Math 10A
Worksheet, Discussion #34; Thursday, 8/2/2018
Instructor name: Roy Zhao

1 Linear Regression

1.1 Concepts

1. Often when given data points, we want to find the line of best fit through them. To them, we want to approximate them with a line $y = ax + b$. We represent this as a solution where we want to solve for a, b. In matrix vector form and data points (x_i, y_i), this is represented as

$$A\vec{x} = \vec{b} \to \begin{pmatrix} x_1 & 1 \\ x_2 & 1 \\ \vdots & \vdots \\ x_n & 1 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}.$$

Often, we cannot find a perfect fit (if not all the points lie on the same line). So we want to find the error. One way to find the error is to take the least square error or $E = \sum(y_i - (ax_i + b))^2$, the sum of the squares of the error. The choice of a, b that minimizes this is

$$\begin{pmatrix} a \\ b \end{pmatrix} = (A^T A)^{-1} A^T \vec{b}.$$

Written out, we have

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

where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ is the average of the x values and $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ is the average of the y values.

1.2 Problems

2. **TRUE** False The matrix $A^T A$ will always be square.

> **Solution:** If A is $m \times n$, then A^T is $n \times m$ so $A^T A$ is $(n \times m)(m \times n) = n \times n$ so it will be square.

3. Consider the set of points $\{(-2, -1), (1, 1), (3, 2)\}$. Calculate the line of best fit.

Solution: We first write it in matrix form as

$$\begin{pmatrix} -2 & 1 \\ 1 & 1 \\ 3 & 1 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} -1 \\ 1 \\ 2 \end{pmatrix}.$$

Then we calculate $A^TA = \begin{pmatrix} 14 & 2 \\ 2 & 3 \end{pmatrix}$ and $A^T\vec{b} = \begin{pmatrix} 9 \\ 2 \end{pmatrix}$. Then

$$\begin{pmatrix} a \\ b \end{pmatrix} = (A^TA)^{-1}A^T\vec{b} = \frac{1}{42-4}\begin{pmatrix} 3 & -2 \\ -2 & 14 \end{pmatrix}\begin{pmatrix} 9 \\ 2 \end{pmatrix} = \begin{pmatrix} 23/38 \\ 10/38 \end{pmatrix}.$$

So the line of best fit is $y = 23/38x + 5/19$.

4. Find the line of best fit and the error of the fit of the points $\{(-1,2),(0,-1),(1,1),(3,2)\}$ and use it to estimate the value at 2.

Solution: We first write it in matrix form as

$$\begin{pmatrix} -1 & 1 \\ 0 & 1 \\ 1 & 1 \\ 3 & 1 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} 2 \\ -1 \\ 1 \\ 2 \end{pmatrix}.$$

Then we calculate $A^TA = \begin{pmatrix} 11 & 3 \\ 3 & 4 \end{pmatrix}$ and $A^T\vec{b} = \begin{pmatrix} 5 \\ 4 \end{pmatrix}$. Then

$$\begin{pmatrix} a \\ b \end{pmatrix} = (A^TA)^{-1}A^T\vec{b} = \frac{1}{44-9}\begin{pmatrix} 4 & -3 \\ -3 & 11 \end{pmatrix}\begin{pmatrix} 5 \\ 4 \end{pmatrix} = \begin{pmatrix} 8/35 \\ 29/35 \end{pmatrix}.$$

So the line of best fit is $y = 8/35x + 29/35$. Thus, $y(2) = 8/35(2) + 29/35 = (16+29)/35 = 45/35 = 9/7$.

5. Consider the set of points $\{(-2,-1),(1,1),(3,2)\}$. Calculate the square error if we estimate it using the line $y = x$. Then calculate the square error if we use the line $y = 0$. Which is a better approximation?

Solution: Using the fit $y = x$ gives the error $(-1-(-2))^2 + (1-1)^2 + (2-3)^2 = 1+0+1 = 2$. Using the approximation $y = 0$ gives the error $(-1-0)^2 + (1-0)^2 + (2-0)^2 = 1+1+4 = 6$. Thus $y = x$ is the better fit.

6. The number of people applying to Berkeley is given in the following table:

Year	2011	2012	2013	2014	2015	2016	2017
Applicants(in 1000s)	53	62	68	74	79	83	85

Predict how many people applied this year (2018).

Solution: We first write it in matrix form as

$$\begin{pmatrix} 2011 & 1 \\ 2012 & 1 \\ 2013 & 1 \\ 2014 & 1 \\ 2015 & 1 \\ 2016 & 1 \\ 2017 & 1 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} 53 \\ 62 \\ 68 \\ 74 \\ 79 \\ 83 \\ 85 \end{pmatrix}.$$

Then we calculate $A^TA = \begin{pmatrix} 28393400 & 14098 \\ 14098 & 7 \end{pmatrix}$ and $A^T\vec{b} = \begin{pmatrix} 1015205 \\ 504 \end{pmatrix}$. Then

$$\begin{pmatrix} a \\ b \end{pmatrix} = (A^TA)^{-1}A^T\vec{b} = \frac{1}{28393400 \cdot 7 - 14098^2} \begin{pmatrix} 7 & -14098 \\ -14098 & 28393400 \end{pmatrix} \begin{pmatrix} 1015205 \\ 504 \end{pmatrix} \approx \begin{pmatrix} 5 \\ -10645 \end{pmatrix}.$$

So the line of best fit is $y = 5x - 10645$. So the estimate for the number of people who will apply in 2018 is $5 \cdot 2018 - 10645 = 93.3$. (The true number was around 89 million)