QBUS6810
Statistical Learning and Data Mining

Tutorial 6 (Written Problems)

Question 1

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

$$E\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta}.$$

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:

$$\begin{aligned}
\widehat{\boldsymbol{\beta}} &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} \qquad \text{and} \\
\boldsymbol{y} &= \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}.
\end{aligned}$$

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:

$$\begin{aligned}
E\widehat{\boldsymbol{\beta}} &= E\Big[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}\Big] \\
&= E\Big[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon})\Big] \\
&= E\Big[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta}\Big] + E\Big[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\varepsilon}\Big] \\
&= (\boldsymbol{X}^T\boldsymbol{X})^{-1}(\boldsymbol{X}^T\boldsymbol{X})\boldsymbol{\beta} + (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T E\boldsymbol{\varepsilon} \\
&= \boldsymbol{\beta}.
\end{aligned}$$

Question 2

Let $y_1, \ldots, y_n$ 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 $\widehat{\theta}$, the maximum likelihood estimator of θ.

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

Solution: The likelihood function is

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

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 $\widehat{\theta}$:

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

Thus,

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

For the sample $0.35, 0.28, 0.99$, we have:

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

Question 3

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

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

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 $j = 1, ..., p$. 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 $N\left(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}\, ,\, \sigma^2\right)$. Using the symbol $\propto$ to denote "proportional to" and leaving out multiplicative constants, we have:

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

Similarly, the prior satisfies

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

Hence, the posterior density has the form

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

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:

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

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.