

QBUS6810

Statistical Learning and Data Mining

Tutorial 6 (Written Problems)

Question 1

Show that the OLS estimator is unbiased, i.e., derive the following:

$$E\hat{\beta} = \beta.$$

Treat the x values as fixed (i.e. non-random) and use the formula for the OLS estimator.

Solution: Recall that the expected value of the error terms in the MLR model is zero. We will make use of the following formulas (and all the corresponding notation) from Lecture 3:

$$\begin{aligned}\hat{\beta} &= (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} \quad \text{and} \\ \mathbf{y} &= \mathbf{X}\beta + \epsilon.\end{aligned}$$

In the following expected value calculations, non-random matrixes are treated as constants, which we can be factored out of the expected values. We have:

$$\begin{aligned}E\hat{\beta} &= E\left[(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}\right] \\ &= E\left[(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T (\mathbf{X}\beta + \epsilon)\right] \\ &= E\left[(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{X}\beta\right] + E\left[(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \epsilon\right] \\ &= (\mathbf{X}^T \mathbf{X})^{-1} (\mathbf{X}^T \mathbf{X})\beta + (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T E\epsilon \\ &= \beta.\end{aligned}$$

Question 2

Consider a method for learning the true regression function, f , in the additive error model. It is known (this will be further discussed in the lectures later in the semester) that the expected value of the amount by which the training MSE underestimates the corresponding test MSE is given by

$$\frac{2}{n} \sum_{i=1}^n \text{Cov}(\hat{Y}_i, Y_i), \quad (1)$$

where Y_i are the response values in the training data, and $\hat{Y}_i = \hat{f}(\mathbf{x}_i)$ are the corresponding fitted values of the learning method.

Show that the quantity in display (1) equals $2\sigma^2/k$ for the k-nearest neighbours regression method. Thus, the training error underestimates the test error by the largest amount in the case $k = 1$, in which the training data is fitted perfectly.

Treat the x values as fixed (i.e. non-random).

Solution: Focus on a particular i and let $\mathbf{x}_{i_1}, \mathbf{x}_{i_2}, \dots, \mathbf{x}_{i_k}$ be the k closest neighbors of \mathbf{x}_i in the training set (i.e. the k points in $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$ that are the closest to \mathbf{x}_i). Note that \mathbf{x}_i is itself one of these k points. Also note that random variables Y_i and Y_{i_j} are independent (and thus, uncorrelated) for $i \neq i_j$. We have:

$$\begin{aligned} \text{Cov}(Y_i, \hat{Y}_i) &= \text{Cov}\left(Y_i, \frac{1}{k}[Y_{i_1} + Y_{i_2} + \dots + Y_{i_k}]\right) \\ &= \text{Cov}\left(Y_i, \frac{Y_{i_1}}{k}\right) + \text{Cov}\left(Y_i, \frac{Y_{i_2}}{k}\right) + \dots + \text{Cov}\left(Y_i, \frac{Y_{i_k}}{k}\right) \\ &= \frac{1}{k} \text{Cov}(Y_i, Y_{i_1}) + \frac{1}{k} \text{Cov}(Y_i, Y_{i_2}) + \dots + \frac{1}{k} \text{Cov}(Y_i, Y_{i_k}) \\ &= \frac{1}{k} \text{Cov}(Y_i, Y_i) \\ &= \frac{1}{k} \text{Var}(Y_i) \\ &= \frac{\sigma^2}{k}. \end{aligned}$$

Therefore,

$$\frac{2}{n} \sum_{i=1}^n \text{Cov}(\hat{Y}_i, Y_i) = \frac{2}{n} \sum_{i=1}^n \frac{\sigma^2}{k} = \frac{2}{n} n \frac{\sigma^2}{k} = \frac{2\sigma^2}{k}.$$

Question 3

Let y_1, \dots, y_n be a sample from a distribution with the density function $p(y; \theta) = \theta y^{\theta-1}$ for $0 < y < 1$, where $\theta > 0$.

Find $\hat{\theta}$, the maximum likelihood estimator of θ .

Compute $\hat{\theta}$ for the sample $y_1 = 0.35$, $y_2 = 0.28$, $y_3 = 0.91$.

Solution: The likelihood function is

$$\begin{aligned}\ell(\theta) &= p(y_1; \theta)p(y_2; \theta) \dots p(y_n; \theta) \\ &= \prod_{i=1}^n \theta y_i^{\theta-1} \\ &= \theta^n \prod_{i=1}^n y_i^{\theta-1}.\end{aligned}$$

Taking the natural log:

$$L(\theta) = \log(\ell(\theta)) = n \log(\theta) + (\theta - 1) \sum_{i=1}^n \log(y_i).$$

The first derivative is

$$\frac{dL(\theta)}{d\theta} = \frac{n}{\theta} + \sum_{i=1}^n \log(y_i).$$

The first derivative is zero at $\hat{\theta}$:

$$\frac{n}{\hat{\theta}} + \sum_{i=1}^n \log(y_i) = 0.$$

Thus,

$$\hat{\theta} = \frac{-n}{\sum_{i=1}^n \log(y_i)}.$$

For the sample 0.35, 0.28, 0.91, we have:

$$\hat{\theta} = \frac{-3}{\log(0.35) + \log(0.28) + \log(0.91)} = 1.24.$$

Question 4

Consider the following penalized least-squares estimator, called the *Ridge regression estimator* (to be discussed in Lecture 6):

$$\hat{\beta}_{\text{ridge}} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^p \beta_j^2 \right\}$$

Note that OLS is a special case of Ridge, corresponding to $\lambda = 0$.

Show that if we set $\lambda = \sigma^2/\tau^2$, the ridge regression estimator is the posterior mode (i.e. the MAP estimator) in a Gaussian linear regression model with the prior on the regression coefficients under which β_j are independent $N(0, \tau^2)$, for $j = 1, \dots, p$. Here we are not putting an informative prior on the intercept β_0 (this is equivalent to using a flat prior density for β_0 , i.e., a density that is proportional to the constant 1).

Solution:

The following derivation of the posterior density is almost identical to the one used at the end of Lecture 5, except there is no β_0 component in the prior this time.

Note that random variables Y_i are independent $N(\beta_0 + \sum_{j=1}^p \beta_j x_{ij}, \sigma^2)$. Using the symbol \propto to denote “proportional to” and leaving out multiplicative constants, we have:

$$\begin{aligned} p(\mathbf{y}|\beta) &\propto \prod_{i=1}^n \exp \left\{ -\frac{1}{2\sigma^2} \left(y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij} \right)^2 \right\} \\ &\propto \exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij} \right)^2 \right\}. \end{aligned}$$

Similarly, the prior satisfies

$$\begin{aligned} p(\beta) &\propto \prod_{j=1}^p \exp \left\{ -\frac{\beta_j^2}{2\tau^2} \right\} \\ &\propto \exp \left\{ -\frac{1}{2\tau^2} \sum_{j=1}^p \beta_j^2 \right\}. \end{aligned}$$

Hence, the posterior density has the form

$$\begin{aligned} p(\boldsymbol{\beta}|\mathbf{y}) &\propto \exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij})^2 \right\} \times \exp \left\{ -\frac{1}{2\tau^2} \sum_{j=1}^p \beta_j^2 \right\} \\ &\propto \exp \left\{ -\frac{1}{2\sigma^2} \left[\sum_{i=1}^n (y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij})^2 + \frac{\sigma^2}{\tau^2} \sum_{j=1}^p \beta_j^2 \right] \right\}. \end{aligned}$$

Recall that the MAP estimator is the maximizer of both the posterior density and the log-posterior density. We will work with the log-posterior for convenience. Because $\lambda = \sigma^2/\tau^2$, the logarithm of the posterior density is:

$$\log [p(\boldsymbol{\beta}|\mathbf{y})] = -\frac{1}{2\sigma^2} \left[\sum_{i=1}^n (y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij})^2 + \lambda \sum_{j=1}^p \beta_j^2 \right] + \text{constant},$$

where the “constant” comes from taking the log of the multiplicative factors that were left out in the expressions above.

Thus, the relevant part of the log-posterior consists of the ridge objective function times a negative multiplier. Hence, maximising the log-posterior is equivalent to *minimising* the ridge objective function. It follows that the MAP estimator is equivalent to the ridge estimator.