Introduction Supervised Learning

Theoretical questions

OLS (Ordinary Least Squares)

We have seen that the OLS estimator is equal to $\beta^* = (X^T X)^{-1} X^T y$ which can be rewritten as $\beta^* = Hy$. Let $\hat{\beta} = Cy$ be another linear unbiased estimator of β where C is a $d \times n$ matrix, e.g., $C = H + D$ where D is a non-zero matrix.

- Demonstrate that OLS is the estimator with the smallest variance: compute $E[\hat{\beta}]$ and $Var(\hat{\beta}) = E[(\hat{\beta} - E[\hat{\beta}])(\hat{\beta} - E[\hat{\beta}])^T]$ and show when and why $Var(\beta^*) < Var(\hat{\beta})$. Which assumption of OLS do we need to use?

Answer

To demonstrate that the OLS estimator has the smallest variance, we need to use the Gauss-Markov Theorem, which states that under the assumptions of the classical linear regression model, the OLS estimator has the smallest variance among all unbiased linear estimators. These assumptions are:

1. Linearity of parameters
2. Random sampling
3. No perfect multicollinearity
4. Zero conditional mean (The error term has a zero conditional mean given any value of the explanatory variables)
5. Homoscedasticity (constant variance) of the errors

Given that $\beta^* = Hy$ and $\hat{\beta} = Cy$, where $C = H + D$, and assuming that H is the matrix which gives us the OLS estimator, we have that $H = (X^T X)^{-1} X^T$.

For β^*:

$$E[\beta^*] = E[Hy] = HE[y] = HX\beta$$

Since $H = (X^T X)^{-1} X^T$, we have $HX = I$, where I is the identity matrix, so $E[\beta^*] = \beta$.

For $\hat{\beta}$:

$$E[\hat{\beta}] = E[Cy] = CE[y] = CX\beta$$

To be an unbiased estimator, $E[\hat{\beta}]$ must equal β, which implies that $CX = I$.

For the variance of $\hat{\beta}$:

$$Var(\hat{\beta}) = E[(\hat{\beta} - E[\hat{\beta}])(\hat{\beta} - E[\hat{\beta}])^T]$$

$$Var(\hat{\beta}) = E[(Cy - CX\beta)(Cy - CX\beta)^T]$$

$$Var(\hat{\beta}) = CE[(y - X\beta)(y - X\beta)^T]C^T$$

Since $E[(y - X\beta)(y - X\beta)^T]$ is the variance of y, which we can denote as $\sigma^2 I$ under the assumption of homoscedasticity and independence, we have:

$$Var(\hat{\beta}) = \sigma^2 CC^T$$

For the variance of the OLS estimator:

$$Var(\beta^*) = \sigma^2 HH^T$$

And since $H = (X^T X)^{-1} X^T$, we have:

$$Var(\beta^*) = \sigma^2 (X^T X)^{-1}$$

We need to show that $Var(\beta^*) < Var(\hat{\beta})$. Since $C = H + D$, we have:

$$Var(\hat{\beta}) = \sigma^2(H + D)(H + D)^T$$
$$Var(\hat{\beta}) = \sigma^2(HH^T + HD^T + DH^T + DD^T)$$

Given that $Var(\beta^*) = \sigma^2 HH^T$, $Var(\hat{\beta}) > Var(\beta^*)$ because $HD^T + DH^T + DD^T$ is a positive semi-definite matrix, and adding this to HH^T will give a matrix with larger diagonal elements (variances), assuming D is not a matrix of zeros (which would violate the assumption that D is a non-zero matrix). This shows that the variance of β^* is less than the variance of $\hat{\beta}$, making β^* the estimator with the smallest variance among all linear unbiased estimators.

Ridge Regression

Suppose that both y and the columns of x are centered (y_c and x_c) so that we do not need the intercept β_0. In this case, the matrix x_c has d (rather than $d+1$) columns. We can thus write the criterion for ridge regression as:

$$\beta^*_{\text{ridge}} = \arg\min_{\beta} \left\{(y_c - x_c\beta)^T(y_c - x_c\beta) + \lambda\|\beta\|^2\right\}$$

- Show that the estimator of ridge regression is biased (that is $E[\beta^*_{\text{ridge}}] \neq \beta$).

Answer: The ridge regression estimator β^*_{ridge} is found by minimizing the penalized residual sum of squares:

$$\beta^*_{\text{ridge}} = \arg\min_{\beta} \left\{(y_c - x_c\beta)^T(y_c - x_c\beta) + \lambda\|\beta\|^2\right\}$$

The solution to this minimization problem is:

$$\beta^*_{\text{ridge}} = (x_c^T x_c + \lambda I)^{-1} x_c^T y_c$$

The expectation of β^*_{ridge}:

$$E[\beta^*_{\text{ridge}}] = E\left[(x_c^T x_c + \lambda I)^{-1} x_c^T y_c\right]$$

Since $y_c = x_c\beta + \epsilon$, where ϵ is the error term, we can substitute y_c into the expectation:

$$E[\beta^*_{\text{ridge}}] = E\left[(x_c^T x_c + \lambda I)^{-1} x_c^T (x_c\beta + \epsilon)\right]$$

Distributing x_c^T we get:

$$E[\beta^*_{\text{ridge}}] = (x_c^T x_c + \lambda I)^{-1} x_c^T x_c\beta + (x_c^T x_c + \lambda I)^{-1} x_c^T E[\epsilon]$$

Assuming $E[\epsilon] = 0$, this simplifies to:

$$E[\beta^*_{\text{ridge}}] = (x_c^T x_c + \lambda I)^{-1} x_c^T x_c\beta$$

$E[\beta^*_{\text{ridge}}]$ will not equal β unless $\lambda = 0$, because the presence of λI in the inverse term modifies the relation between $x_c^T x_c$ and β. Specifically, when $\lambda > 0$, the term $(x_c^T x_c + \lambda I)^{-1} x_c^T x_c$ acts as a shrinkage operator, pulling the estimates of β towards zero.

Therefore, the estimator β^*_{ridge} is biased because the expectation of the estimator does not equal the true parameter value, i.e., $E[\beta^*_{\text{ridge}}] \neq \beta$.

- Recall that the SVD decomposition is $x_c = UDV^T$. Write down by hand the solution β^*_{ridge} using the SVD decomposition. When is it useful using this decomposition? Hint: do you need to invert a matrix?

Answer: Substituting the SVD of x_c into the expression for β^*_{ridge} we get:

$$\beta^*_{\text{ridge}} = (VDU^TUDV^T + \lambda I)^{-1}VDU^Ty_c$$

Since $U^TU = I$ and $VV^T = I$, where I is the identity matrix, we can simplify this to:

$$\beta^*_{\text{ridge}} = (VD^2V^T + \lambda I)^{-1}VDU^Ty_c$$

We can take advantage of the diagonal structure of D^2 and the orthogonal matrices U and V to compute the ridge estimator more efficiently:

$$\beta^*_{\text{ridge}} = V(D^2 + \lambda I)^{-1}DV^Ty_c$$

This is possible because the inverse of a diagonal matrix $D^2 + \lambda I$ is easy to compute; it's simply the reciprocal of the diagonal elements.

Using the SVD decomposition is particularly useful in ridge regression for a couple of reasons:

1. Numerical stability: When $x_c^T x_c$ is close to singular or ill-conditioned (which can happen when multicollinearity is present or when d is large), directly computing its inverse as required in the standard ridge regression formula can be numerically unstable. The SVD approach avoids this problem because the inverse of a diagonal matrix (with the regularization term added) is always well-conditioned.
2. Computational efficiency: Computing the inverse of a matrix is computationally expensive and can be slow if the matrix is large. However, because SVD provides us with matrices U, D, and V, where D is diagonal, we only need to compute the inverse of the diagonal elements of $D^2 + \lambda I$, which is straightforward and fast.

- Remember that $Var(\beta^*_{\text{OLS}}) = \sigma^2(X^TX)^{-1}$. Show that $Var(\beta^*_{\text{OLS}}) \geq Var(\beta^*_{\text{ridge}})$.

Answer: The variance of the OLS estimator is:

$$Var(\beta^*_{\text{OLS}}) = \sigma^2(X^TX)^{-1}$$

For the ridge regression estimator, the solution can be written using the SVD as $\beta^*_{\text{ridge}} = V(D^2+\lambda I)^{-1}DV^Ty$. The variance of the ridge regression estimator is:

$$Var(\beta^*_{\text{ridge}}) = \sigma^2V(D^2 + \lambda I)^{-2}V^T$$

Now, we need to show that:

$$\sigma^2(X^TX)^{-1} \geq \sigma^2V(D^2 + \lambda I)^{-2}V^T$$

Using the SVD of X, we have $X = UDV^T$, so $X^TX = VD^2V^T$. Replacing this into the variance of the OLS estimator gives us:

$$Var(\beta^*_{\text{OLS}}) = \sigma^2(VD^2V^T)^{-1}$$

Multiplying both sides by VD^2V^T to remove the inverse, we get:

$$VD^2V^T \cdot Var(\beta^*_{\text{OLS}}) = \sigma^2 I$$

Since VD^2V^T is a positive semi-definite matrix, $Var(\beta^*_{\text{OLS}})$ must also be a positive semi-definite matrix. This implies that:

$$VD^2V^T \cdot Var(\beta^*_{\text{OLS}}) \geq \sigma^2 I$$

Similarly, for ridge regression, we have:

$$V(D^2 + \lambda I)^{-2} \cdot Var(\beta^*_{\text{ridge}}) = \sigma^2 I$$

Multiplying both sides by $(D^2 + \lambda I)^2$ we get:

$$Var(\beta^*_{\text{ridge}}) = \sigma^2 V(D^2 + \lambda I)^{-2} V^T$$

Given that $(D^2 + \lambda I)$ is a diagonal matrix with each diagonal element $d_i^2 + \lambda$ being greater than d_i^2, the inverse of $(D^2 + \lambda I)$ will have diagonal elements less than or equal to the inverse of D^2. Thus:

$$V(D^2 + \lambda I)^{-2} V^T \leq VD^{-2}V^T$$

Multiplying through by σ^2 we find:

$$\sigma^2 V(D^2 + \lambda I)^{-2} V^T \leq \sigma^2 V(D^2 V^T)^{-1}$$

$$Var(\beta^*_{\text{ridge}}) \leq Var(\beta^*_{\text{OLS}})$$

Therefore, the variance of the OLS estimator is greater than or equal to the variance of the ridge regression estimator.

- When λ increases what happens to the bias and to the variance? Hint: Compute MSE $= E[(y_0 - x_0^T \beta^*_{\text{ridge}})^2]$ at the test point (x_0, y_0) with $y_0 = x_0^T \beta + \epsilon_0$ being the true model and β^*_{ridge} the ridge estimate.

Answer: To examine what happens to the bias and variance as λ increases, let's consider the mean squared error (MSE) at the test point (x_0, y_0). The MSE can be decomposed into the sum of the variance and the square of the bias, plus the variance of the error term:

$$MSE = Var(x_0^T \beta^*_{\text{ridge}}) + [Bias(x_0^T \beta^*_{\text{ridge}})]^2 + Var(\epsilon_0)$$

Given that $y_0 = x_0^T \beta + \epsilon_0$, where x_0 is a new observation and ϵ_0 is the error term associated with the new observation, the bias of the ridge estimate at this test point is:

$$Bias(x_0^T \beta^*_{\text{ridge}}) = E[x_0^T \beta^*_{\text{ridge}}] - x_0^T \beta$$

As λ increases, the ridge estimator β^*_{ridge} will shrink towards zero. This increases the bias term $E[x_0^T \beta^*_{\text{ridge}}] - x_0^T \beta$ since the expected value of $x_0^T \beta^*_{\text{ridge}}$ will be further from $x_0^T \beta$.

Regarding variance, the ridge estimate's variance is given by:

$$Var(\beta^*_{\text{ridge}}) = \sigma^2 V(D^2 + \lambda I)^{-2} V^T$$

As λ increases, the diagonal elements of the matrix $(D^2 + \lambda I)$ increase, which leads to a decrease in the diagonal elements of the inverse matrix $(D^2 + \lambda I)^{-2}$. Consequently, the variance $Var(x_0^T \beta^*_{\text{ridge}})$ decreases.

As λ increases: - The bias $Bias(x_0^T \beta^*_{\text{ridge}})$ increases because the ridge regression estimate is shrunk further towards zero, causing it to deviate more from the true parameter β. - The variance $Var(x_0^T \beta^*_{\text{ridge}})$ decreases because the regularization term λ penalizes the magnitude of the coefficients, thus reducing the estimator's sensitivity to fluctuations in the training data.

The MSE will balance these two effects, and the optimal value of λ (in terms of predictive performance) is one that achieves a good trade-off between bias and variance. This is the essence of the bias-variance trade-off in the context of ridge regression.

- Show that $\beta^*_{\text{ridge}} = \frac{\beta^*_{\text{OLS}}}{1+\lambda}$ when $X^T X = I_d$

Answer: The OLS estimator β^*_{OLS} is given by:

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

The ridge regression estimator β^*_{ridge} is given by:

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

Since $X^T X = I_d$, the OLS estimator simplifies to:

$$\beta^*_{\text{OLS}} = I_d^{-1} X^T y$$
$$\beta^*_{\text{OLS}} = X^T y$$

Now, considering the ridge regression estimator:

$$\beta^*_{\text{ridge}} = (I_d + \lambda I_d)^{-1} X^T y$$

Since $I_d + \lambda I_d$ is a diagonal matrix with each diagonal entry equal to $1 + \lambda$, its inverse is a diagonal matrix with each diagonal entry equal to $\frac{1}{1+\lambda}$. Thus, we have:

$$\beta^*_{\text{ridge}} = \frac{1}{1+\lambda} I_d X^T y$$

Since $I_d X^T y$ is just $X^T y$, we obtain:

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

It looks like you've provided a description of the Elastic Net regularization method and its advantages over using Lasso or Ridge regularization individually. Here's the transcription of the content and the benefits of Elastic Net:

Elastic Net

Using the previous notation, we can also combine Ridge and Lasso in the so-called Elastic Net regularization:

$$\beta^*_{\text{ENet}} = \arg\min_{\beta}\{(y_c - x_c\beta)^T(y_c - x_c\beta) + \lambda_2\|\beta\|^2 + \lambda_1\|\beta\|_1\}$$

Calling $\alpha = \frac{\lambda_2}{\lambda_1+\lambda_2}$, solving the previous Eq. is equivalent to:

$$\beta^*_{\text{ENet}} = \arg\min_{\beta}\{(y_c - x_c\beta)^T(y_c - x_c\beta) + \lambda(\alpha\sum_{j=1}^{d}\beta_j^2 + (1-\alpha)\sum_{j=1}^{d}|\beta_j|)\}$$

- This regularization overcomes some of the limitations of the Lasso, notably:
  - If $d > N$ Lasso can select at most N variables $\rightarrow$ ENet removes this limitation.
  - If a group of variables are highly correlated, Lasso randomly selects only one variable $\rightarrow$ with ENet correlated variables have a similar value (grouped).
  - Lasso solution paths tend to vary quite drastically $\rightarrow$ ENet regularizes the paths.
  - If $N > d$ and there is high correlation between the variables, Ridge tends to have a better performance in prediction $\rightarrow$ ENet combines Ridge and Lasso to have better (or similar) prediction accuracy with less (or more grouped) variables.

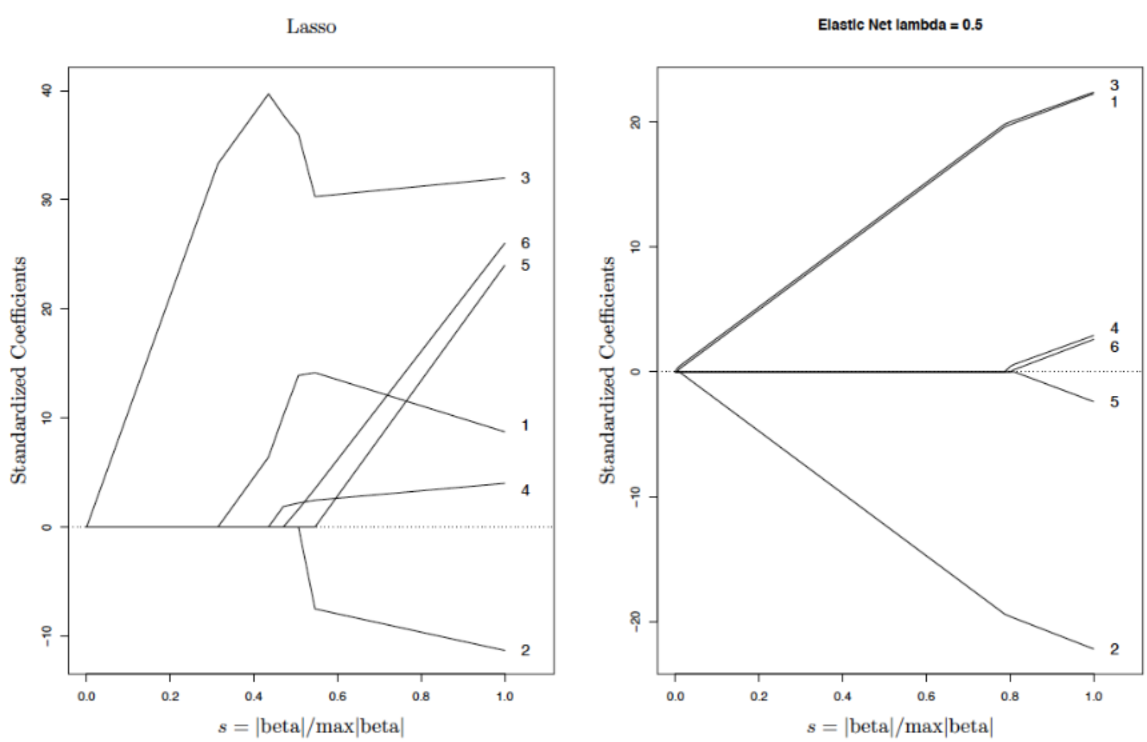

Figure 1: alt text

- Compute by hand the solution of Eq.2 supposing that $X_c^T X_c = I_d$ and show that the solution is:

$$\beta^*_{\text{ENet}} = \frac{(\beta^*_{\text{OLS}})_j \pm \frac{\lambda_1}{2}}{1+\lambda_2}$$

Answer :

To arrive at the Elastic Net solution using a thresholding approach, we start with the objective function given in Equation 2, taking into consideration that $X_c^T X_c = I_d$ (the identity matrix):

$$\beta^{ENet} = \arg\min_{\beta}\{(y_c - X_c\beta)^T(y_c - X_c\beta) + \lambda_2||\beta||_2^2 + \lambda_1||\beta||_1\}$$

Because $X_c^T X_c = I_d$, the objective function simplifies to:

$$\beta^{ENet} = \arg\min_{\beta}\{||y_c - X_c\beta||_2^2 + \lambda_2||\beta||_2^2 + \lambda_1||\beta||_1\}$$

The solution for the Ridge regression part (where $\lambda_1 = 0$) with orthogonal predictors is:

$$\beta^{Ridge} = \frac{\beta^{OLS}}{1+\lambda_2}$$

For Lasso regression, which uses an L1 penalty, we apply soft-thresholding to each coefficient. The soft-thresholding function for a given j-th coefficient, when $X_c^T X_c = I_d$, is defined as:

$$S_{\lambda_1}((\beta^{OLS})_j) = \text{sign}((\beta^{OLS})_j)(|(\beta^{OLS})_j| - \frac{\lambda_1}{2})_+$$

Here, $(x)_+$ means $\max(0, x)$, and $\text{sign}(x)$ is the sign function, which is $+1$ for $x > 0$, 0 for $x = 0$, and -1 for $x < 0$.

In the Elastic Net, which combines both L1 and L2 penalties, the solution for each coefficient incorporates both the soft-thresholding from Lasso and the shrinkage from Ridge. The soft-thresholding operator is applied first, followed by the shrinkage due to the Ridge penalty:

$$\beta_j^{ENet} = \frac{S_{\lambda_1}((\beta^{OLS})_j)}{1+\lambda_2}$$

Substituting the soft-thresholding function we get:

$$\beta_j^{ENet} = \frac{\text{sign}((\beta^{OLS})_j)(|(\beta^{OLS})_j| - \frac{\lambda_1}{2})_+}{1+\lambda_2}$$

Now, we must account for the positive and negative scenarios depending on the sign of $(\beta^{OLS})_j$. If $(\beta^{OLS})_j > \frac{\lambda_1}{2}$, then $\text{sign}((\beta^{OLS})_j) = +1$, and if $(\beta^{OLS})_j < -\frac{\lambda_1}{2}$, then $\text{sign}((\beta^{OLS})_j) = -1$. If $|(\beta^{OLS})_j| \leq \frac{\lambda_1}{2}$, then the soft-thresholding output will be zero.

Thus, the final formula for each non-zero β_j^{ENet} is:

$$\beta_j^{ENet} = \frac{(\beta_j^{OLS}) \pm \frac{\lambda_1}{2}}{1+\lambda_2}$$

The $\pm$ depends on the sign of the original OLS coefficient $(\beta^{OLS})_j$, which reflects the Lasso's characteristic of either subtracting or adding $\frac{\lambda_1}{2}$ after thresholding, and then applying the Ridge shrinkage of $\frac{1}{1+\lambda_2}$.