

The Next Level in AP* Statistics: The Concepts Behind the Formulae

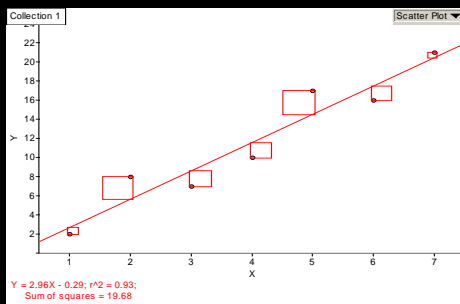
John Diehl
Hinsdale Central HS
Illinois

Robin Levine-Wissing
Glenbrook North HS
Illinois

USCOTS 2007 Breakout Session

How are the Slope and y-intercept Computed for Least-Squares Regression Lines?

A Diagram of Least-Squares



The Minimization Problem:

$$\begin{aligned} & \sum (y - \hat{y})^2 \\ &= \sum [y - (ax + b)]^2 \\ &= \sum [y^2 - 2y(ax + b) + (ax + b)^2] \\ &= \sum [y^2 - 2axy - 2by + a^2x^2 + 2abx + b^2] \\ &= \sum y^2 - 2a \sum xy - 2b \sum y + a^2 \sum x^2 + 2ab \sum x + nb^2 \end{aligned}$$

A Specific Example:

x	y	x ²	y ²	xy
1	2	1	4	2
2	8	4	64	16
3	7	9	49	21
4	10	16	100	40
5	17	25	289	85
6	16	36	256	96
7	21	49	441	147
28	81	140	1203	407

The function:

■ Minimize

$$Q = 1203 - 2a(407) - 2b(81) + a^2(140) + 2ab(28) + 7b^2$$

$$Q = 1203 - 814a - 162b + 140a^2 + 56ab + 7b^2$$

$$Q = 140a^2 + 56ab + 7b^2 - 814a - 162b + 1203$$

Calculus!

$$\frac{\partial Q}{\partial a} = 280a + 56b - 814$$

$$\frac{\partial Q}{\partial b} = 56a + 14b - 162$$

Keep going!

- If $\frac{\partial Q}{\partial a} = 0$ and $\frac{\partial Q}{\partial b} = 0$

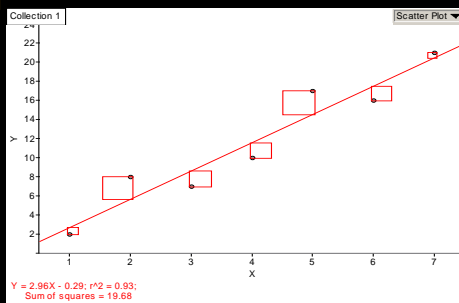
- Then $280a + 56b = 814$
 $56a + 14b = 162$

$$56a = 166$$

$$a = 166/56 = 2.96$$

$$b = 0.29$$

The actual graph



What are the Expressions in the Formulae for r^2 ?

Introduction to Statistics and Data Analysis (Peck, Olsen, Devore)

$$r^2 = 1 - \frac{SS_{Resid}}{SST_o}$$

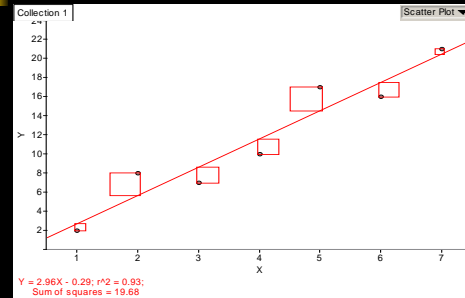
The Practice of Statistics (Yates, Moore, Starnes)

$$r^2 = \frac{SST - SSE}{SST}$$

Sum of Squared Residuals (Errors)

$$SS_{Resid} = SSE = \sum (y - \hat{y})^2$$

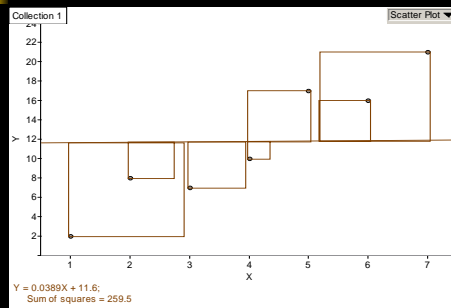
The picture of SSResid



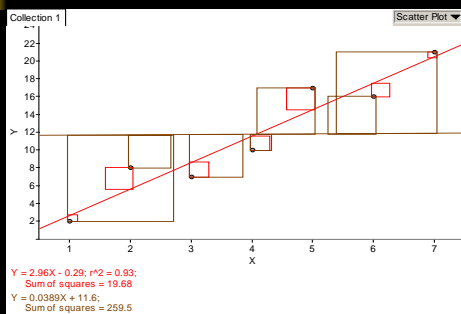
Sum of Squares Total

$$SSTo = SST = \sum (y - \bar{y})^2$$

The picture of SSTo



The picture of both sums

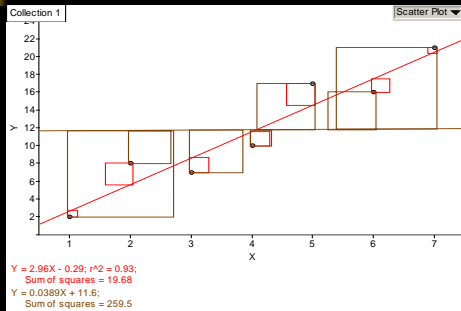


For our example:

$$r^2 = \frac{SST - SSE}{SST} = 1 - \frac{SSE}{SST}$$

$$= \frac{259.5 - 19.7}{259.5} = 1 - \frac{19.7}{259.5} = .93$$

There is another SS in this picture



“Predicted” variation

$$SST = \sum (y - \bar{y})^2$$

$$SSE = \sum (y - \hat{y})^2$$

$$SSP = \sum (\hat{y} - \bar{y})^2$$

The connection:

$$SST = SSE + SSP$$

$$\sum (y - \bar{y})^2 = \sum (y - \hat{y})^2 + \sum (\hat{y} - \bar{y})^2$$

$$\therefore r^2 = \frac{SST - SSE}{SST} = \frac{SSP}{SST}$$

Why do we use $\frac{\sigma}{\sqrt{n}}$ as the Standard Deviation of the Sampling Distribution of the Sample Mean?

Formulas for Independent Random Variables

$$\sigma_{X+Y}^2 = \sigma_X^2 + \sigma_Y^2$$


$$\sigma_{X/n}^2 = \frac{\sigma_X^2}{n^2}$$

If we use the same distribution, X


$$\sigma_{\frac{X+X+\dots+X}{n}}^2 = \frac{\sigma_X^2 + \sigma_X^2 + \dots + \sigma_X^2}{n^2}$$


$$= \frac{n\sigma_X^2}{n^2} = \frac{\sigma_X^2}{n}$$

$$\therefore \sigma_{\bar{X}}^2 = \frac{\sigma_X^2}{n} \text{ and } \sigma_{\bar{X}} = \frac{\sigma_X}{\sqrt{n}}$$



Why do we use $n-1$ in the formula for sample variance?


$$\sum \left(\frac{X - \mu}{\sigma} \right)^2 = \sum Z^2 = \chi^2(n)$$


$$\begin{aligned} \sum \left(\frac{X - \mu}{\sigma} \right)^2 &= \sum \left[\frac{(X - \bar{X}) + (\bar{X} - \mu)}{\sigma} \right]^2 \\ &= \sum \left(\frac{X - \bar{X}}{\sigma} \right)^2 + n \left(\frac{\bar{X} - \mu}{\sigma} \right)^2 \\ &\therefore \sum \left(\frac{X - \bar{X}}{\sigma} \right)^2 = \chi^2(n-1) \end{aligned}$$

$$\begin{aligned} E(\chi^2(n-1)) &= n-1 \\ E\left(\sum \left(\frac{X - \bar{X}}{\sigma} \right)^2\right) &= n-1 \\ E\left(\sigma^2 \sum \left(\frac{X - \bar{X}}{\sigma} \right)^2\right) &= \sigma^2(n-1) \\ E\left(\sum (X - \bar{X})^2\right) &= \sigma^2(n-1) \\ E\left(\frac{1}{n-1} \sum (X - \bar{X})^2\right) &= \sigma^2 \end{aligned}$$