Homework 4

posted November 29, 2017
due December 13, 2017

1. Four points A, B, C, D are located are colinear and the distance between them are measured approximatively yielding the following results:

$$AD = 89, AC = 67, BD = 53, AB = 35, \text{ and } CD = 20.$$

   We need to determine the length of the segments $r_1 = AB$, $r_2 = BC$, and $r_3 = CD$.

   The results are inconsistent because if we use the last three equations

$$\begin{aligned} r_1 + r_2 + r_3 &= 89 \\ r_1 + r_2 &= 67 \\ r_2 + r_3 &= 53 \\ r_1 = 35 \\ r_3 = 20 \end{aligned}$$

   we have $r_1 = 35$, $r_2 = 33$ and $r_3 = 20$. However, the first two equations yield $x_1 + x_2 + x_3 - 89 = -1$ and $x_1 + x_2 - 67 = 1$.

   Write the above system in matrix form $A\mathbf{r} = \mathbf{b}$, where $\mathbf{r} = \begin{pmatrix} r_1 \\ r_2 \\ r_3 \end{pmatrix}$, $A \in \mathbb{R}^{5\times 3}$ and $\mathbf{b} \in \mathbb{R}^5$ and determine $\mathbf{r}$ such that $\parallel A\mathbf{r} - \mathbf{b} \parallel$ is minimal.

2. Let $\mathbf{b} \in \mathbb{R}^m - \{\mathbf{0}_m\}$ and $\mathbf{y} \in \mathbb{R}^m$. Prove that $\parallel \mathbf{b}r - \mathbf{y} \parallel$ is minimal when $r = \frac{(\mathbf{y},\mathbf{b})}{\|\mathbf{b}\|^2}$.

Let $B = (\mathbf{b}^1 \; \cdots \; \mathbf{b}^n) \in \mathbb{R}^{m\times n}$ be a matrix that contains input data of m experiments. The rows of this matrix are denoted by $\mathbf{u}_1, \ldots, \mathbf{u}_m$, where $\mathbf{u}_i \in \mathbb{R}^n$ contains the input values of the variable for the i^{th} experiment. The *average* of B is the vector $\tilde{\mathbf{u}} = \frac{1}{m} sum_{i=1}^m \mathbf{u}_i$. The matrix is *centered* if $\tilde{\mathbf{u}} = \mathbf{0}'_n$. Note that $\tilde{u} = \frac{1}{m}\mathbf{1}'_m B$. Note that the matrix

$$\hat{B} = \left(I_m - \frac{1}{m}\mathbf{1}_m\mathbf{1}'_m \right) B$$

because

$$\begin{aligned}
\frac{1}{m}\mathbf{1}'_m \hat{B} &= \frac{1}{m}\mathbf{1}'_m \left(I_m - \frac{1}{m}\mathbf{1}_m\mathbf{1}'_m\right) B \\
&= \frac{1}{m}\left(\mathbf{1}'_m - \frac{1}{m}\mathbf{1}'_m\mathbf{1}_m\mathbf{1}'_m\right) B \\
&= \frac{1}{m}\left(\mathbf{1}'_m - \mathbf{1}'_m\right) B = 0.
\end{aligned}$$

The matrix $H_m = I_m - \frac{1}{m}\mathbf{1}_m\mathbf{1}'_m \in \mathbb{R}^{m\times m}$ is the *centering matrix*.

If the measurement scales of the variables $x_1, \ldots, x_n$ involved in the series are very different due to different measurement units, some variables may influence inappropriately the certain regression processes. The *standard deviation* of a vector $\mathbf{b} \in \mathbb{R}^m$ is $s(\mathbf{b}) = \sqrt{\frac{1}{m-1}\sum_{b_i - \tilde{b}}}$, where $\tilde{b} = \frac{1}{m}\sum_{i=1}^m b_i$. To scale a matrix we need to replace each column $\mathbf{b}^j$ by $\frac{1}{s(\mathbf{b}^j)}\mathbf{b}^j$.

3. Prove that the centering matrix H_n is symmetric and idempotent.

4. Let $B \in \mathbb{R}^{m\times n}$ and $\mathbf{y} \in \mathbb{R}^m$ the data used in linear regression. Suppose that B is centered and define the matrix $\hat{B} = \begin{pmatrix} B \\ \sqrt{\lambda} I_n \end{pmatrix} \in \mathbb{R}^{(m+n)\times n}$ and $\hat{\mathbf{y}} = \begin{pmatrix} \mathbf{y} \\ \mathbf{0}_n \end{pmatrix} \in \mathbb{R}^{m+n}$. Prove that the ordinary regression applied to this data amounts to ridge regression.

5. Study the GLmnet Vignette (a description of the `glmnet` R package) which is posted on the web site. Install this package and also, the package `ggplot2`. The dataset `diamonds` is a part of `ggplot2`. This data gives the price of a diamond as a function of the carat weight, cut, color, etc.

   Apply at least two type of regression to this dataset using the glmnet package. Of course you will have to install and upload both `glmnet` and `ggplot2`.