HOMEWORK 1

Problem 1. In the lecture, we discussed a simple linear regression model, i.e., linear regression with only one predictor variable. Now, we will consider a linear regression model with k predictor variables and an outcome variable. As discussed in the lecture, this model will involve $k + 1$ parameters (or coefficients) which we are interested in estimating. In order to analyze this (general) multiple linear regression model, it is useful to consider a matrix representation of linear regression.

Suppose the data consists of n observations. The outcome variable can be denoted as a column vector ($n \times 1$ matrix) as follows: Suppose there are k predictor variables and a

$$\mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

constant term (corresponding to the intercept). These can be represented with a $n \times (k + 1)$ matrix as follows. Note that x_{ij} in the matrix represents the i^{th} observation of the j^{th} predictor variable. As mentioned previously, there are $(k + 1)$ coefficients or parameters to

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & \dots & x_{1k} \\ 1 & x_{21} & \dots & x_{2k} \\ \vdots & \vdots & \dots & \vdots \\ 1 & x_{n1} & \dots & x_{nk} \end{pmatrix}$$

be estimated, k coefficients for each of the predictor variables and one coefficient for the intercept. These can be represented with a $(k + 1) \times 1$ matrix, i.e. a column vector with $(k + 1)$ rows as follows.

$$\beta = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_k \end{pmatrix}$$

Finally, let ϵ denote a vector of errors. Then, observe that the multiple linear regression model in matrix form is given by:

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon \tag{1}$$

i. While working with matrices, it always helps to pay attention to the dimensions. Confirm that the dimensions of the matrices on both sides of equation (1) match.

ii. We will now estimate the $(k+1)$ parameters, i.e., the vector β. The procedure is the same as before. We will estimate the parameters by writing down the sum of squared residuals (which is $n \times MSE$) and then minimizing it by taking the first derivative and equating to zero.[1] As a first step, write down the sum of squared residuals (or MSE) in matrix form. Hint: The residuals are given by $\mathbf{e} = \mathbf{Y} - \mathbf{X}\beta$ and sum of squared residuals is given by $\mathbf{e}^T\mathbf{e}$.

iii. Let β^* be the estimated vector of parameters obtained by differentiating the sum of squared residuals obtained in step two with respect to β. Show that $(\mathbf{X}^T\mathbf{X})\beta^* = \mathbf{X}^T\mathbf{Y}$.

iv. Finally, show that,

$$\beta^* = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\left(\mathbf{X}^T\mathbf{Y}\right) \tag{2}$$

Problem 2. A dataset **advertising.csv** has been uploaded on myCourses. Download the dataset. Using Python and using equation (2), perform a linear regression with Sales as the outcome variable and all other columns as predictor variables. Note that equation (2) involves inverting a matrix, so you are free to use appropriate functions and libraries in Python to perform this computation.

Finally, compare the parameter estimates obtained from this procedure with the estimates obtained from a black-box implementation of linear regression in scikit-learn.

[1] We ignore the second order conditions.