SECTION 4.2 – Least Squares Regression

Least-Squares Regression Criterion

The **least-squares regression line** is the line that minimizes the sum of the squared errors (or residuals). This line minimizes the sum of the squared vertical distance between the observed values of y and those predicted by the line, $\hat{y}$ (read "y-hat"). We represent this as "minimize Σ residuals^2".

Line of Best Fit (aka The Least-Squares Regression Line)

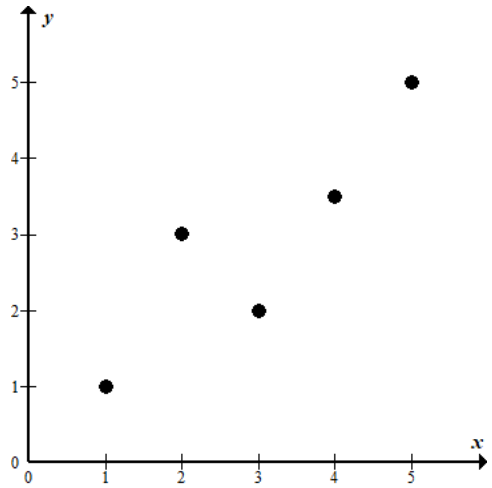
The **Line of Best Fit** or **The Least-Squares Regression Line** refers to a line through a scatter plot of data points that best expresses the relationship between those points. The least-squares regression criterion is the method to arrive at the geometric equation for the line.

☺ **Example #1**:

Consider the following five points:

x	1	2	3	4	5
y	1	3	2	3.5	5

a) By hand, draw a scatter diagram treating x as the explanatory variable and y as the response variable.



b) Select 3 different sets of two points from the scatter diagram and find the equation of the line containing the points selected.

So, let's choose the following:

Line *A*: Use points (2, 3) & (5, 5)
Line *B*: Use points (2, 3) & (4, 3.5)
Line *C*: Use points (1, 1) & (5, 5)

Now, in order to find the equation of a line given two points, we will need some formulas from Elementary Algebra:

Slope between two points: $m = \dfrac{y_2 - y_1}{x_2 - x_1}$ Point-Slope Form: $y - y_1 = m(x - x_1)$

SECTION 4.2 – Least Squares Regression

b) continued

Line *A*: Use points (2, 3) & (5, 5)

Slope: $m = \frac{5-3}{5-2} = \frac{2}{3}$

Point-Slope Form:

$$y - y_1 = m(x - x_1)$$
$$y - 3 = \frac{2}{3}(x - 2)$$
$$y - 3 = \frac{2}{3}x - \frac{4}{3}$$
$$y = \frac{2}{3}x + \frac{5}{3}$$

Line *B*: Use points (2, 3) & (4, 3.5)

Slope: $m = \frac{3.5-3}{4-2} = \frac{0.5}{2} = \frac{1}{4}$

Point-Slope Form:

$$y - y_1 = m(x - x_1)$$
$$y - 3 = \frac{1}{4}(x - 2)$$
$$y - 3 = \frac{1}{4}x - \frac{1}{2}$$
$$y = \frac{1}{4}x + \frac{5}{2}$$

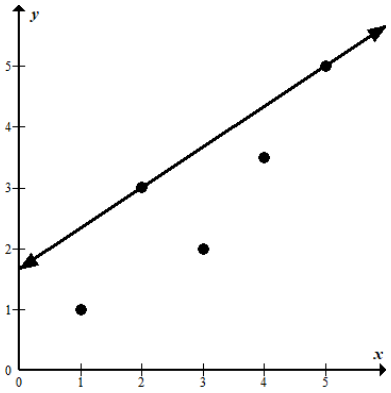
Line *C*: Use points (1, 1) & (5, 5)

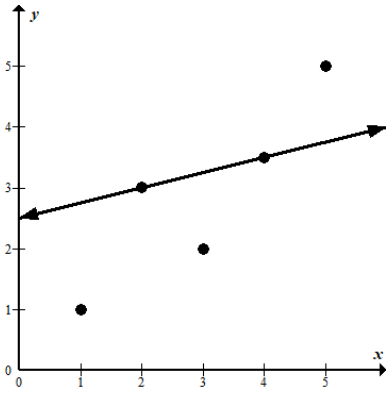
Slope: $m = \frac{5-1}{5-1} = \frac{4}{4} = 1$

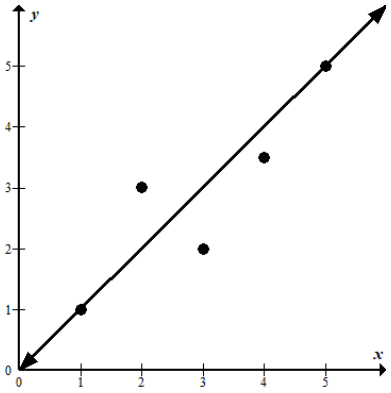
Point-Slope Form:

$$y - y_1 = m(x - x_1)$$
$$y - 1 = 1(x - 1)$$
$$y - 1 = x - 1$$
$$y = x$$

c) Graph the lines found in part **b)** on the scatter diagram.







SECTION 4.2 – Least Squares Regression

d) Compute the sum of the squared residuals for each line found in part **b)**.

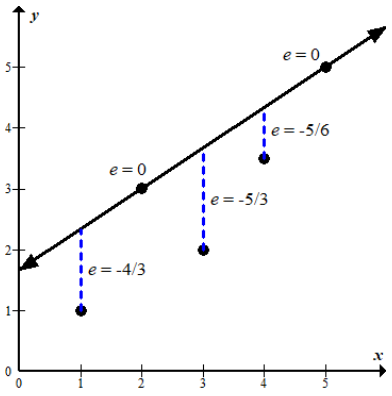
To measure quantitatively how well a line fits the data, we first consider the residuals or errors, *e*, made in using the line to predict the *y*-values of the data points.

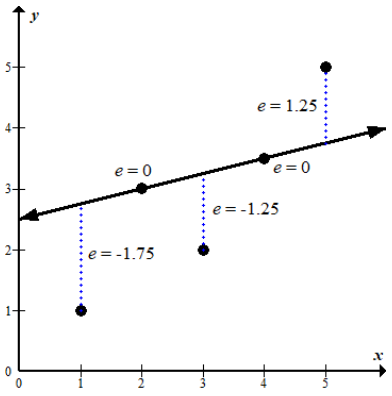
In line *A*,
the predicted values for our data, $(1,1),(2,3),(3,2),(4,3.5)$ and $(5,5)$, was $(1,\frac{7}{3}),(2,3),(3,\frac{11}{3}),(4,\frac{13}{3})$ and $(5,5)$.
(Here, we just simply substituted the x-values in the equation of line *A* which was $y=\frac{2}{3}x+\frac{5}{3}$).

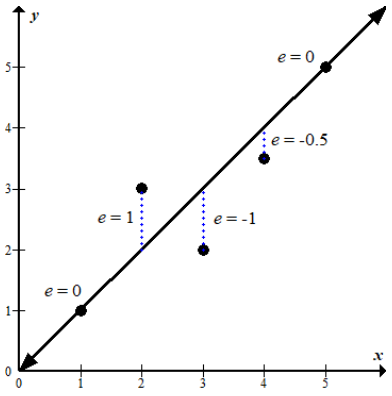
In line *B*,
the predicted values for our data, $(1,1),(2,3),(3,2),(4,3.5)$ and $(5,5)$, was $(1,2.75),(2,3),(3,3.25),(4,3.5)$ and $(5,3.75)$.
(Here, we just simply substituted the x-values in the equation of line *B* which was $y=\frac{1}{4}x+\frac{5}{2}$).

In line *C*,
the predicted values for our data, $(1,1),(2,3),(3,2),(4,3.5)$ and $(5,5)$, was $(1,1),(2,2),(3,3),(4,4)$ and $(5,5)$.
(Here, we just simply substituted the x-values in the equation of line *C* which was $y=x$).

Now, we calculate each error, *e*, or the residual, which is the signed vertical distance from the point on the line to each data point from the scatter diagram.







SECTION 4.2 – Least Squares Regression

d) continued

After calculating each error, e, or the residual, from there, those distances are squared, then summed.

Line *A*: $y = \frac{2}{3}x + \frac{5}{3}$				
x	y	$\hat{y}$	residual $= e = y - \hat{y}$	$e^2 = (y - \hat{y})^2$
1	1	7/3	– 4/3	16/9
2	3	3	0	0
3	2	11/3	– 5/3	15/9
4	3.5	13/3	– 5/6	25/36
3	5	5	0	0
				157/36 ≈ 4.36

Line *B*: $y = \frac{1}{4}x + \frac{5}{2}$				
x	y	$\hat{y}$	residual $= e = y - \hat{y}$	$e^2 = (y - \hat{y})^2$
1	1	2.75	– 1.75	3.0625
2	3	3	0	0
3	2	3.25	– 1.25	1.5625
4	3.5	3.5	0	0
3	5	3.75	1.25	1.5625
				6.1875

Line *C*: $y = x$				
x	y	$\hat{y}$	residual $= e = y - \hat{y}$	$e^2 = (y - \hat{y})^2$
1	1	1	0	0
2	3	2	– 1	1
3	2	3	1	1
4	3.5	4	0.5	0.5
3	5	5	0	0
				2.5

e) Which of the three lines is the better fit for the data.

The line with the smallest sum of squared errors is considered the line of best fit amongst those three lines. Thus, this process is called the least-squares regression criterion.

So, from the results above, line *C* has a smaller sum of squared errors or residuals, 2.5. Thus, line *C* is a better line of fit over both line *A* and line *B*. However, is line *C* the best line of fit? Probably not. The following formulas on the next page will instruct us on how to find the line of best fit.

SECTION 4.2 – Least Squares Regression

The Equation of the Least-Squares Regression Line

The equation of the least-squares regression line is given by

$$\hat{y} = b_1 x + b_0$$

where b_1 is the slope and b_0 is the *y*-intercept of the least-squares regression line.

Conceptual Formulas

$$b_1 = r \cdot \frac{s_y}{s_x}$$

Computational Formulas

$$b_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n}}{\sum x_i^2 - \frac{(\sum x_i)^2}{n}}$$

and

$$b_0 = \bar{y} - b_1 \bar{x}$$

$$b_0 = \frac{1}{n}(\sum y_i - b_1 \sum x_i)$$

Note: In the conceptual formulas, $\bar{x}$ is the sample mean and s_x is the sample standard deviation of the explanatory variable *x*; $\bar{y}$ is the sample mean and s_y is the sample standard deviation of the response variable *y*.

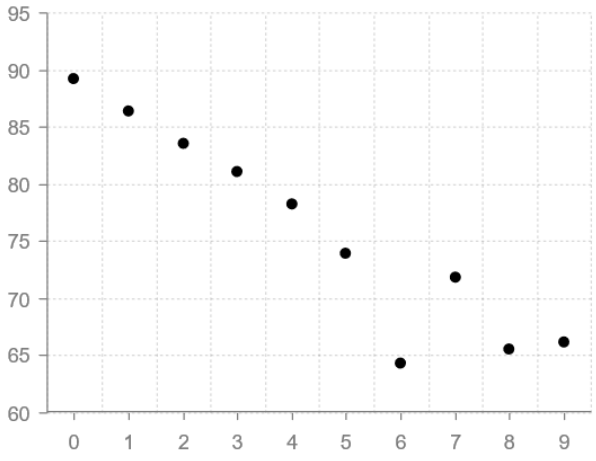
In the first conceptual formula, S_{xx} and S_{xy} are simply different indicators to denote the numerator and denominator, respectively.

☺ **Example #2**:

Attending Classes. The following data represent the number of days absent, *x*, and the final grade, *y*, for a sample of college students in a general education course at a large state university.

No. of absences, *x*	0	1	2	3	4	5	6	7	8	9
Final grade, *y*	89.2	86.4	83.5	81.1	78.2	73.9	64.3	71.8	65.5	66.2

a) Graph the scatter diagram based off the data points.



SECTION 4.2 – Least Squares Regression

b) Find the least-squares regression equation.

Since, the data points do appear to be scattered about a line, we will determine a regression line.
Here, we are going to use the computational formula.

First, construct the following table to the right.

No. of absences x	Final Grade y	xy	x^2
0	89.2	0	0
1	86.4	86.4	1
2	83.5	167	4
3	81.1	243.3	9
4	78.2	312.8	16
5	73.9	369.5	25
6	64.3	385.8	36
7	71.8	502.6	49
8	65.5	524	64
9	66.2	595.8	81
45	760.1	3187.2	285

Second, compute the slope b_1.

$$b_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n}}{\sum x_i^2 - \frac{(\sum x_i)^2}{n}} = \frac{3187.2 - \frac{(45)(760.1)}{10}}{285 - \frac{(45)^2}{10}} = \frac{3187.2 - 3420.45}{285 - 202.5} = \frac{233.25}{82.5} = -2.82\overline{72} \approx -2.83$$

Now, compute the y-intercept b_0.

$$b_0 = \frac{1}{n}(\sum y_i - b_1 \sum x_i) = \frac{1}{10}[760.1 - (-2.83)\cdot(45)] = \frac{1}{10}[760.1 - (-127.35)] = \frac{1}{10}[887.45] = 88.745 \approx 88.7$$

So, the regression line is $\hat{y} = -2.83x + 88.7$. This is considered the best fit line for the data.

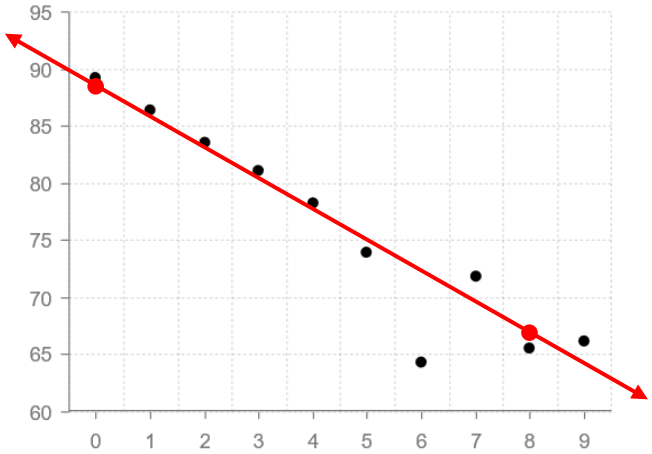
c) Interpret the slope and y-intercept, if appropriate.

Solution: → The slope is –2.83, so that means the final grade tends to decrease 2.83% for every 1 absence.

d) Draw the least squares regression line on the scatter diagram of the data.

Graphing by slope-intercept form can be rather tedious, since more than likely, both the slope and y-intercept are decimals in the tenths, hundredths, thousandths, etc. So, it is best to use a table using only 2 points: One being the intercept, if convenient, and another point of your choice, preferably a point that is somewhat towards the right (or end) of the scatter diagram. So, here we decided to choose $x = 0$ and $x = 8$ and solve for $\hat{y}$.

x	$\hat{y}$
0	88.7
8	66.06



e) Predict the final grade for a student who misses five class periods and compute the residual.

SECTION 4.2 – Least Squares Regression

☺ **Exercises**:

1) Recall the data set from Example #1:

x	1	2	3	4	5
y	1	3	2	3.5	5

a) Find the least-squares regression equation for the data.
b) Compute the correlation coefficient.
c) Determine whether there is a linear relation between x and y.