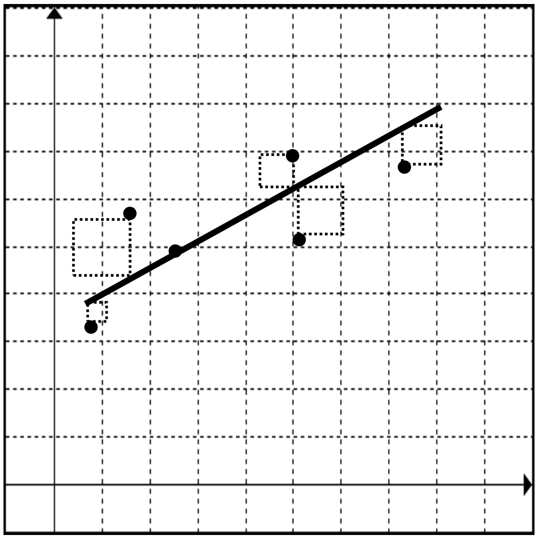
# MDM 4U
# LINEAR REGRESSION

When the linear correlation is strong, it is fairly easy to "eyeball" a good estimate of the line of best fit on a scatter plot. An analytic method using a ***least-squares fit*** gives more accurate results.

Consider the line of best fit in the scatter plot to the right.



The vertical dashed line from each point to the line shows the **residual** or **vertical deviation** of each data point from the line of best fit. Notice that residuals are positive for points above the line and negative for points below the line. The boxes show the squares of the residuals.

The line of best, using the least-squares fit, is determined by...

- Sum of residuals is zero – the positive and negative residuals cancel out.
- Sum of the squares of the residuals has the least possible value.

With some algebra, it can be shown that the equation of the line of best fit can be determined using the following formulas.

$$y = ax + b \qquad \text{where } a = \frac{n(\sum xy)-(\sum x)(\sum y)}{n(\sum x^2)-(\sum x)^2} \quad \text{and} \quad b = \bar{y} - a\bar{x}$$

For the data set below, complete the table, then use the formulas provided to determine the equation of the line of best fit for the hours of study time and percentage on tests.

| Hours of study time, $x$ | Percentage on a test, $y$ | $x^2$ | $y^2$ | $xy$ |
|---|---|---|---|---|
| 2 | 62 | | | |
| 4 | 80 | | | |
| 1 | 52 | | | |
| 6 | 92 | | | |
| 4 | 75 | | | |
| 8 | 95 | | | |
| Σ= | Σ= | Σ= | Σ= | Σ= |