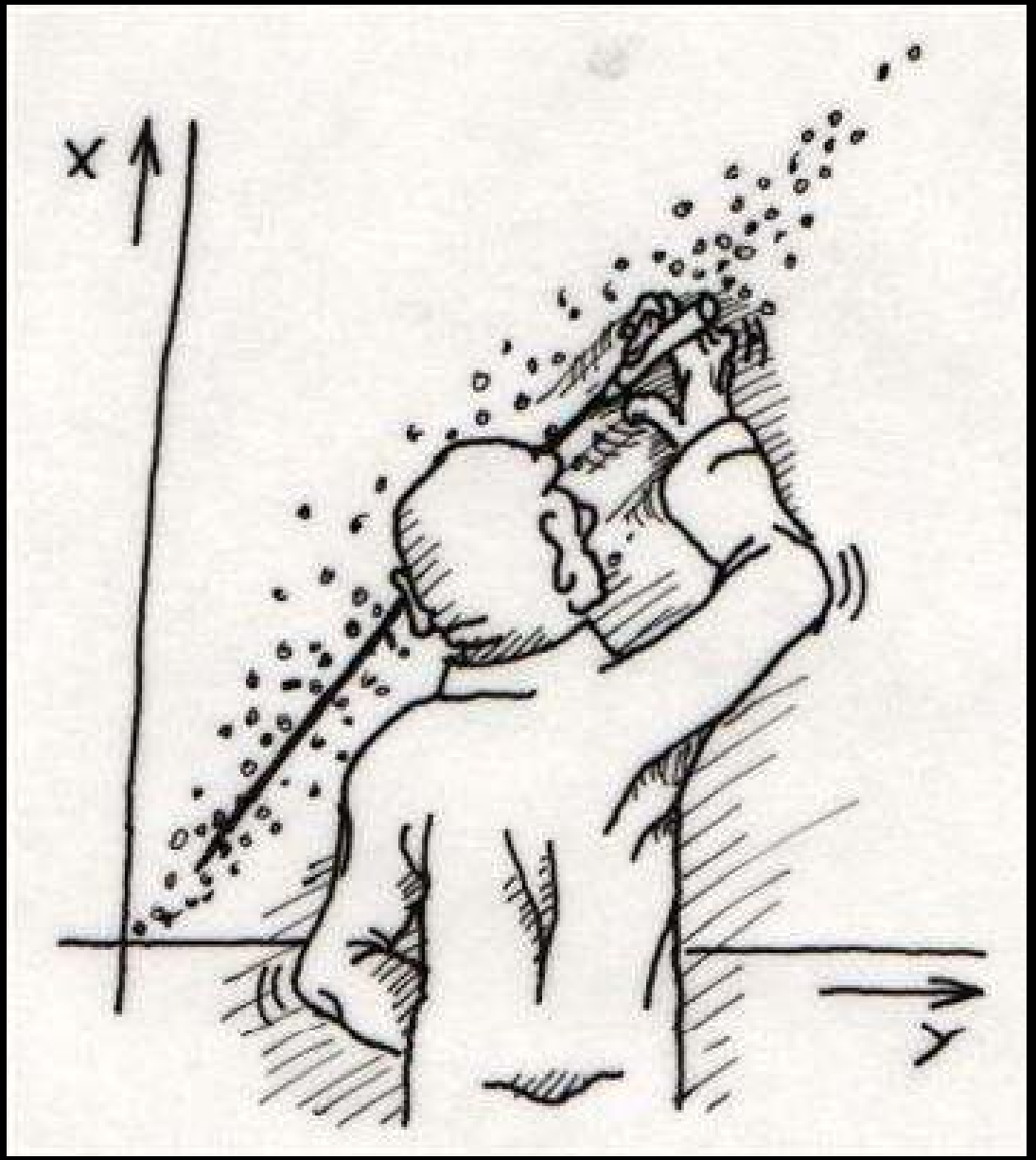
REGRESSION



Linear Regression

Linear regression is a statistical procedure that uses relationships to predict unknown *Y* scores based on the *X* scores from a correlated variable.

Predicted Y Scores

- The symbol Y' stands for a predicted Y score

- Each Y' is our best prediction of the Y score at a corresponding X, based on the linear relationship that is summarized by the regression line

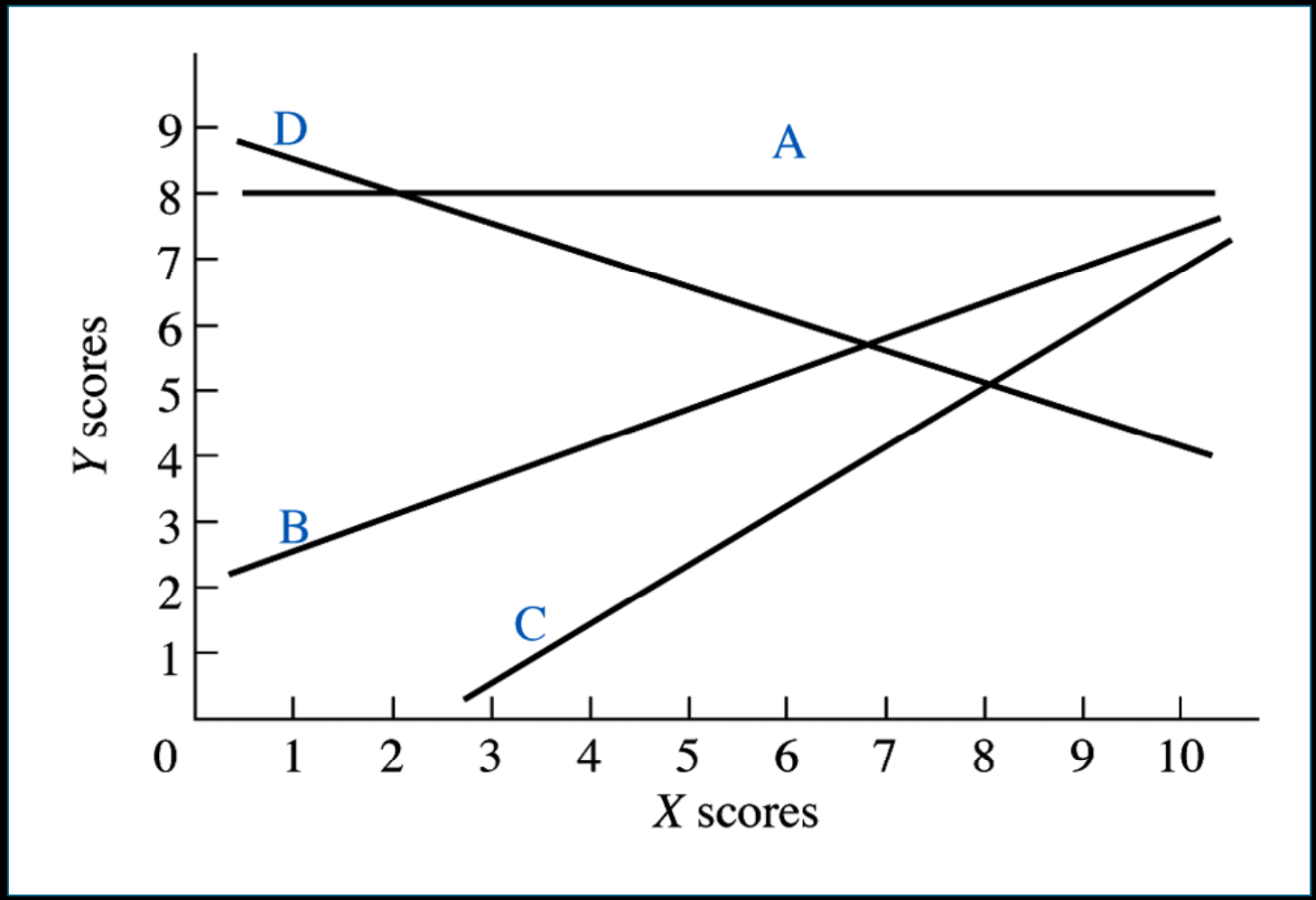
Linear Regression Line

- The linear regression line is the straight line that summarizes the linear relationship in the scatterplot by, on average, passing through the center of the *Y* scores at each *X*.

Slope and Intercept

- The slope is a number that indicates how slanted the regression line is and the direction in which it slants.
- The *Y*-intercept is the value of *Y* at the point where the regression line intercepts, or crosses, the *Y* axis (that is, when *X* equals 0).

Regression Lines Having Different Slopes and *Y* Intercepts



The Linear Regression Equation

- This equation indicates that the predicted Y values are equal to the slope (b) times a given X value and that this product then is added to the Y-intercept (a)

$$Y' = bX + a$$

Computing the Slope

- The formula for the slope (b) is

$$b = \frac{N(\Sigma XY) - (\Sigma X)(\Sigma Y)}{N(\Sigma X^2) - (\Sigma X)^2}$$

- Since we usually first compute r, the values of the elements of this formula already are known

Computing the *Y*-Intercept

- The formula for the *Y*-intercept (a) is

$$a = \overline{Y} - (b)(\overline{X})$$

Example 1

- For the following data set, calculate the linear regression equation.

X	*Y*
1	8
2	6
3	6
4	5
5	1
6	3

Plotting the Regression Line

- We compute predicted scores of Y from our X scores using the regression equation and plot them accordingly.

Example 2
Predicted *Y* Value

- Using the linear regression equation from example 1, determine the predicted *Y* score for $X = 4$.

Errors in Prediction

Variance

- The variance of the Y scores around Y' is the average squared difference between the actual Y scores and their corresponding predicted Y' scores.
  - $(S^2_{Y'})$ is one way to describe the average error when using linear regression to predict Y scores.

$$S^2_{Y'} = S^2_Y(1 - r^2)$$

The Standard Error of the Estimate

The standard error of the estimate is similar to the standard deviation of the *Y* scores around their Y' scores and is the clearest way to describe the "average" error when using Y' to predict *Y* scores.

$$S_{Y'} = S_Y \sqrt{1 - r^2}$$

Example 3

- Using the same data set, calculate the standard error of the estimate.

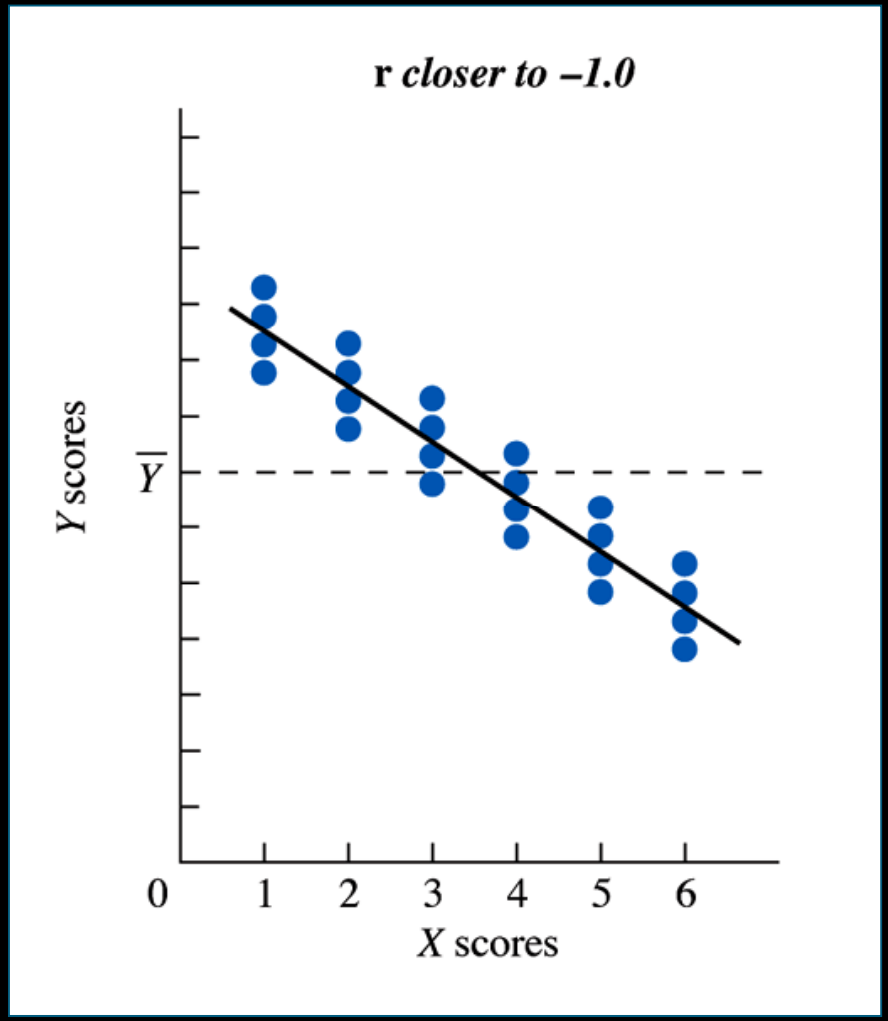
X	*Y*
1	8
2	6
3	6
4	5
5	1
6	3

The Strength of a Relationship and Prediction Error

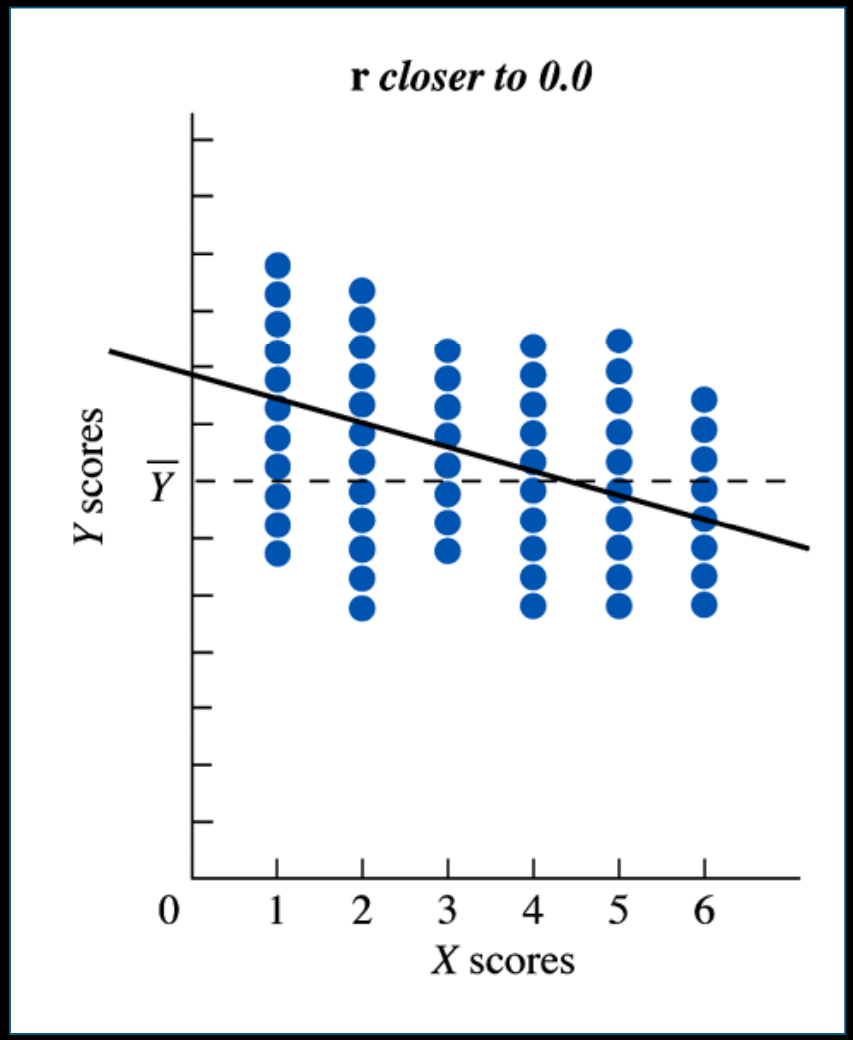
The Strength of a Relationship

As the strength of the relationship—and the absolute value of r—increases, the actual Y scores are closer to their corresponding Y' scores, producing less prediction error and smaller values of $S_{Y'}^2$ and $S_{Y'}$.

Scatterplot of a Strong Relationship



Scatterplot of a Weak Relationship



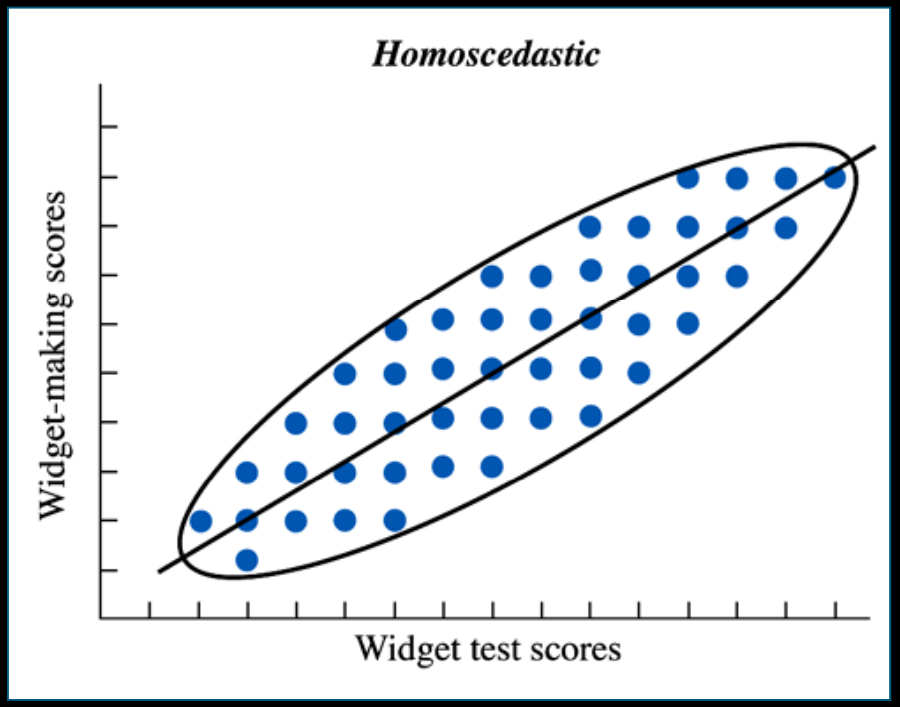
Assumptions of Linear Regression

Assumption 1

- The first assumption of linear regression is that the data are homoscedastic

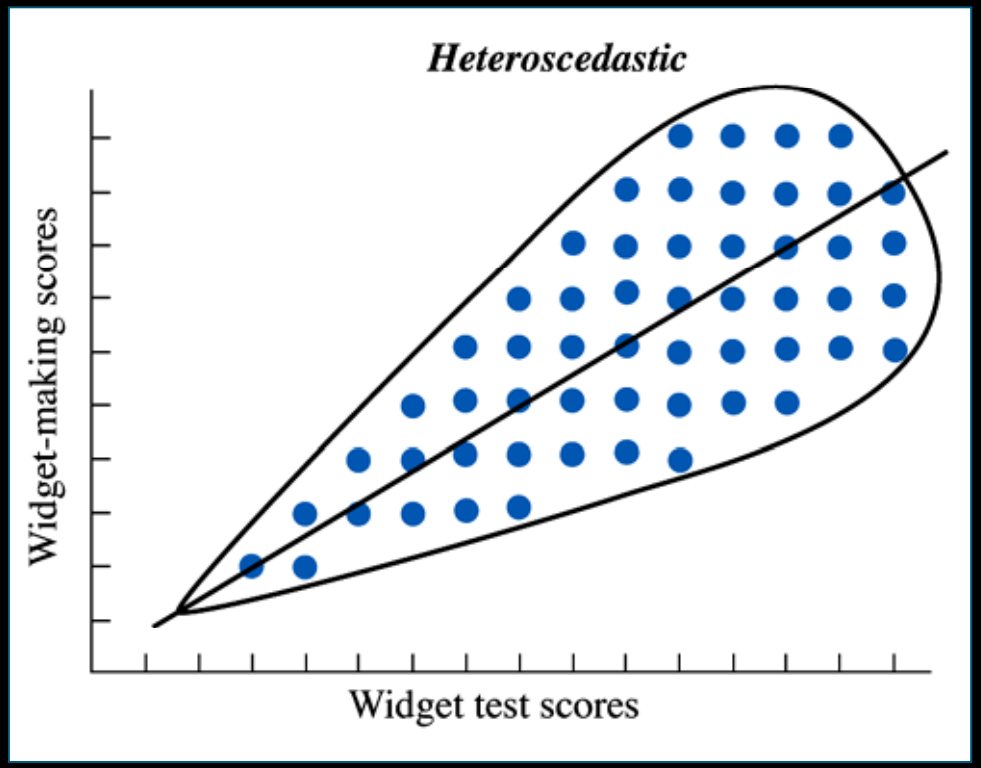
Homoscedasticity

- Homoscedasticity occurs when the *Y* scores are spread out to the same degree at every *X*.



Heteroscedasticity

- Heteroscedasticity occurs when the spread in *Y* is not equal throughout the relationship.



Assumption 2

- The second assumption of linear regression is that the *Y* scores at each *X* form an approximately normal distribution

Scatterplot Showing Normal Distribution of Y Scores at Each

The Proportion of Variance Accounted For

Proportion of Variance Accounted For

The *proportion of variance accounted for* is the proportional improvement in predictions achieved by using a relationship to predict scores, compared to if we do not use the relationship.

Proportion of Variance Accounted For

- When we do not use the relationship, we use the overall mean of the Y scores $(\overline{Y})$ as everyone's predicted Y.
- The error here is the difference between the actual Y scores and the $\overline{Y}$ that we predict they got $(Y - \overline{Y})$.
- When we do not use the relationship to predict scores, our error is S_Y^2.

Proportion of Variance Accounted For

- When we do use the relationship, we use the corresponding Y' as determined by the linear regression equation as our predicted value
- The error here is the difference between the actual Y scores and the Y' that we predict they got $(Y - Y')$
- When we do use the relationship to predict scores, our error is $S^2_{Y'}$

Proportion of Variance Accounted For

The computational formula for the proportion of variance in *Y* that *is* accounted for by a linear relationship with *X* is r^2 . Remember that the formula for computing *r* is

$$r = \frac{N(\Sigma XY) - (\Sigma X)(\Sigma Y)}{\sqrt{[N(\Sigma X^2) - (\Sigma X)^2][N(\Sigma Y^2) - (\Sigma Y)^2]}}$$

Proportion of Variance Not Accounted For

The computational formula for the proportion of variance in *Y* that *is not* accounted for by a linear relationship with *X* is $1 - r^2$

Example 4

- Using the same data set, calculate the proportion of variance accounted for and the proportion of variance not accounted for.

X	*Y*
1	8
2	6
3	6
4	5
5	1
6	3

Example 5

A researcher measures how positive a persons mood is and how creative he or she is, obtaining the interval scores on the table:

Participant	Mood (X)	Creativity (Y)
1	10	7
2	8	6
3	9	11
4	6	4
5	5	5
6	3	7
7	7	4
8	2	5
9	4	6
10	1	4

Example 5

Participant	Mood (X)	Creativity (Y)
1	10	7
2	8	6
3	9	11
4	6	4
5	5	5
6	3	7
7	7	4
8	2	5
9	4	6
10	1	4

A) Compute the statistic that summarizes this relationship

Example 5

Participant	Mood (X)	Creativity (Y)
1	10	7
2	8	6
3	9	11
4	6	4
5	5	5
6	3	7
7	7	4
8	2	5
9	4	6
10	1	4

B) What is the predicted creativity score for anyone scoring 3 on mood?

Example 5

Participant	Mood (X)	Creativity (Y)
1	10	7
2	8	6
3	9	11
4	6	4
5	5	5
6	3	7
7	7	4
8	2	5
9	4	6
10	1	4

C) If your prediction is in error, what is the amount of error you expect to have?


THE
END