Assignment 0

All the datasets mentioned in the assignment are available in the courseworks website except the last one.

1. Given 100 number as follows :

364	142	865	945	453	556	602	78	784	562
197	589	34	27	338	19	431	678	73	378
524	810	84	646	666	457	100	833	929	91
730	790	916	770	996	357	435	310	698	816
116	651	532	970	552	297	268	332	175	271
751	124	696	275	564	112	169	998	64	864
592	63	412	270	535	114	450	792	39	910
413	565	537	209	370	233	96	557	471	467
261	23	762	775	741	199	786	127	276	662
60	362	240	327	874	746	81	859	133	629

Try sorting them by hand and then write a description of your process of sorting. Suppose there is another human being who would do exactly what you tell him to do, write directions for him to sort the numbers. Finally, write an R script that can read the numbers (the file "sort_data.txt") and sort them. How does your algorithm compare in computational complexity with selection sort and merge-sort algorithms discussed in class?

2. There are two matrices X, Y and a vector β as follows:

$$X = \begin{bmatrix} -1 & -2 & 1 \\ 4 & 0 & -6 \\ -8 & -7 & 9 \end{bmatrix}, Y = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix}, \beta = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$$

Find $XY, YX, X\beta, X^T, X^{-1}$, the rank of X. Do these calculations by hand. And then write R code to do them in R.

3. Download the dataset "prostate_data.csv" from courseworks. The data comes from a study of relation between the level of prostate specific antigen (PSA) and a number of clinical measures. For example, lcavol, lweight and lbph represent cancer volume, prostate weight and benign prostatic hyperplasia amount respectively. Excluding the categorical variables “svi” and “gleason”, do simple linear regressions for each pairs within the remaining seven variables. Plot the data and regression lines for each pair. Try to put all the plots together into one picture.

4. Recall the least square estimator problem in the quiz. Now let’s derive the formula for it. Consider the simple linear regression model:

$$y_i = a_0 x_i + b_0 + \epsilon_i \quad (1 \leq i \leq n)$$

The model says that there is almost a linear relation between y_i and x_i (perturbed by noise ϵ_i). We have the data $(x_i, y_i)_{i=1}^n$, then how do we find the best straight line to capture the true relation (or how to estimate a_0 and b_0) between x_i and y_i. Intuitively, if there is no noise term ϵ_i, we can simply find a and b such that $\sum_{i=1}^n (y_i - a x_i - b)^2 = 0$. In the presence of noise, we can estimate a and b by minimizing $\sum_{i=1}^n (y_i - a x_i - b)^2$ as a denoising process. The least square estimator is exactly

$$(\hat{a}, \hat{b}) = \mathrm{argmin}_{a,b \in \mathbb{R}} \sum_{i=1}^{n} (y_i - a x_i - b)^2$$

Prove

$$\hat{a} = \frac{\sum_{i=1}^n (x_i - \bar{x}) y_i}{\sum_{i=1}^n (x_i - \bar{x})^2}, \quad \hat{b} = \bar{y} - \hat{a}\bar{x}$$

where $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i, \quad \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$

5. Suppose that a tuberculosis (TB) skin test is 95 percent accurate. That is, if the patient is TB-infected, then the test will be positive with probability 0.95, and if the patient is not infected, then the test will be negative with probability 0.95. From research study, we know 1 in 1000 of the subjects in the population is infected. Now suppose that a person comes to do the skin test, what is the probability that he would get a positive test result? If we know he got a positive test result, then what is the probability that he is infected?

6. Go to the website `http://www.kaggle.com`. Find the competition named “Titanic: Machine Learning from Disaster”. Download the file “train.csv”. Try to summarize the dataset with plots and summary statistics. Show what you find by the descriptive procedure.