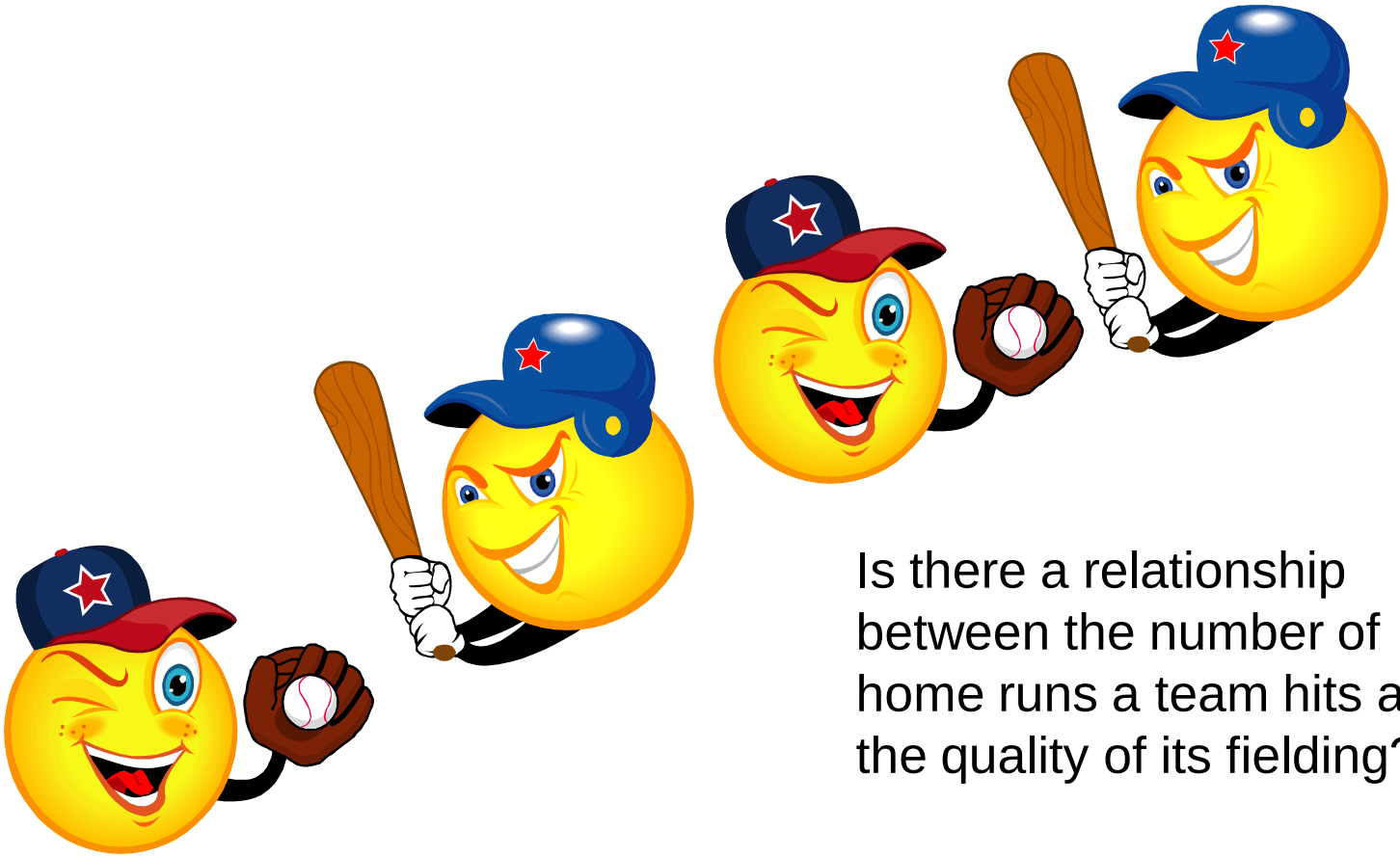
Fitting the Model: The Least Squares Approach

The Baseball Example

The Problem:

Can home runs be used to predict errors?



Is there a relationship between the number of home runs a team hits and the quality of its fielding?

The Data:

Home Runs (x)	Errors (y)	x_i^2	$x_i y_i$
158	126	24964	19908
155	87	24025	13485
139	65	19321	9035
191	95	36481	18145
124	119	15625	14756
$\sum x_i = 767$	$\sum y_i = 492$	$\sum x_i^2 = 120416$	$\sum x_i y_i = 75329$

The Solution:

Formulas for the Least Squares Estimates

$$\text{Slope: } \hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}} = \frac{\sum x_i y_i - \frac{\left(\sum x_i\right)\left(\sum y_i\right)}{n}}{\sum x_i^{\,2} - \frac{\left(\sum x_i\right)^2}{n}} = \frac{75,329 - \frac{(767)(492)}{5}}{120,416 - \frac{(767)^2}{5}}$$

$$= -.0521$$

$$y - \text{intercept: } \beta_0 = \overline{y} - \hat{\beta}_1 \overline{x} = 98.4 - (-.0521)(153.4) = 107.2$$

The Analysis:

We have the linear model of: $y = 107.2 - .0521x$

These results suggest that teams which hit more home runs are (slightly) better fielders (maybe not what we expected). There are, however, only five observations in the sample. It is important to take a closer look at the assumptions we made and the results we got. This is because the line has a negative slope or a negative linear relationship. So as the homeruns increase the number of errors decrease.