Lecture 37 Simple Linear Regression - The Idea

BIO210 Biostatistics

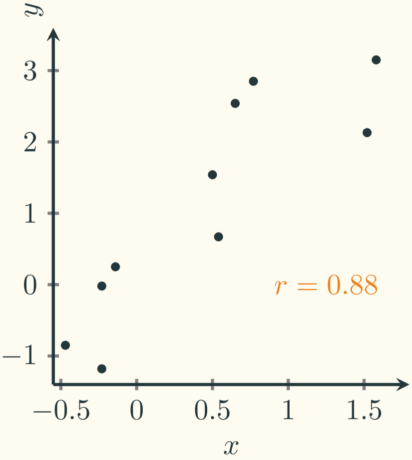
Xi Chen

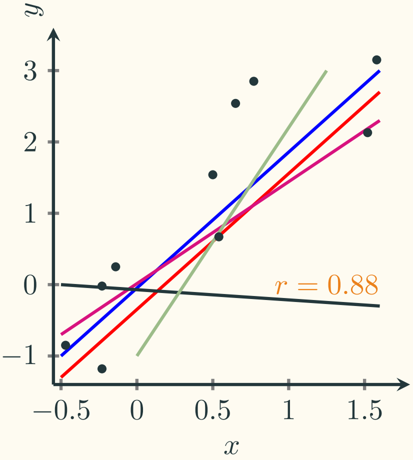
Spring, 2024

School of Life Sciences
Southern University of Science and Technology

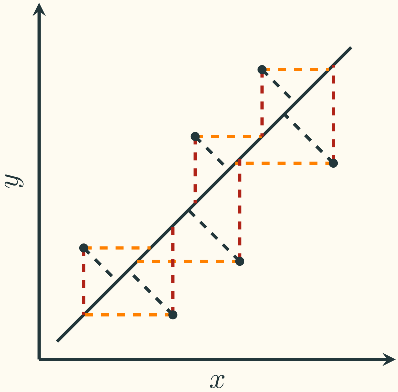
南方科技大学生命科学学院
SUSTech · SCHOOL OF
LIFE SCIENCES

Linear Regression





Best Fit Line

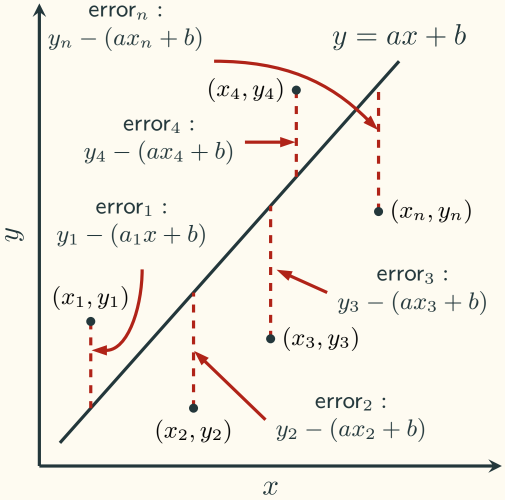


Our goal: minimise the "difference" between the data points and the line.

- Deming regression, PCA etc.
- Errors-in-variables models
- Ordinary least squares (OLS) regression

In practice: minimise squared distance.

Best Fit Line

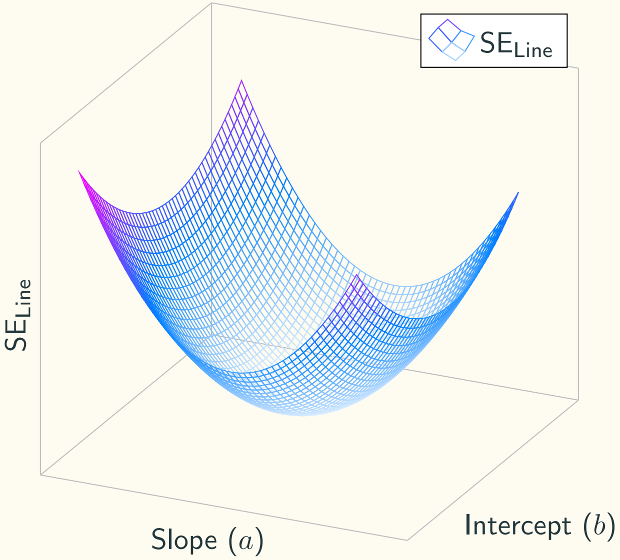


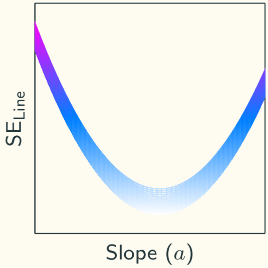
Squared error against the line (SE$_{\text{line}}$):

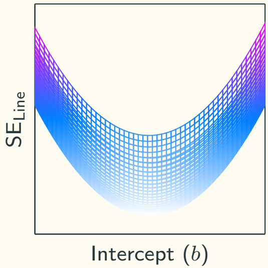
$$
\begin{aligned}
\text{SE}_{\text{line}} = & [\, y_1 - (ax_1 + b)\,]^2 + \\
& [\, y_2 - (ax_2 + b)\,]^2 + \\
& [\, y_3 - (ax_3 + b)\,]^2 + \\
& [\, y_4 - (ax_4 + b)\,]^2 + \\
& \vdots \\
& [\, y_n - (ax_n + b)\,]^2
\end{aligned}
$$

Find a, b to minimise this sum of squares

The Best Fit Line







Minimise SE_{line}

$$
\begin{aligned}
SE_{line} &= \sum_{i=1}^{n} [\, y_i - (ax_i + b)\,]^2 \\
&= \sum_{i=1}^{n} [\, y_i^2 - 2y_i(ax_i + b) + (ax_i + b)^2\,] \\
&= \sum_{i=1}^{n} (y_i^2 - 2ax_iy_i - 2by_i + a^2x_i^2 + 2abx_i + b^2) \\
&= \sum_{i=1}^{n} y_i^2 - 2a\sum_{i=1}^{n} x_iy_i - 2b\sum_{i=1}^{n} y_i + a^2\sum_{i=1}^{n} x_i^2 + 2ab\sum_{i=1}^{n} x_i + nb^2
\end{aligned}
$$

Minimise SE_{line}

$$SE_{line} = \left(\sum_{i=1}^{n} x_i^2\right) \cdot a^2 + 2\left(b\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} x_i y_i\right) \cdot a + \left(\sum_{i=1}^{n} y_i^2 - 2b\sum_{i=1}^{n} y_i + nb^2\right)$$

$$\Rightarrow \frac{\partial SE_{line}}{\partial a} = \left(2\sum_{i=1}^{n} x_i^2\right) \cdot a + 2\left(b\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} x_i y_i\right)$$

$$SE_{line} = n \cdot b^2 + \left(2a\sum_{i=1}^{n} x_i - 2\sum_{i=1}^{n} y_i\right) \cdot b + \left(a^2\sum_{i=1}^{n} x_i^2 - 2a\sum_{i=1}^{n} x_i y_i + \sum_{i=1}^{n} y_i^2\right)$$

$$\Rightarrow \frac{\partial SE_{line}}{\partial b} = 2n \cdot b + \left(2a\sum_{i=1}^{n} x_i - 2\sum_{i=1}^{n} y_i\right)$$

Minimise SE_{line}

$$\text{Let } \frac{\partial SE_{line}}{\partial a} = 0 \Rightarrow \left(2\sum_{i=1}^{n} x_i^2\right) \cdot a + 2\left(b\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} x_i y_i\right) = 0$$

$$\Rightarrow n \cdot \overline{x^2} \cdot a + b \cdot n \cdot \bar{x} - n \cdot \overline{xy} = 0$$

$$\Rightarrow \overline{x^2} \cdot a + b \cdot \bar{x} - \overline{xy} = 0$$

$$\text{Let } \frac{\partial SE_{line}}{\partial b} = 0 \Rightarrow 2n \cdot b + \left(2a\sum_{i=1}^{n} x_i - 2\sum_{i=1}^{n} y_i\right) = 0$$

$$\Rightarrow n \cdot b + a \cdot n \cdot \bar{x} - n \cdot \bar{y} = 0$$

$$\Rightarrow b + a \cdot \bar{x} - \bar{y} = 0$$

Best Fit Line: $y = ax + b$

Best fit line: $y = ax + b$

To minimise SE_{line}, we need:

$$\begin{cases} a \cdot \overline{x^2} + b \cdot \bar{x} - \overline{xy} = 0 \\ b + a \cdot \bar{x} - \bar{y} = 0 \end{cases} \Rightarrow \begin{cases} a \cdot \dfrac{\overline{x^2}}{\bar{x}} + b = \dfrac{\overline{xy}}{\bar{x}} \\ a \cdot \bar{x} + b = \bar{y} \end{cases}$$

Points $(\bar{x}, \bar{y})$ & $\left(\dfrac{\overline{x^2}}{\bar{x}}, \dfrac{\overline{xy}}{\bar{x}}\right)$ are on the best fit line !

Best Fit Line: $y = ax + b$

Best fit line: $y = ax + b$

$$\begin{cases} a \cdot \dfrac{\overline{x^2}}{\bar{x}} + b = \dfrac{\overline{xy}}{\bar{x}} \\ \\ a \cdot \bar{x} + b = \bar{y} \end{cases} \Rightarrow a\left(\bar{x} - \frac{\overline{x^2}}{\bar{x}}\right) = \bar{y} - \frac{\overline{xy}}{\bar{x}} \qquad \Rightarrow a = \frac{\bar{y} - \overline{xy}/\bar{x}}{\bar{x} - \overline{x^2}/\bar{x}} = \frac{\bar{x} \cdot \bar{y} - \overline{xy}}{(\bar{x})^2 - \overline{x^2}}$$

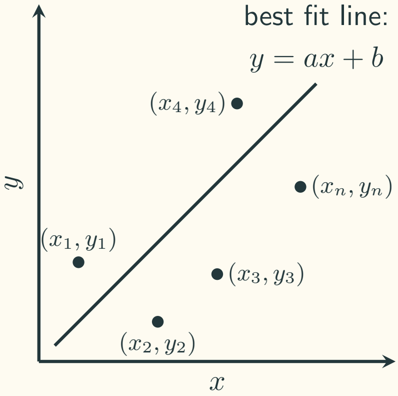
- More widely used form: $a = \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \dfrac{Cov(x, y)}{Cov(x, x)}$

Two Forms of The Slope

$$\begin{aligned}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) &= \sum_{i=1}^{n}(x_iy_i-\bar{x}\cdot y_i-\bar{y}\cdot x_i+\bar{x}\cdot\bar{y})\\ &= \sum_{i=1}^{n}x_iy_i-\bar{x}\sum_{i=1}^{n}y_i-\bar{y}\sum_{i=1}^{n}x_i+\sum_{i=1}^{n}\bar{x}\cdot\bar{y}\\ &= n\cdot\overline{xy}-\bar{x}\cdot n\cdot\bar{y}-\bar{y}\cdot n\cdot\bar{x}+n\cdot\bar{x}\cdot\bar{y}\\ &= n\cdot\overline{xy}-n\cdot\bar{x}\cdot\bar{y}\end{aligned}$$

$$\sum_{i=1}^{n}(x_i-\bar{x})^2=\sum_{i=1}^{n}\left[x_i^2-2\cdot\bar{x}\cdot x_i+(\bar{x})^2\right]=\sum_{i=1}^{n}x_i^2-2\cdot\bar{x}\sum_{i=1}^{n}x_i+\sum_{i=1}^{n}(\bar{x})^2$$
$$=n\cdot\bar{x}-2n\cdot(\bar{x})^2+n\cdot(\bar{x})^2=n\cdot\bar{x}-n\cdot(\bar{x})^2$$

Relationship Between The Slope & Pearson's r

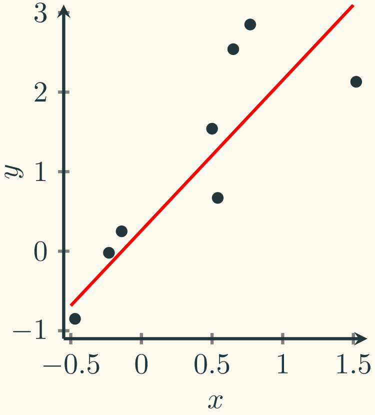


Using OLS regression:

$$\begin{cases} a = \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\ b = \bar{y} - a \cdot \bar{x} \end{cases}$$

$$\begin{aligned} a &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}} \\ &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}} \cdot \frac{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \\ &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \cdot \frac{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}} \\ &= r \cdot \frac{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2/(n-1)}}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2/(n-1)}} = r \cdot \frac{s_y}{s_x} \end{aligned}$$

Best Fit Line



x	y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})\cdot(y-\bar{y})$	$(x-\bar{x})^2$
0.5	1.54	0.05	0.43	0.0215	0.0025
-0.14	0.25	-0.59	-0.86	0.5074	0.3481
0.65	2.54	0.2	1.43	0.286	0.04
1.52	2.13	1.07	1.02	1.0914	1.1449
-0.23	-1.18	-0.68	-2.29	1.5572	0.4624
-0.23	-0.02	-0.68	-1.13	0.7684	0.4624
1.58	3.15	1.13	2.04	2.3052	1.2769
0.77	2.85	0.32	1.74	0.5568	0.1024
-0.47	-0.85	-0.92	-1.96	1.8032	0.8464
0.54	0.67	0.09	-0.44	-0.0396	0.0081

$$\bar{x} = 0.45,\ \bar{y} = 1.11,\ s_x = 0.72,\ s_y = 1.55$$

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