Best Fit Line

Jonathan Auerbach

9/28/2018

Scientists often want to summarize one variable as a simple function of another variable

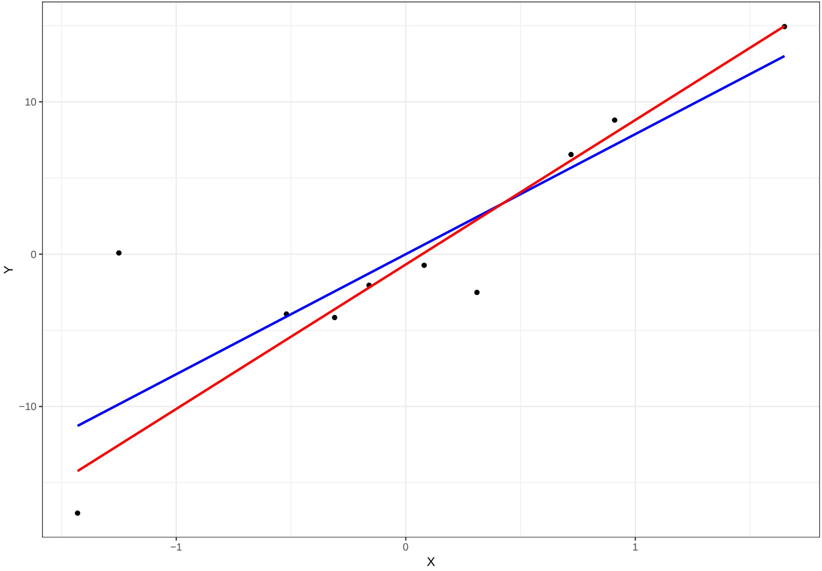
- Suppose you observed n pairs of random variables: X and Y.
- For example, you observe the heights of 10 child/parent pairs, and you want to communicate to a new parent how tall their child will likely be.
- You could list all 10 observed pairs you observed:

$$(X_1, Y_1), (X_2, Y_2), (X_3, Y_3), (X_4, Y_4), (X_5, Y_5),$$

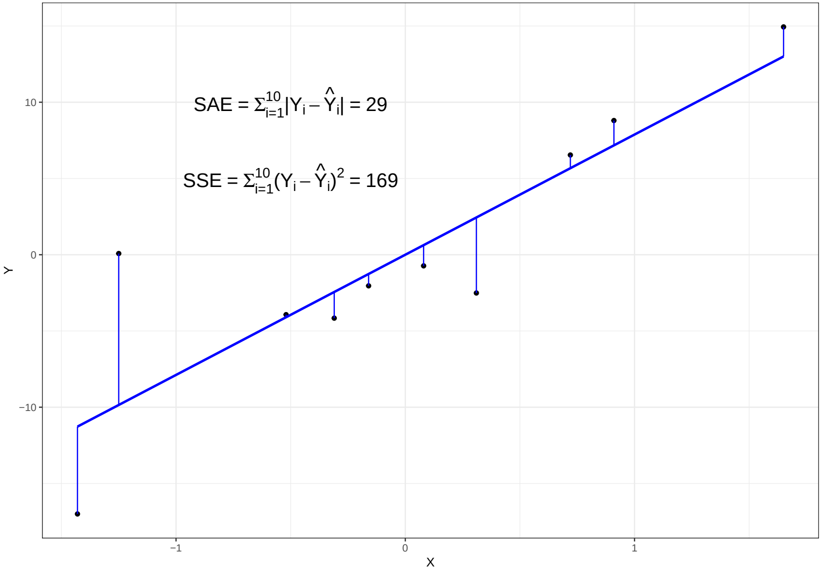
$$(X_6, Y_6), (X_7, Y_7), (X_8, Y_8), (X_9, Y_9), (X_{10}, Y_{10})$$

- A simple summary of Y as a function of X is the straight line: $Y_i = \alpha + \beta X_i$

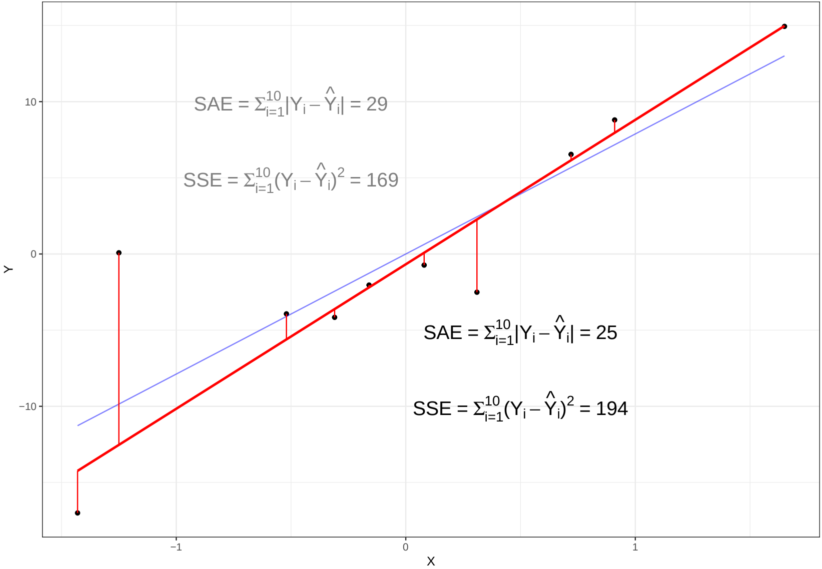
Which line is the best fit line? i.e. from which line would you make predictions, $\hat{Y}$, closest to the observed values Y?



Consider two measures of discrepancy: Sum of Absolute Errors (SAE) and Sum of Squared Errors (SSE)



The red line is the slope that results in the best SAE, and the blue line is the slope that results in the best SSE



Squared error is often used as an approximation to an arbitrary "smooth" measure of error

- Suppose we only had one observation: Y. How good is the prediction $\hat{Y}$?
- Let $f(\hat{Y})$ be any "smooth" measure of error. f takes a prediction as its argument, compares it to the actual outcome: Y, and returns a measure of discrepancy ≥ 0.
- We assume the discrepancy is 0 only if the prediction is the same as the outcome. $f(Y) = 0$ and $f'(Y) = 0$
- A taylor expansion of $f(\hat{Y})$ around Y gives the following approximation:

$$
\begin{aligned}
f(\hat{Y}) &\approx f(Y) + f'(Y)(Y - \hat{Y}) + \frac{1}{2}f''(Y)(Y - \hat{Y})^2 \\
&= 0 + 0 * (Y - \hat{Y}) + \frac{1}{2}f''(Y) \cdot (Y - \hat{Y})^2 \\
&\propto_{\hat{Y}} (Y - \hat{Y})^2
\end{aligned}
$$

The slope that minimizes the Sum of Squared Error (SSE) can be solved for directly

- Choose $\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} (Y_i - \beta X_i)^2$

$$
\begin{aligned}
0 &\overset{\text{set}}{=} \frac{d}{d\beta} \sum_{i=1}^{n} (Y_i - \beta X_i)^2 \\
&= \sum_{i=1}^{n} \frac{d}{d\beta} (Y_i - \beta X_i)^2 \\
&= \sum_{i=1}^{n} 2(Y_i - \beta X_i)(-X_i) \\
&= -2 \sum_{i=1}^{n} Y_i X_i + 2\beta \sum_{i=1}^{n} X_i^2 \\
\Rightarrow \hat{\beta} &= \frac{\sum_{i=1}^{n} Y_i X_i}{\sum_{i=1}^{n} X_i^2}
\end{aligned}
$$

- Minimum since second derivative: $2 \sum_{i=1}^{n} X_i^2 \geq 0$

References