Least Squares

Setup:
- Given data points from observations, tests
- Try to make sense of them as (roughly!) lying on a line
- Optimize the “sameness” of the line and the data points.

Decreasing one error usually increases another error
Minimize combined errors
Another option: add up residuals (errors) squared
\((y_1 - y_2)^2\), \(y_1\) is data’s value, \(y_2\) is point on the line
Both have uses; we’ll focus on this one
Optimization
Least Squares

Least Squares - First Example

Minimize:

\[(1 - (2m + b))^2 + (3 - (4m + b))^2 + (2 - (6m + b))^2\]

Expand and simplify:

\[SSR = 14 - 52m - 12b + 3b^2 + 56m^2 + 24mb\]

Problem: two variables.

\[R(b) = 3b^2 + (24m - 12)b + (14 - 52m + 56m^2)\]
\[R(m) = 56m^2 + (24b - 52)m + (14 - 12b + 3b^2)\]

SSR: Sum of Squared Residuals

Given one variable, we can find the other that makes R minimal. (Only CP of parabola)

- \[R'(b) = 6b + 24m - 12\], so R is minimum when \(b = 2 - 4m\)
- \[R'(m) = 112m + 24b - 52\], so R is minimum when:
  - \(112m = 52 - 24b\)
  - Then \(b = 2 - 4m = 2 - 4\left(\frac{13 - 6b}{28}\right)\)
  - That is, \(b = 2 - \frac{13 - 6b}{7}\), so \(b = 1\)
  - Then \(m = \frac{13 - 2}{14} = \frac{11}{14} = \frac{1}{2}\)
- That is, our line with the least sum of squared residuals is

\[y = \frac{1}{4}x + 1\]
Example 1:

Find a line passing through the origin that has the least sum of squared residuals for the following points:

\((-1, 2)\) \(\left(0, \frac{1}{2}\right)\) \((2, -2)\)
Formulas: Line through the origin

Best fit: $y = ax$

Given the data points $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, the line through the origin with the least sum of squared residuals is given by $y = ax$, where

$$a = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$$

Example: $(-1, 2), (0, 1/2), (2, -2)$:

$$a = \frac{\sum_{i=1}^{3} x_i y_i}{\sum_{i=1}^{3} x_i^2} = \frac{(-1)(2) + (0)(1/2) + (2)(-2)}{(-1)^2 + 0^2 + (-2)^2} = \frac{-6}{5}$$

$$y = \frac{-6}{5}x$$

Notes

Formulas: Line through the origin

Best fit: $y = ax$

Given the data points $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, the line through the origin with the least sum of squared residuals is given by $y = ax$, where

$$a = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$$

Example: $(2, -4), (3, -5), (4, 1), (5, -7)$:

Notes

Formulas: Line

Best fit: $y = ax + b$

Given the data points $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, the line with the least sum of squared residuals is given by $y = ax + b$, where:

- $a = \frac{P_{avg} - \bar{y} \cdot \bar{x}}{X_{avg}^2 - \bar{x}^2}$
- $b = \bar{y} - a \bar{x}$
- $P_{avg} = \frac{1}{n} \sum_{i=1}^{n} x_i y_i$
- $X_{avg} = \frac{1}{n} \sum_{i=1}^{n} x_i$

"average of the products minus the product of the averages"

"average of the squares minus square of the average"

Example: $(-1, 2), (0, 1/2), (2, -2)$:

- $\bar{x} = \frac{-1 + 0 + 2}{3} = \frac{1}{3}$
- $\bar{y} = \frac{1}{3} \sum_{i=1}^{3} y_i = \frac{1}{3} (2 + 1/2 + (-2)) = \frac{1}{3}$
- $P_{avg} = \frac{(-1)(2) + (0)(1/2) + (2)(-2)}{3} = -2$
- $X_{avg}^2 = \frac{(-1)^2 + 0^2 + (-2)^2}{3} = \frac{5}{3}$

$$28y = 17 - 37x$$

Notes
Least Squares Formula

\[ y = \frac{6}{5}x \]

\[ 28y = 17 - 37x \]

Larger Data Sets

A data set contains yearly enrollment in (1) school lunch programs and (2) summer food programs for all 50 states and the District of Columbia. Using the data from 2015, can we predict summer enrollment in 2017 from school lunch enrollment?

2015 data in a google sheet - link
Worked example with linear regression - link

Source: Food data atlas – United States Department of Agriculture
Economic Research Service