MATH 10A - Fall 2018 GSI Section 26
Professor Stankova 11/29/18

Least Squares Approximation

Line of Best Fit with 2-Points: (20 minutes) 0:10 – 0:30

Given the two points, $(1,2)$ and $(3,-4)$, calculate an equation for the line of best fit, $y = ax + b$

1) Plot the two points in a two-dimensional graph. I believe in your ability to do this.
2) Use the points to define the vectors: $\vec{x}, \vec{y}, \vec{w}, \& \vec{u}$.

$$\vec{x} = \begin{pmatrix}1\\3\end{pmatrix}, \vec{y} = \begin{pmatrix}2\\-4\end{pmatrix}, \vec{w} = \begin{pmatrix}f(1)\\f(3)\end{pmatrix}, \& \vec{u} = \begin{pmatrix}1\\1\end{pmatrix}$$

3) Using these vectors, define a system of equations that will lead to the least squares formula.

$$\begin{Bmatrix}\vec{x} \cdot (\vec{w} - \vec{y}) = 0\\ \vec{u} \cdot (\vec{w} - \vec{y}) = 0\end{Bmatrix} \Rightarrow \begin{Bmatrix}\vec{x} \cdot \vec{w} = \vec{x} \cdot \vec{y}\\ \vec{u} \cdot \vec{w} = \vec{u} \cdot \vec{y}\end{Bmatrix} \Rightarrow \begin{pmatrix}x_1 & x_2\\ 1 & 1\end{pmatrix}\begin{pmatrix}w_1\\w_2\end{pmatrix} = \begin{pmatrix}x_1 & x_2\\ 1 & 1\end{pmatrix}\begin{pmatrix}y_1\\y_2\end{pmatrix}$$

$$\begin{pmatrix}1 & 3\\ 1 & 1\end{pmatrix}\begin{pmatrix}w_1\\w_2\\w_3\end{pmatrix} = \begin{pmatrix}1 & 3\\ 1 & 1\end{pmatrix}\begin{pmatrix}2\\-4\end{pmatrix}$$

4) Rewrite this system of equations as an equality containing the transposed matrix, A^T, and the vectors $\vec{w}$ & $\vec{y}$.

$$A^T\vec{w} = A^T\vec{y}$$

5) Using the definition of $\vec{w} = a\vec{x} + b\vec{u}$, rewrite the equality from (4) as the Normal Equation, containing a vector of the coefficients to be estimated, a & b.

$$\vec{w} = a\vec{x} + b\vec{u} \Rightarrow \vec{w} = \begin{pmatrix}x_1 & 1\\ x_2 & 1\end{pmatrix}\begin{pmatrix}a\\b\end{pmatrix} = A\begin{pmatrix}a\\b\end{pmatrix}$$

$$\text{Normal Equation: } A^TA\begin{pmatrix}a\\b\end{pmatrix} = A^T\vec{y} \Rightarrow \begin{pmatrix}1 & 3\\ 1 & 1\end{pmatrix}\begin{pmatrix}1 & 1\\ 3 & 1\end{pmatrix}\begin{pmatrix}a\\b\end{pmatrix} = \begin{pmatrix}1 & 3\\ 1 & 1\end{pmatrix}\begin{pmatrix}2\\-4\end{pmatrix}$$

6) Use matrix multiplication to simplify the normal equation.

$$\begin{pmatrix}10 & 4\\ 4 & 2\end{pmatrix}\begin{pmatrix}a\\b\end{pmatrix} = \begin{pmatrix}-10\\-2\end{pmatrix}$$

7) Take the inverse of a certain $2x2$ matrix and multiply both sides of the simplified Normal Equation to solve for a & b. Plug these values into the line of best fit and plot it on your graph.

$$B = \frac{1}{4}\begin{pmatrix}2 & -4\\ -4 & 10\end{pmatrix} \qquad \begin{pmatrix}a\\b\end{pmatrix} = \frac{1}{4}\begin{pmatrix}2 & -4\\ -4 & 10\end{pmatrix}\begin{pmatrix}-10\\-2\end{pmatrix} = \begin{pmatrix}-3\\5\end{pmatrix} \qquad y = -3x + 5$$

8) What do you notice about the line relative to your original points? Will this happen for every least squares approximation with only two points?

Line of best fit passes through points – yes, given only two points the process will find the line passing through them.

Line of Best Fit with 3-Points: (20 minutes) 0:30 – 0:50

Given the three points, $(1,2)$, $(3,-4)$, and $(5,-1)$, calculate the new line of best fit.

1) Skip to the Normal Equation and add the new point to the matrices and vector from (5) above.

$$A^T A \begin{pmatrix} a \\ b \end{pmatrix} = A^T \vec{y} \quad \Rightarrow \quad \begin{pmatrix} 1 & 3 & 5 \\ 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 3 & 1 \\ 5 & 1 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} 1 & 3 & 5 \\ 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} 2 \\ -4 \\ -1 \end{pmatrix}$$

2) Use matrix multiplication again to simplify and take the inverse.

$$\begin{pmatrix} 35 & 9 \\ 9 & 3 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} -15 \\ -3 \end{pmatrix} \qquad B^{-1} = \frac{1}{24} \begin{pmatrix} 3 & -9 \\ -9 & 35 \end{pmatrix}$$

3) Solve for the new a & b and find the new line of best fit equation. Plot the new point and line on your graph.

$$\begin{pmatrix} a \\ b \end{pmatrix} = \frac{1}{24} \begin{pmatrix} 3 & -9 \\ -9 & 35 \end{pmatrix} \begin{pmatrix} -15 \\ -3 \end{pmatrix} \quad \Rightarrow \quad \begin{pmatrix} a \\ b \end{pmatrix} = \frac{1}{24} \begin{pmatrix} -22 \\ 30 \end{pmatrix} \quad \Rightarrow \quad \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} -\frac{11}{12} \\ \frac{5}{4} \end{pmatrix} \qquad y = -\frac{11}{12}x + \frac{5}{4}$$

4) Intuitively, which line has the smallest error for the first two points? For the three points? Which would you prefer to use for predicting a 4th point (assuming the same data generating process)?

Line from the first problem: $y = -3x + 5$. Line from the second problem: $y = -\frac{11}{12}x + \frac{5}{4}$. Second.

5) Calculate the sum of squared residuals for the three points using the line made from those points (Hint: $\vec{s} = |\vec{w} - \vec{y}|$).

$$\vec{s} = |\vec{w} - \vec{y}| = \sqrt{(f(x_1) - y_1)^2 + (f(x_2) - y_2)^2 + (f(x_3) - y_3)^2}$$

$$\vec{w} = \begin{pmatrix} f(x_1) \\ f(x_2) \\ f(x_3) \end{pmatrix} = \begin{pmatrix} -\frac{11}{12}x_1 + \frac{5}{4} \\ -\frac{11}{12}x_2 + \frac{5}{4} \\ -\frac{11}{12}x_3 + \frac{5}{4} \end{pmatrix} = \begin{pmatrix} \frac{1}{3} \\ -\frac{3}{2} \\ -\frac{10}{3} \end{pmatrix}$$

$$\vec{s} = \sqrt{\left(\frac{1}{3} - 2\right)^2 + \left(-\frac{3}{2} + 4\right)^2 + \left(-\frac{10}{3} + 1\right)^2} = \sqrt{\frac{511}{36}} = 3.768$$

6) Calculate the sum of squared residuals for the three points using the line made from the first two points only.

$$\vec{w} = \begin{pmatrix} f(x_1) \\ f(x_2) \\ f(x_3) \end{pmatrix} = \begin{pmatrix} -3x_1 + 5 \\ -3x_2 + 5 \\ -3x_3 + 5 \end{pmatrix} = \begin{pmatrix} 2 \\ -4 \\ -10 \end{pmatrix}$$

$$\vec{s} = \sqrt{(2-2)^2 + (-4+4)^2 + (-10+1)^2} = \sqrt{81} = 9$$

Optional Break (5 minutes) 0:50 – 0:55

Line of Best Fit with 4-Points: (15 minutes) 0:55 – 1:10

Given four new points, $(2,3), (-1,-2), (4,8), (-3,1)$, calculate the line of best fit.

1) Skip directly to the symmetric, square matrix B, in the equation: $B\begin{pmatrix}a\\b\end{pmatrix} = A^T\vec{y}$.
   a. Hint: $\begin{pmatrix}b_{11} & b_{12}\\ b_{21} & b_{22}\end{pmatrix}$, where $b_{12} = b_{21}$, and $b_{22} = n$.

$$b_{11} = \sum_{i=1}^{n} x_i^2\ , b_{12} = \sum_{i=1}^{n} x_i \qquad A^T\vec{y} = \vec{Y} = \begin{matrix}Y_1\\Y_2\end{matrix} \qquad Y_1 = \sum_{i=1}^{n} x_i y_i \qquad Y_2 = \sum_{i=1}^{n} y_i$$

$$\begin{pmatrix}30 & 2\\ 2 & 4\end{pmatrix}\begin{pmatrix}a\\b\end{pmatrix} = \begin{pmatrix}2 & -1 & 4 & -3\\ 1 & 1 & 1 & 1\end{pmatrix}\begin{pmatrix}3\\-2\\8\\1\end{pmatrix} \Rightarrow \begin{pmatrix}30 & 2\\ 2 & 4\end{pmatrix}\begin{pmatrix}a\\b\end{pmatrix} = \begin{pmatrix}37\\10\end{pmatrix}$$

2) Use an augmented matrix and Gaussian elimination to solve for $\begin{pmatrix}a\\b\end{pmatrix}$.

$$\left(\begin{array}{cc|c}30 & 2 & 37\\ 2 & 4 & 10\end{array}\right) \xRightarrow{II-\frac{1}{15}I} \left(\begin{array}{cc|c}30 & 2 & 37\\ 0 & \frac{58}{15} & \frac{113}{15}\end{array}\right) \Rightarrow\ 58b = 113 \Rightarrow b = \frac{113}{58} = 1.948$$

$$30a + 2(1.948) = 37 \Rightarrow 870a = 960 \ \Rightarrow\ a = 1.103$$

3) Plot the points and line of best fit to confirm your solution.

$$y = 1.103a + 1.948$$

4) Use the line of best fit to calculate $\vec{w}$, then use this to calculate $\vec{s}$. Plot the points (x_i, w_i) on your graph, along with dotted lines representing the residuals for each point.

$$\vec{w} = \begin{pmatrix}f(x_1)\\f(x_2)\\f(x_3)\\f(x_4)\end{pmatrix} = \begin{pmatrix}1.103x_1 + 1.948\\1.103x_2 + 1.948\\1.103x_3 + 1.948\\1.103x_4 + 1.948\end{pmatrix} = \begin{pmatrix}4.155\\0.8446\\6.3618\\-1.362\end{pmatrix}$$

$$\vec{s} = \sqrt{(4.155-3)^2 + (0.8446+2)^2 + (6.3618-8)^2 + (-1.362-1)^2} = \sqrt{17.6885} = 4.2057$$

Alternative Method for Line of Best Fit: (15 minutes) 1:10 – 1:25

Use the alternative formulas to find the line of best fit for the four points above:

$$a = \frac{\overline{xy} - \bar{x}\,\bar{y}}{\overline{x^2} - \bar{x}^{\,2}} \qquad b = \bar{y} - a\,\bar{x} \qquad \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \quad \overline{xy} = \frac{\sum_{i=1}^{n} x_i\, y_i}{n} \qquad \overline{x^2} = \frac{\sum_{i=1}^{n} x_i^2}{n}$$

1) Calculate the sample mean of the x-values, $\bar{x}$, and y-values,$\bar{y}$.

$$\bar{x} = \frac{2-1+4-3}{4} = \frac{1}{2} = 0.5 \;\; \bar{y} = \frac{3-2+8+1}{4} = \frac{5}{2} = 2.5$$

2) Calculate the sample mean of the product of the x- and y-values, $\overline{xy}$.

$$\overline{xy} = \frac{6+2+32-3}{4} = \frac{37}{4} = 9.25$$

3) Calculate the sample mean of the x-values squared, $\overline{x^2}$.

$$\overline{x^2} = \frac{4+1+16+9}{4} = \frac{15}{2} = 7.5$$

4) Use these means to calculate a and b. Compare with the equation you found using matrices.

$$a = \frac{9.25-0.5*2.5}{7.5-0.25} = \frac{8}{7.25} = 1.103 \qquad b = 2.5 - 1.103 * 0.5 = 2.5 - 0.5517 = 1.948$$

$$y = 1.103x + 1.948$$

5) What technologies do we have to calculate these values and conduct regression analyses?

Some calculators, Excel, Google Sheets, R, Python, Stata, Matlab, etc. etc.

Bonus! (if time allows)

For those with knowledge of linear regression in statistics, what is the corollary formula?

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \qquad \hat{Y} = \widehat{\beta_0} + \widehat{\beta_1} x_i$$

What is the correlation coefficient? Variance? Standard error?

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n\,\sigma_x\,\sigma_y} \qquad \sigma_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \quad \sigma_x = \sqrt{\sigma_x^2}$$

How can we use these to arrive at our best fit equation?

$$\beta_1 = \frac{r\,\sigma_y}{\sigma_x} \qquad \beta_0 = y - \beta_1 \bar{x}$$