Tutorial 4

1. *Matrix Approach to Linear Regression*

   Consider the following simple linear relationship between response y and one input feature, x:

   $$y \approx \beta_0 + \beta_1 x.$$

   Given a data set of n points $(x_1, y_1), ..., (x_n, y_n)$, the model above is then

   $$y_i \approx \beta_0 + \beta_1 x_i, \quad i = 1, ..., n. \quad (*)$$

   To rewrite (*) in matrix form, we have

   $\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$, then the right-hand side of (*) is $\mathbf{X}\boldsymbol{\beta}$.

   The residual sum of squares, $RSS = \sum_{i=1}^{n} [\mathrm{y}_i - (\beta_0 + \beta_1 \mathrm{x}_i)]^2$ is actually equal to

   $$\begin{aligned} & (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \\ &= [y_1 - (\beta_0 + \beta_1 x_1), y_2 - (\beta_0 + \beta_1 x_2), ..., y_n - (\beta_0 + \beta_1 x_n)] \begin{bmatrix} y_1 - (\beta_0 + \beta_1 x_1) \\ y_2 - (\beta_0 + \beta_1 x_2) \\ \vdots \\ y_n - (\beta_0 + \beta_1 x_n) \end{bmatrix} \end{aligned}$$

   Minimizing $RSS = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$ w.r.t. $\boldsymbol{\beta}$, we have $\widehat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$, where A^{-1} is the inverse of the square matrix A.

   (a) Consider data set `Colleges.txt`. Write a function in R **using the matrix approach** to perform a simple linear regression of percentage of applicants accepted (Acceptance) on the median combined math and verbal SAT score of students (SAT).

   Compare the results with the answers in part (b) of Question 1.

   (b) If data set of n points has two input features, x^1, x^2, by matrix approach, the estimate of coefficient is still $\widehat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$.

      i. Specify matrix $\mathbf{y}$, $\mathbf{X}$ and $\boldsymbol{\beta}$.

      ii. Use your function in part (a) to perform a multivariate linear regression of percentage of applicants accepted (`Acceptance`) on `SAT` and `Top.10p` - percentage of students in the top 10% of their high school graduating class.

2. A dataset on house selling price was randomly collected [1], `house_selling_prices_FL.csv`. It's our interest to model how y = selling price (dollar) is dependent on x = the size of the house (square feet). A simple linear regression model (y regress on x) was fitted, called Model 1.

   The given data has another variable, NW, which specifies if a house is in the part of the town considered less desirable (NW = 0).

[1] *Statistics: The Art and Science of Learning from Data*, 4th, Agresti, Franklin, Klingenberg

(a) Derive the correlation between x and y.

(b) Derive a scatter plot of y against x. Give your comments on the association of y and x.

(c) Derive R^2 of Model 1. Verify that $\sqrt{R^2} = |cor(y, x)|$. In which situation we can have $\sqrt{R^2} = cor(y, x)$?

(d) Form a model (called Model 2) which has two regressors (x and NW). Write down the equation of Model 2.

(e) Report the coefficient of variable NW in Model 2. Interpret it.

(f) Estimate the price of a house where its size is 4000 square feet and is located at the more desirable part of the town.

(g) Report the R^2 of Model 2. Interpret it.