Extra exercise 3.1
Linear regression by OLS

The purpose with this exercise is partly to learn about the behaviour of the ordinary least squares estimator, but also to set up a simulation experiment that can be used for comparisons between several methods that will be introduced later.

Consider the linear regression model

$$
\begin{aligned}
Y &= f(X) + \varepsilon \qquad (1) \\
&= \beta_0 + \sum_{j=1}^{j=p} X_j \beta_j + \varepsilon,
\end{aligned}
$$

where $p = 15$,
$X = X_1, \ldots, X_p{}^T$,
$\beta_0 = 3$,
$\beta_j = 2$ for $j = 1, \ldots, 5$,
$\beta_j = 1$ for $j = 6, \ldots, 10$,
$\beta_j = 0$ for $j = 11, \ldots, 15$, and
$\mathrm{Var}(\varepsilon) = 25$.

Furthermore, assume that

$$
X = N_p(0, \Sigma), \qquad (2)
$$

i.e. multivariate normal with zero mean and covariance matrix Σ, where all the diagonal elements of Σ are 1 and all non-diagonal elements are 0.8. The Σ is also the correlation matrix with all correlations equal to 0.8.

Generate $N = 20$ observations of the inputs from their multivariate normal distribution (2) and put them into a matrix $\mathbf{X}^{train}$, and then generate $N = 1000$ observations and put them into a matrix $\mathbf{X}^{test}$. You can use the function `mvrnorm` from the R package `MASS`.

Perform now the following simulation experiment with 1000 repetitions, conditioned on the inputs you already have generated:

- For the given $\mathbf{X}^{train}$, simulate the corresponding output vector $\mathbf{y}^{train}$ from the model (1). Together, $\mathbf{y}^{train}$ and $\mathbf{X}^{train}$ constitute the training set.
- Simulate also an output vector $\mathbf{y}^{test}$ conditioned on $\mathbf{X}^{test}$. Together, these constitute a test set.

- Estimate the β vector by ordinary least squares (OLS), for instance by using the `lm` function in R. Furthermore, predict $\mathbf{y}^{test}$ by $\hat{f}(\mathbf{X}^{test}) = \mathbf{X}^{test}\hat{\beta}$.

- Compute the bias, variance and mean squared error of each of the β coefficients by averaging over the simulations, and compare with the theoretical values. Compute also the bias, variance and mean squared error (MSE) of $\hat{f}(\mathbf{X}^{test})$ averaged over the values of the inputs in the test set and over all simulations and finally compute the corresponding prediction error.

Repeat the simulation experiment, first with $N = 100$ and then with $N = 1000$ observations in the training set. Note what happen with the variances and mean squared errors when N in the training sets increases.