• Instructor: He Wang Email: he.wang@northeastern.edu

Further Reading on $1:

We can use Matlab to do some computation to verify our calculations. We list some basic MATLAB functions about linear system here.

Input matrices and vectors to MATLAB:

```
A=[1 2 7;-2 5 4;5 0 15];
b=[1; 0; 3];
c = [2 1 8];
```

mldivide Solve systems of linear equations $A\vec{x} = B$ for $\vec{x}$. Use

```
x = A\B  or x = mldivide(A,B)
```

rref Reduced row echelon form (Gauss-Jordan elimination)

```
rref(A)
```

mrdivide Solve systems of linear equations $\vec{x}A = C$ for $\vec{x}$.

```
x = C/A  or x = mrdivide(C,A)
```

size(A): number of elements of A and size of A

```
size(A)
```

zeros(m,n) - Produces an $m \times n$ matrix of 0's.

ones(m,n) - Produces an $m \times n$ matrix of 1's.

https://www.mathworks.com/help/matlab/linear-algebra.html

Further reading: More real world applications.

Most of program languages deal with vectors. (Advantages: Programming using vectors is shorter and much faster than using entries.)

Linear Regression in Statistics We want to model the linear relationship between predictor variables $\vec{x}$ and a target variable y.

Example 1. (Study time and final scores) We want to model the relation between hours/week (x) spent studying and final scores y by students. Our goal is to find a function

$$y = c_0 + c_1 x$$

with parameters c_0 and c_1.

Suppose we have data of 6 students (4, 80), (5, 85), (5.5, 85), (6, 90), (6.5, 95), (7, 92). Hence the linear system is

$$\begin{cases} c_0 + 4c_1 & = 80 \\ c_0 + 5c_1 & = 85 \\ c_0 + 5.5c_1 & = 85 \\ c_0 + 6c_1 & = 90 \\ c_0 + 6.5c_1 & = 95 \\ c_0 + 7c_1 & = 92 \end{cases}$$

So, we need to solve $A\vec{c} = \vec{b}$ where $A = \begin{bmatrix} 1 & 4 \\ 1 & 5 \\ 1 & 5.5 \\ 1 & 6 \\ 1 & 6.5 \\ 1 & 7 \end{bmatrix}$ and $\vec{b} = \begin{bmatrix} 80 \\ 85 \\ 85 \\ 90 \\ 95 \\ 92 \end{bmatrix}$

We already know how to solve a consistent linear system. However, this linear system is inconsistent. In §5.4, we will use least-squares method to approach solutions for this. The least-squares solutions is given by $\vec{c} = \begin{bmatrix} 60.9571 \\ 4.7429 \end{bmatrix}$

Now, you want to see how many hours to spend in the class is better. So, we test when $x = 4.5$ $x = 6$ and $x = 7.5$.

Let $B = \begin{bmatrix} 1 & 4.5 \\ 1 & 6 \\ 1 & 7.5 \\ 1 & 8 \end{bmatrix}$ and calculate $B\vec{c} = \begin{bmatrix} 82.3 \\ 89.4 \\ 96.5 \\ 98.9 \end{bmatrix}$

Of course, we know that this is not the precise value since the final grade are affect by many other factors, however, this give us the first approach for the prediction.

Example 2. (House price)

We want to find a formula to predict the final price y (in \$) of each house in a town. The prices are affected by LotArea (x_1 in ft^2), HouseLivingArea (x_2 in ft^2), GarageArea (x_3 in ft^2), YearBuild (x_4).

Our goal is to find a function

$$y = c_0 + c_1 x_1 + c_2 x_2 + c_3 x_3 + c_4 x_4$$

with parameters c_0 c_1, c_2, c_3 and c_4. Suppose we have the data from 10 houses in Ames, Iowa.

x_1	x_2	x_3	x_4	y
8450	1710	548	2003	208500
9600	1262	460	1976	181500
11250	1786	608	2001	223500
9550	1717	642	1915	140000
14260	2198	836	2000	250000
14115	1362	480	1993	143000
10084	1694	636	2004	307000
10382	2090	484	1973	200000
6120	1774	468	1931	129900
7420	1077	205	1939	118000

So, we need to solve $A\vec{c} = \vec{b}$ where

$$A = \begin{bmatrix} 1 & 8450 & 1710 & 548 & 2003 \\ 1 & 9600 & 1262 & 460 & 1976 \\ 1 & 11250 & 1786 & 608 & 2001 \\ 1 & 9550 & 1717 & 642 & 1915 \\ 1 & 14260 & 2198 & 836 & 2000 \\ 1 & 14115 & 1362 & 480 & 1993 \\ 1 & 10084 & 1694 & 636 & 2004 \\ 1 & 10382 & 2090 & 484 & 1973 \\ 1 & 6120 & 1774 & 468 & 1931 \\ 1 & 7420 & 1077 & 205 & 1939 \end{bmatrix} \text{ and } \vec{b} = \begin{bmatrix} 208500 \\ 181500 \\ 223500 \\ 140000 \\ 250000 \\ 143000 \\ 307000 \\ 200000 \\ 129900 \\ 118000 \end{bmatrix}$$

The least-squares solution is given by $\vec{c} = \begin{bmatrix} -2512046.85480899 \\ -8.87544782059107 \\ 6.51030372812289 \\ 195.509869623197 \\ 1356.09378125158 \end{bmatrix}$

Now, let us use another 10 houses to see whether or not our function is good.

x_1 x_2 x_3 x_4 y

```
11200  1040   384  1965   129500
11924  2324   736  2005   345000
12968  912  352  1962   144000
10652  1494   840  2006   279500
10920  1253   352  1960   157000
6120   854  576  1929   132000
11241  1004   480  1970   149000
10791  1296   516  1967   90000
13695  1114   576  2004   159000
7560   1339   294  1958   139000
```

We will use matrix multiplication $B\vec{c}$ to predict the house prices. (This is much faster than evaluate one by one in programming.)

$$B\vec{c} = \begin{bmatrix} 1 & 11200 & 1040 & 384 & 1965 \\ 1 & 11924 & 2324 & 736 & 2005 \\ 1 & 12968 & 912 & 352 & 1962 \\ 1 & 10652 & 1494 & 840 & 2006 \\ 1 & 10920 & 1253 & 352 & 1960 \\ 1 & 6120 & 854 & 576 & 1929 \\ 1 & 11241 & 1004 & 480 & 1970 \\ 1 & 10791 & 1296 & 516 & 1967 \\ 1 & 13695 & 1114 & 576 & 2004 \\ 1 & 7560 & 1339 & 294 & 1958 \end{bmatrix} \begin{bmatrix} -2512046.85480899 \\ -8.87544782059107 \\ 6.51030372812289 \\ 195.509869623197 \\ 1356.09378125158 \end{bmatrix} \approx \begin{bmatrix} 135118 \\ 260115 \\ 108269 \\ 287690 \\ 125953 \\ 167713 \\ 160070 \\ 168935 \\ 203881 \\ 142283 \end{bmatrix}$$

The real selling price is $bb = \begin{bmatrix} 129500 \\ 345000 \\ 144000 \\ 279500 \\ 157000 \\ 132000 \\ 149000 \\ 90000 \\ 159000 \\ 139000 \end{bmatrix}$. The difference is $B\vec{c} - bb = \begin{bmatrix} 5618 \\ -84884 \\ -35730 \\ 8190 \\ -31046 \\ 35713 \\ 11070 \\ 78935 \\ 44881 \\ 3283 \end{bmatrix}$.

We can see that the prediction is "ok" but there is room to make it better. We know that the price of house can be affect by many other factors we did not consider here, like, the number of bathrooms, the school district, how safe is the district, near public transportation or not, etc. In addition, we only used 10 houses to get the formula, if we have large data, the prediction will also be better.

There is a Kaggle completion on this topic. Further Reading:

https://www.kaggle.com/c/house-prices-advanced-regression-techniques http://jse.amstat.org/v19n3/decock.pdf

Matlab code for Example 1.

```

%% Example 1: Study time and grade example
clear all

A=[1 4;
1  5 ;
1 5.5 ;
1 6 ;
1 6.5 ;
1 7 ]

b=[80;
85;
85;
90;
95;
92]

%% Least Squares solution
c=A\b

%% Least Squares solution(alternative method)
c=(transpose(A)*A)^(-1)*(transpose(A)*b)


%% Predict

B=[1 4.5;
1 6 ;
1 7.5 ;
1 8 ]

B*c
```

Matlab code for Example 2.

```


%% Example 2:  House Price
clear all

A=[1  8450  1710  548 2003;
1 9600  1262  460 1976;
1 11250 1786  608 2001;
1 9550  1717  642 1915;
1 14260 2198  836 2000;
1 14115 1362  480 1993;
1 10084 1694  636 2004;
1 10382 2090  484 1973;
1 6120  1774  468 1931;
```

```
1 7420  1077  205 1939]

b=[
208500;
181500;
223500;
140000;
250000;
143000;
307000;
200000;
129900;
118000
]
%% Least Squares solution
c=A\b

%% Least Squares solution(alternative method)
c=(transpose(A)*A)^(-1)*(transpose(A)*b)

%% Test: Predict house
B=[1  11200 1040  384 1965;
1 11924 2324  736 2005;
1 12968 912 352 1962;
1 10652 1494  840 2006;
1 10920 1253  352 1960;
1 6120  854 576 1929;
1 11241 1004  480 1970;
1 10791 1296  516 1967;
1 13695 1114  576 2004;
1 7560  1339  294 1958]

v=B*c

bb=[129500;
345000;
144000;
279500;
157000;
132000;
149000;
90000;
159000;
139000]
%% difference
di=v-bb

%%
clear all


```

Example 3. (Salary) Model the relation between salary (y) and the top degree, number of years working experiences, number of certificates, age, etc.