IMA205 TP3 Theoretical Questions

Luiz Fernando Moreira Teixeira

March 18, 2025

OLS

Questions

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

Answers

Let the linear model be

$$y = X\beta + \varepsilon, \quad E[\varepsilon] = 0, \quad \mathrm{Var}(\varepsilon) = \sigma^2 I.$$

A linear estimator is

$$\tilde{\beta} = C\,y.$$

For unbiasedness we require

$$E[\tilde{\beta}] = C\,X\beta = \beta \quad \Longrightarrow \quad C\,X = I.$$

Its variance is

$$\mathrm{Var}(\tilde{\beta}) = \sigma^2\,C\,C^T.$$

The OLS estimator is

$$\beta^* = (X^TX)^{-1}X^Ty,$$

with variance

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

Any other unbiased estimator can be written as

$$C = (X^TX)^{-1}X^T + D, \quad \text{with } D\,X = 0.$$

Then one may show that

$$\mathrm{Var}(\tilde{\beta}) - \mathrm{Var}(\beta^*) = \sigma^2\,D\,D^T \succeq 0.$$

Thus, the OLS estimator has the smallest variance among all linear unbiased estimators.

Ridge regression

Questions

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

3. Recall that the SVD decomposition is $\mathbf{x_c} = UDV^T$. 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?

4. Remember that $\mathrm{Var}(\beta^*_{\text{OLS}}) = \sigma^2(\mathbf{x}^T\mathbf{x})^{-1}$. Show that $\mathrm{Var}(\beta^*_{\text{OLS}}) \geq \mathrm{Var}(\beta^*_{\text{ridge}})$.

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

6. Show that $\beta^*_{\text{ridge}} = \frac{\beta^*_{\text{OLS}}}{1+\lambda}$ when $\mathbf{x_c}^T\mathbf{x_c} = I_d$.

Answers:

- **2)**
The ridge estimator is
$$\beta^*_{\text{ridge}} = (X_c^T X_c + \lambda I)^{-1} X_c^T y_c.$$
Taking expectations,
$$E[\beta^*_{\text{ridge}}] = (X_c^T X_c + \lambda I)^{-1} X_c^T X_c \, \beta.$$
Since $(X_c^T X_c + \lambda I)^{-1} X_c^T X_c \neq I$ for $\lambda > 0$, the estimator is biased.

- **3)**
Let $X_c = UDV^T$ so that
$$X_c^T X_c = V D^T D V^T.$$
Then,
$$\beta^*_{\text{ridge}} = V (D^T D + \lambda I)^{-1} D^T U^T y_c.$$
This form is useful when $X_c^T X_c$ is singular or ill–conditioned, as the addition of λI guarantees invertibility.

- **4)**
For OLS,
$$\text{Var}(\beta^*_{\text{OLS}}) = \sigma^2 (X_c^T X_c)^{-1}.$$
For ridge regression, one finds
$$\text{Var}(\beta^*_{\text{ridge}}) = \sigma^2 (X_c^T X_c + \lambda I)^{-1} X_c^T X_c \, (X_c^T X_c + \lambda I)^{-1},$$
which is smaller (in the positive semidefinite sense).

- **5)**
As λ increases, the bias increases (since $E[\beta^*_{\text{ridge}}]$ shrinks toward zero), while the variance decreases, lowering the overall mean squared error at an appropriate choice of λ.

- **6)**
If $X_c^T X_c = I$ then
$$\beta^*_{\text{ridge}} = (I + \lambda I)^{-1} X_c^T y_c = \frac{1}{1+\lambda} \, \beta^*_{\text{OLS}}.$$

Elastic Net

Questions

7. Compute by hand the solution of Eq.2 supposing that $\mathbf{x_c}^T \mathbf{x_c} = I_d$ and show that the solution is:

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

Answers

The Elastic Net estimator is defined as

$$\beta^*_{\text{ElNet}} = \arg\min_{\beta} \left\{ (y_c - X_c\beta)^T (y_c - X_c\beta) + \lambda_2 \|\beta\|_2^2 + \lambda_1 \|\beta\|_1 \right\}.$$

Define

$$\lambda = \lambda_1 + \lambda_2 \quad \text{and} \quad \alpha = \frac{\lambda_2}{\lambda_1 + \lambda_2}.$$

Then the problem can be written as

$$\min_{\beta} \left\{ (y_c - X_c\beta)^T (y_c - X_c\beta) + \lambda \Big[\alpha \, \|\beta\|_2^2 + (1-\alpha) \, \|\beta\|_1 \Big] \right\}.$$

Assume $X_c^T X_c = I$ so that the OLS solution is $\beta^*_{\text{OLS}} = X_c^T y_c$. Then, the minimization decouples for each coordinate j:

$$\min_{\beta_j} \left\{ (\beta_{\text{OLS},j} - \beta_j)^2 + \lambda_2 \, \beta_j^2 + \lambda_1 \, |\beta_j| \right\}.$$

Collecting the quadratic terms gives

$$(1+\lambda_2)\beta_j^2 - 2\beta_{\text{OLS},j}\beta_j + \lambda_1 \, |\beta_j|.$$

The minimizer is obtained via soft–thresholding:

$$\beta_{\text{ElNet},j} = \text{sgn}(\beta_{\text{OLS},j}) \, \max\left\{ |\beta_{\text{OLS},j}| - \frac{\lambda_1}{2(1+\lambda_2)}, \, 0 \right\}.$$

This is the closed–form solution in the special case of orthonormal predictors.