Statistical Reasoning
Scatterplots and Correlation

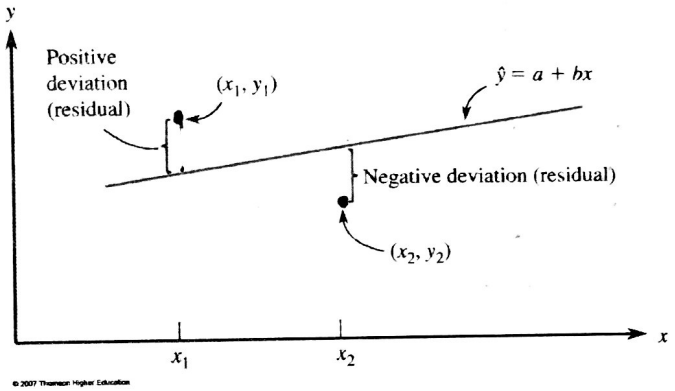
Name: Date: Class:

Regression Error: Residuals

In examining any graph of data, we look first for an overall pattern or shape and then for deviations in the pattern. With regression, the fitted line is the pattern and residuals are the deviations.

Residual = observed (point) value − predicted (line) value

$= y - \hat{y}$ (where y is the actual y-value and y-hat is the regressions y-value)

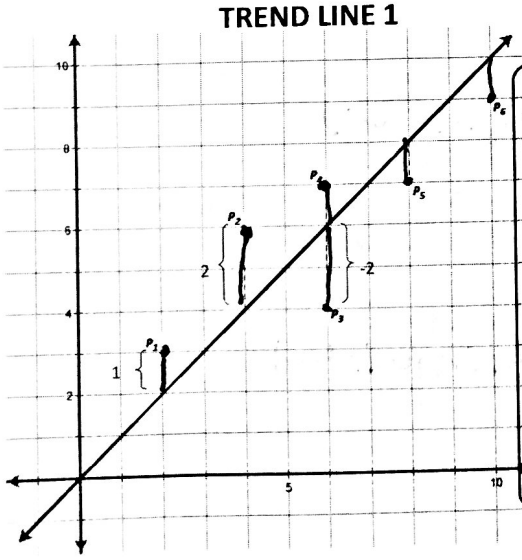


We use residuals to check the usefulness of our regression line

Most trend lines that are considered to be a "good fit" will be balanced such that the total **RESIDUAL** above and below the trend line is equal. **RESIDUAL** can be defined as the difference between the actual value (y) and expected value($\hat{y}$). A more succinct definition, **RESIDUAL** can be described as the vertical distance each data point is away from the trend line (with signed difference for above and below the trend line).

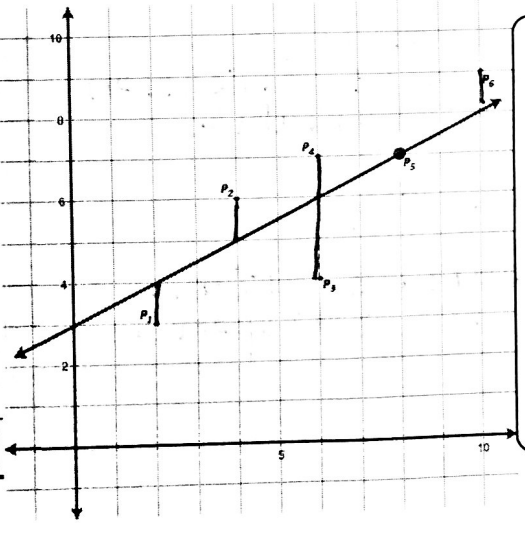
Find the **RESIDUAL**s for each of the TREND LINES below (the SCATTER PLOT is the **same** in each graph).

TREND LINE 1



Data Point	Residual
P_1	1
P_2	2
P_3	−2
P_4	1
P_5	−1
P_6	−1
Sum of Residuals	0

TREND LINE 2



Data Point	Residual
P_1	−1
P_2	1
P_3	−2
P_4	1
P_5	0
P_6	1
Sum of Residuals	0

What do all 2 trend lies above have in common?

To better analyze which trend line is best, it is common to consider comparing the sum of the squ of the residuals. Which trend line do you think is the best based on this new information? Is it the o you expected?

TREND LINE 1

Data Point	Residual	Squared	Residual Squared
P_1	1	1	1
P_2	2	4	4
P_3	−2	4	4
P_4	1	1	
P_5	−1	1	
P_6	−1	1	
Sum	0	12	

TREND LINE 2

Data Point	Residual	Residual Squared
P_1	−1	1
P_2	1	1
P_3	−2	4
P_4	1	1
P_5	0	0
P_6	1	1
Sum	0	8

Correlation ≠ Causation ! Just because two variables have a strong correlation does not mean that one causes the other— there may be an alternative explanation including chance.

For Example:

a) Given the question - Do storks cause babies? Stork population vs. babies born – in a small town in Germany, both populations increased over time
b) Given the question - Does smoking cause cancer?—difficult to prove although correlation exists

- association —general term, often used for categorical variables as well as numeric;
- correlation only applies to quantitative variables and is a statistical measure of direction and strength (r)
- lurking variable —a variable other than the two we are studying (x and y that influences both variables of interest, explaining the relationship between the two (often c common response —ice cream sales and number of people who drown both respond to hot weather)
- confounding variables—two variables are confounded when their effects on a response cannot be distinguished ; difficu to determine whether either of them is causal.

Also need to look at individual points that lie outside the pattern:

Outlier – a point that lies far from the fitted line and so produces a large residual. Outliers are points that do not follow the pattern of the other points in the dataset. These points typically have y-values that are either much above or below the least squares line. That is, they have large residuals in absolute value.

Find the equation of the least-squares regression line for the data set with and without the "outlier". Sketch the lines on the graph. What changed? Check the residuals. What do you notice?

Unusual Observations – "Internal" Outlier (II)

X	Y
2	5
3	4
3	7
5	5
5	8
7	5
8	8
10	7
10	10
6	**14**