb{V}(\beta^*_{\text{ridge}}) \to 0$).

Derivation of Ridge estimator and OLS estimator expression under $X^TX = I$

As already seen on previous sections:

$$\beta^*_{\text{ridge}} = (X^TX + \lambda I)^{-1}X^Ty$$

$$\beta^*_{\text{OLS}} = (X^TX)^{-1}X^Ty$$

Assuming $X^TX = I$, these last expressions become:

$$
\begin{aligned}
\beta^*_{\text{ridge}} &= (X^TX + \lambda I)^{-1}X^Ty \\
&= (I + \lambda I)^{-1}X^Ty \\
&= ((1+\lambda)I)^{-1}X^Ty
\end{aligned}
$$

$$
\begin{aligned}
\beta^*_{\text{OLS}} &= (X^TX)^{-1}X^Ty \\
&= X^Ty
\end{aligned}
$$

Therefore:

$$\beta^*_{\text{ridge}} = \frac{\beta^*_{\text{OLS}}}{1+\lambda}$$

Elastic Net

Analyzing the Elastic Net estimator $\beta^*_{\text{ElNet}} = \text{argmin}_\beta \parallel y - X\beta \parallel_2^2 + \lambda_2 \parallel \beta \parallel_2^2 + \lambda_1 \parallel \beta \parallel_1$ and denoting $f(\beta) = \parallel y - X\beta \parallel_2^2 + \lambda_2 \parallel \beta \parallel_2^2 + \lambda_1 \parallel \beta \parallel_1$, we have:

$$
\begin{aligned}
f(\beta) &= \parallel y - X\beta \parallel_2^2 + \lambda_2 \parallel \beta \parallel_2^2 + \lambda_1 \parallel \beta \parallel_1 \\
&= (y - X\beta)^T(y - X\beta) + \lambda_2\beta^T\beta + \lambda_1 \parallel \beta \parallel_1 \\
&= (y^T - \beta^TX^T)(y - X\beta) + \lambda_2\beta^T\beta + \lambda_1 \parallel \beta \parallel_1 \\
&= (y^Ty - y^TX\beta - \beta^TX^Ty + \beta^TX^TX\beta) + \lambda_2\beta^T\beta + \lambda_1 \parallel \beta \parallel_1 \\
\therefore \partial f(\beta) &= 2X^TX\beta - 2X^Ty + 2\lambda_2\beta \pm \lambda_1 \\
\Rightarrow \beta^*_{\text{ElNet}} &= (X^TX + \lambda_2 I)^{-1}(X^Ty \mp \frac{\lambda_1}{2})
\end{aligned}
$$

Assuming $X^TX = I$, this last expression becomes:

$$
\begin{aligned}
\beta^*_{\text{ElNet}} &= (X^TX + \lambda_2 I)^{-1}X^T(y \mp \frac{\lambda_1}{2}) \\
&= (I + \lambda_2 I)^{-1}X^T(y \mp \frac{\lambda_1}{2}) \\
&= ((1+\lambda_2)I)^{-1}X^T(y \mp \frac{\lambda_1}{2})
\end{aligned}
$$

Therefore:

$$\beta^*_{\text{ElNet}} = \frac{\beta^*_{\text{OLS}} \mp \frac{\lambda_1}{2}}{1+\lambda_2}$$