Linear Regression

I211: Information infrastructure II
Linear Relationships

Consider 2 points of the form \((x, y)\). They are \((2, 3)\) and \((4, 8)\).

Question: how to make a line out of these two points?

\[
y = k \cdot x + n
\]
Consider 2 points of the form \((x, y)\). They are \((2, 3)\) and \((4, 8)\).

Question: how to make a line out of these two points?

\[ y = k \cdot x + n \]

Rise:
\[ k = \frac{p}{q} \]
Consider 2 points of the form \((x, y)\). They are \((2, 3)\) and \((4, 8)\).

Question: how to make a line out of these two points?

Intercept: 
\[ n = -2 \cdot k + 3 \]
Matlab Code

clear
clc

% define 2 data points a and b
a = [2 3];
b = [4 8];

% calculate rise and intercept
k = (b(2) - a(2))/(b(1) - a(1));
n = -k * a(1) + a(2);

% calculate points of the line
x = 0 : 0.1 : 8;
y = k * x + n;

% plot the line and two data points
plot(x, y, a(1), a(2), '*'
    b(1), b(2), '*')
axis([0 8 -4 18])
Linear relationship given two points \( a = (a_1, a_2) \) and \( b = (b_1, b_2) \) is calculated as:

\[
y - a_2 = \frac{b_2 - a_2}{b_1 - a_1} \cdot (x - a_1)
\]
Consider now 3 points of the form \((x, y)\). They are \(a = (2, 3)\), \(c = (3, 5.5)\) and \(b = (4, 8)\).

Let’s calculate rise using points \(a\) and \(c\):

\[
k_{ac} = \frac{5.5 - 3}{3 - 2} = 2.5
\]

Let’s calculate rise using points \(b\) and \(c\):

\[
k_{bc} = \frac{5.5 - 8}{3 - 4} = 2.5
\]

**Conclusion:**

If three (or more) points are co-linear, it does not matter which two we use to construct a line.
3 Data Points that are not Co-Linear

Consider 3 different points of the form \((x, y)\). They are \(a = (2, 3)\), \(b = (4, 8)\), and \(c = (3.5, 5.5)\). Clearly, we cannot draw a line that goes through all three data points.

Let’s calculate rise using points \(a\) and \(c\):

\[
k_{ac} = \frac{5.5 - 3}{3.5 - 2} = 1.67
\]

Let’s calculate rise using points \(b\) and \(c\):

\[
k_{bc} = \frac{5.5 - 8}{3.5 - 4} = 5
\]
Let's Get Back to Original Problem

If we had 1st component of vector c and used the line to predict its 2nd component, e would be a prediction error.
What if a and c are Used for Line?

Square error for each point:

\[
\begin{align*}
\text{d}^2(a, \text{line}) &= 0 \\
\text{d}^2(b, \text{line}) &= 1.67^2 = 2.78 \\
\text{d}^2(c, \text{line}) &= 0
\end{align*}
\]
Given points $a$, $b$ and $c$, find line $y = k \cdot x + n$ such that $(e_1^2 + e_2^2 + e_3^2)$ is minimized!
Application: Measuring Acceleration

Idea: apply force \((F)\) to an object and measure its acceleration \((a)\).

Use many different forces \((F_1, F_2, \ldots)\) and measure acceleration for each force \((a_1, a_2, \ldots)\). This will be our data set collection process.

Picture from: https://wiki.brown.edu/confluence/display/PhysicsLabs/LAB+3+-+FORCE+AND+ACCELERATION
Look at Data Points

Step:

- Collect data points (a, F)
- Do linear regression
- If you have a good instrument and do it right, your rise should be equal to object’s mass and intercept should be zero
Problem

Given:

\[ \text{data set } D = \{(x, y)_i, i = 1..m\} \]

Find:

best \( k \) and \( n \) (from the line equation \( y = k \cdot x + n \))

Let’s expand \( x \) into vector \( x = [1 \ x]^T \).

Let’s use \( \beta = [n \ k]^T \)

Now,

\[ [n \ k] \cdot \begin{bmatrix} 1 \\ x \end{bmatrix} = n + k \cdot x \quad \rightarrow \quad \beta = \begin{bmatrix} n \\ k \end{bmatrix}, \quad x = \begin{bmatrix} 1 \\ x \end{bmatrix} \]

\[ y = \beta^T \cdot x = k \cdot x + n \]
Problem, more generally

Given:

data set $D = \{(x, y), i = 1..m\}$

Find:

best $\beta$ such that $y' = \beta^T \cdot x$

Solution (without derivation):

$$\beta = (X^T X)^{-1} X^T y$$
Problem, more generally

What is matrix $X$?
What is vector $y$?

Add 1 to each point

$$D = \begin{bmatrix}
3 & 5 & 2.3 \\
7 & 4 & 7.2 \\
2 & 7 & 2.5 \\
5 & 4 & 5.2 \\
9 & 4 & 8.4
\end{bmatrix}$$

$$X = \begin{bmatrix}
1 & 3 & 5 \\
1 & 7 & 4 \\
1 & 2 & 7 \\
1 & 5 & 4 \\
1 & 9 & 4
\end{bmatrix}$$

$$y = \begin{bmatrix}
2.3 \\
7.2 \\
2.5 \\
5.2 \\
8.4
\end{bmatrix}$$
Back to a Familiar Example

\[ b = (4, 8) \]

\[ a = (2, 3) \]

\[ c = (3.5, 5.5) \]

\[ y = k \cdot x + n \]

\[
\begin{bmatrix}
1 & 2 \\
1 & 3.5 \\
1 & 4 \\
\end{bmatrix}
\]

\[
\begin{bmatrix}
3 \\
5.5 \\
8 \\
\end{bmatrix}
\]
clear
clc

% define 2 data points a and b
a = [2 3];
b = [4 8];
c = [3.5 5.5];

% define X and y
X = [1 2; 1 4; 1 3.5];
y = [3; 8; 5.5];

% calculate optimal coefficients
beta = inv((X' * X)) * X' * y;

% calculate points of the line
x = 0 : 0.1 : 8;
y = beta(2) * x + beta(1);

% plot the line and two data points
plot(x, y, a(1), a(2), '*', ...
    b(1), b(2), '*', c(1), c(2), '*')
axis([0 8 -4 18])

\[
\beta = \begin{bmatrix}
-1.8077 \\
2.3077
\end{bmatrix}
\]

\[
y = \beta^T \cdot x = 2.3 \cdot x - 1.8
\]
Calculate Errors

\[
y = k \cdot x + n
\]

Squared errors:

\[
d^2(a, \text{line}) = e_1^2 = 0.19^2 = 0.04
\]
\[
d^2(b, \text{line}) = e_3^2 = 0.58^2 = 0.33
\]
\[
d^2(c, \text{line}) = e_2^2 = 0.77^2 = 0.59
\]

We minimized sum of squared errors!!!

\[
e_1^2 + e_2^2 + e_3^2 = 0.96
\]