Least squares lines

EG Data points $(2,1)$, $(5,2)$, $(7,3)$, $(8,3)$

Determine line that "best fits".

$$y = a + bx$$ need to determine a, b

for perfect fit:

$(2,1)$: $1 = a + 2b$

$(5,2)$: $2 = a + 5b$

$(7,3)$: $3 = a + 7b$

$(8,3)$: $3 = a + 8b$

$$\begin{bmatrix} 1 & 2 \\ 1 & 5 \\ 1 & 7 \\ 1 & 8 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \\ 3 \\ 3 \end{bmatrix}$$

design matrix X — observation vector y

system $X \begin{bmatrix} a \\ b \end{bmatrix} = y$ inconsistent

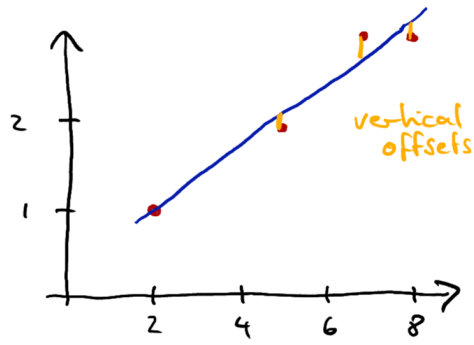
but: for "best fit" take least squares solution

$$X^T X \begin{bmatrix} a \\ b \end{bmatrix} = X^T y$$ normal equations

$$\begin{bmatrix} 4 & 22 \\ 22 & 142 \end{bmatrix} \quad \begin{bmatrix} 9 \\ 57 \end{bmatrix} \quad \rightsquigarrow \quad \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} 2/7 \\ 5/14 \end{bmatrix}$$

least squares line: $y = \frac{2}{7} + \frac{5}{14}x$

what kind of "best fit"?



vertical offsets

minimize error $X \begin{bmatrix} a \\ b \end{bmatrix} - y$

differences between y-values on our line and y-values of data points

$\Leftrightarrow$ minimize $\sum_i (a + bx_i - y_i)^2$

residual sum of squares

SS_{res}

x: explanatory variable

y: dependent variable

$\Leftrightarrow$ "minimizing error in predicting y through x"