Homework 1

Stats 413: Applied Regression Analysis

due date: 11:59pm (Eastern time), September 18, 2020

For problems that require programming, please properly comment your code and submit it together with needed output.

1. **United Nations Data** (copied from Problem 1.1 in Applied Linear Regression)
   The data in the file *UN11* from http://users.stat.umn.edu/~sandy/alr4ed/data/ contains several variables, including ppgdp, the gross national product per person in U.S. dollars, and fertility, the birth rate per 1000 females, both from the year 2009. The data are for 199 localities, mostly UN member countries, but also other areas such as Hong Kong that are not independent countries. The data were collected from United Nations (2011). We will study the dependence of fertility on ppdgp.

   (a) Identify the predictor and the response.

   (b) Draw the scatterplot of fertility on the vertical axis versus ppgdp on the horizontal axis and summarize the information in this graph. Does a straight-line mean function seem to be plausible for a summary of this graph?

   (c) Draw the scatterplot of log(fertility) versus log(ppgdp) using natural logarithms. Does the simple linear regression model seem plausible for a summary of this graph? If you use a different base of logarithms, use scatterplot to show that the shape of the graph won't change, but the values on the axes will change.

2. **Professor ratings Data** (copied from Problem 1.6 in Applied Linear Regression)
   The data *Rateprof* can be found from http://users.stat.umn.edu/~sandy/alr4ed/data/. In the website and online forum RateMyProfessors.com, students rate and comment on their instructors. Launched in 1999, the site includes millions of ratings on thousands of instructors. The data file includes the summaries of the ratings of 364 instructors at a large campus in the Midwest (Bleske-Rechek and Fritsch, 2011). Each instructor included in the data had at least 10 ratings over a several year period. Students provided ratings of 1-5 on quality, helpfulness, clarity, easiness of instructor's courses, and rater-Interest in the subject matter covered in the instructor's courses. The data file provides the averages of these five ratings.

   Draw the scatterplot matrix for these five ratings and summarize the information available from these plots.

3. **Height and weight data** (Prob. 2.1 in Applied Linear Regression)

   The data *Htwt* from http://users.stat.umn.edu/~sandy/alr4ed/data/ give "ht" = height in centimeters and "wt" = weight in kilograms for a sample of $n = 10$ 18-year-old girls. Interest is in predicting weight from height.

(a) Draw a scatter plot of "wt" on the vertical axis versus "ht" on the horizontal axis. On the basis of this plot, does a simple linear regression model make sense for these data? Why or why not?

(b) Compute estimates of the slope and the intercept for the regression of "wt" on "ht". Draw the fitted line on your scatterplot.

(c) Obtain the estimate of σ^2 and find the estimated standard errors of $\hat{\beta}_0$ and $\hat{\beta}_1$.

4. **Simple linear regression**

Prove that in the case of simple linear regression with a sample $(x_i, y_i), i = 1, \ldots, n$, the lest squares line always passes through the point $(\bar{x}, \bar{y})$, where $\bar{x} = n^{-1} \sum_{i=1}^{n} x_i$ and $\bar{y} = n^{-1} \sum_{i=1}^{n} y_i$.

5. **Multi-task regression** (by Andrew Ng)

Thus far, we only considered regression with scalar-valued responses. In some applications, the response is itself a vector: $\mathbf{y}_i \in \mathbb{R}^{m \times 1}$. We posit the relationship between the features/predictors ($\mathbf{x}_i \in \mathbb{R}^{d \times 1}$) and the vector-valued response $\mathbf{y}_i$ is linear:

$$\mathbf{y}_i^T = \mathbf{x}_i^T B^* + error, \text{ for } i = 1, \cdots, n$$

where $B^* \in \mathbb{R}^{d \times m}$ is a matrix of regression coefficients. Here note that for the linear regression model in class, the dimension of response variable $\mathbf{y}_i$ is $m = 1$.

(a) Express the sum of squared residuals (also called residual sum of squares, RSS) in matrix notation (*i.e.* without using any summations). Similarly to the linear regression model, the RSS is defined as

$$RSS(B) = \sum_{i=1}^{n} (\mathbf{y}_i^T - \mathbf{x}_i^T B)(\mathbf{y}_i^T - \mathbf{x}_i^T B)^T.$$

Hint: *work out how to express the RSS in terms of the data matrices*

$$\mathbf{X} = \begin{bmatrix} - & \mathbf{x}_1^T & - \\ & \vdots & \\ - & \mathbf{x}_n^T & - \end{bmatrix} \in \mathbb{R}^{n \times d}, \quad \mathbf{Y} = \begin{bmatrix} - & \mathbf{y}_1^T & - \\ & \vdots & \\ - & \mathbf{y}_n^T & - \end{bmatrix} \in \mathbb{R}^{n \times m}.$$

Also note that for a matrix $A = (a_{ij})_{n \times m}$ with its ith row vector denoted by $\mathbf{a}_i$, we have $tr(AA^\top) = \sum_{i=1}^{n} \mathbf{a}_i \mathbf{a}_i^\top = \sum_{1 \le i,j \le n} a_{ij}^2$.

(b) Find the matrix of regression coefficients that minimizes the RSS.

(c) Instead of minimizing the RSS, we break up the problem into m regression problems with scalar-valued responses. That is, we fit m linear models of the form

$$(\mathbf{y}_i)_k = \mathbf{x}_i^T \beta_k + error,$$

where $(\mathbf{y}_i)_k$ denotes the kth element in the vector $\mathbf{y}_i$ and $\beta_k \in \mathbb{R}^d$. How do the regression coefficients from the m separate regressions compare to the matrix of regression coefficients that minimizes the SSR in (b).