Assigned: Wednesday, March 25, 2015
Due: Week of April 20, 2015

MTH 371: Group Project 2
Least Squares Approximation and Sports Ranking

GENERAL GROUP PROJECT GUIDELINES:

- Group project assignments should be a collaborative effort. All should participate in discussion and solution writing.
- Two weeks after the project is assigned, your group will meet with Dr. Vidden to discuss. All members must be present. Your grade will be determined at the end of the meeting.
- Each student should keep group project solutions in a dedicated notebook. Bring this notebook to your weekly meeting to discuss your findings. For coded solutions, bring a laptop to your weekly meeting. Have the laptop ready before the start of the meeting.

With this project, we consider a least squares approach for ranking sports teams. This method known as the Massey method and was developed as an undergraduate thesis by Kenneth Massey in 1997. Since then, this method has progressed and is now used by the Bowl Championship Series system to select NCAA football bowl matchups.

1. In this class, we have solved nonsingular systems $A\vec{x} = \vec{b}$ with n equations and n unknowns. If on the other hand there are *more equations than unknowns*, the system probably has no solution. Instead, we might wish to find an approximate solution $A\vec{x} \approx \vec{b}$. The *least squares problem* is to minimize the residual in the ℓ^2 norm, $\|\vec{b} - A\vec{x}\|_2$. Recall, $\|\cdot\|_2$ denotes the Euclidean norm. For this problem, we solve the least squares problem using calculus. Assume dimensions $A_{m\times n}, \vec{x}_n$, and $\vec{b}_m$.
   (a) Why is this problem called the *least squares problem*?
   (b) To minimize $\|\vec{b} - A\vec{x}\|_2$, differentiate in each component x_k and set equal to zero. Why must there only be one minimum here?
   (c) From (b), show that for each $k = 1, \ldots, n$,
   $$\sum_{i=1}^{m} a_{ik} \left(\sum_{j=1}^{n} a_{ij} x_j \right) = \sum_{i=1}^{m} a_{ik} b_i.$$
   (d) From (c), show that $A^T A\vec{x} = A^T\vec{b}$. This systems is called the *normal equations*.

2. Consider the following set of basketball matchups for teams $T_1, \ldots, T_5$.

	Duke(T_1)	Miami(T_2)	UNC(T_3)	UVA(T_4)	VT(T_5)	Record	Point Differential
Duke(T_1)		7-52	21-24	7-38	0-45	0-4	-124
Miami(T_2)	52-7		34-16	25-17	27-7	4-0	91
UNC(T_3)	24-21	16-34		7-5	3-30	2-2	-40
UVA(T_4)	38-7	17-25	5-7		14-52	1-3	-17
VT(T_5)	45-0	7-27	30-3	52-14		3-1	90

(a) For each game, we have the equation $r_i - r_j = y_k$ where y_k is the margin of victory for game k, and r_i and r_j are the ratings of teams i and j respectively. In other words, the difference in ratings of the two teams ideally predicts the margin of victory of a match between the teams. Construct the system $A\vec{r} = \vec{y}$ for the above matchups and show it has no solution.

(b) Construct the normal equations $M\vec{r} = \vec{p}$ which solve the least squares problem. What is M? $\vec{p}$?

(c) Matrix M of part (b) is called the Massey matrix. M turns out to be rather simple. In fact, it needs not be computed. It can be formed by assigning diagonal entries M_{ii} as the total number of games played by team i. For off diagonal entries $M_{ij}, i \neq j$, these are the negation of the matchups between teams i and j. Explain how this happens.

(d) Also, vector $\vec{p}$ is rather nice as well. What happens here?

(e) Unfortunately, because of the structure of matrix M, system $M\vec{r} = \vec{p}$ has no unique solution. Show M is singular (hint: what is the row sum of M?). Show if $\vec{r}$ solves the system, so does $\vec{r} + \vec{c}$ for any constant vector $\vec{c}$, so there are in fact infinite solutions.

(f) Massey's workaround to this problem is to add another constraint to the system stating that all ratings r_i must sum to 0. Replace the last equation in your least squares system with this requirement. Solve this final system. Make sure you construct M in Scilab in a *general way* using the nice structure of M as seen above. Your code should be easily adaptable to adding more games. Your final results should match the following table.

Team	r_i	rank
Duke	-24.8	5th
Miami	18.2	1st
UNC	-8.0	4th
UVA	-3.4	3rd
VT	18.0	2nd

3. BONUS extra credit problems.

(a) Visit the website http://www.masseyratings.com/ and download date from your favorite sport. Rank the teams using Massey's method.

(b) Read the Colley Method paper on D2L and implement this method for the above basketball matchups.