BINUS UNIVERSITY

GREATER JAKARTA • BANDUNG • MALANG

STAT6171001 Basic Statistics

Regression

Session 10

Raymond Bahana

rbahana@binus.edu

People
Innovation
Excellence

Session Learning Outcomes

Upon completion of this session, students are expected to be able to

- LO 2. Analyze a problem by using the basic concept of descriptive and inferential statistics
- LO 3. Design a descriptive and inferential statistics solution to meet a given set of computing requirements in the context of computer science
- LO4. Produce descriptive and inferential statistics solutions

Topics

- Regression

Regression

Introduction

- This topic is about prediction.

- Statisticians are often called upon to develop methods to predict one variable from other variables.

- For example, one might want to predict college grade point average from high school grade point average. Or, one might want to predict income from the number of years of education.

Introduction

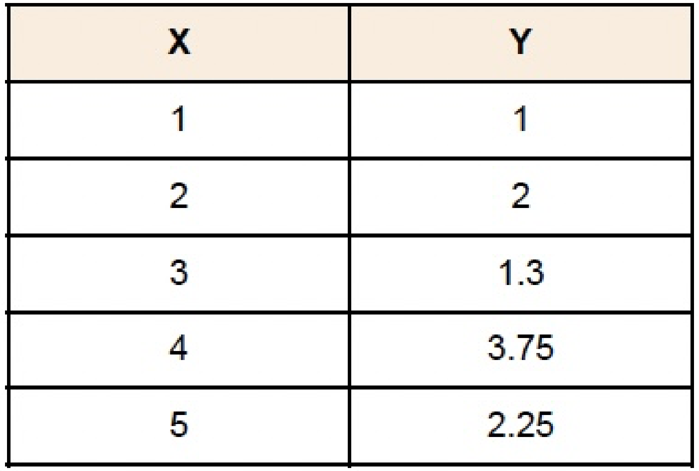
- If two variables are correlated, description can lead to prediction.
- For example, if computer skills and GPAs are related, level of computer skills can be used to predict GPAs.
- Predictive accuracy increases with the strength of the underlying correlation.
- Also discussed is a prevalent phenomenon

Linear Regression

- In simple linear regression, we predict scores on one variable from the scores on a second variable.
- The variable we are predicting is called the criterion variable and is referred to as Y.
- The variable we are basing our predictions on is called the predictor variable and is referred to as X.
- When there is only one predictor variable, the prediction method is called simple regression.
- In simple linear regression, the topic of this section, the predictions of Y when plotted as a function of X form a straight line.

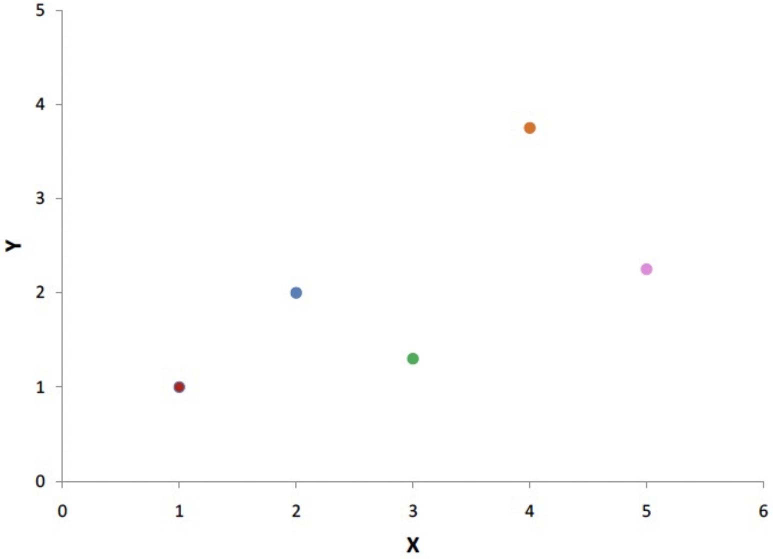
Linear Regression

- The example data in table are plotted in the figure.
- You can see that there is a positive relationship between X and Y. If you were going to predict Y from X, the higher the value of X, the higher your prediction of Y.



X	Y
1	1
2	2
3	1.3
4	3.75
5	2.25

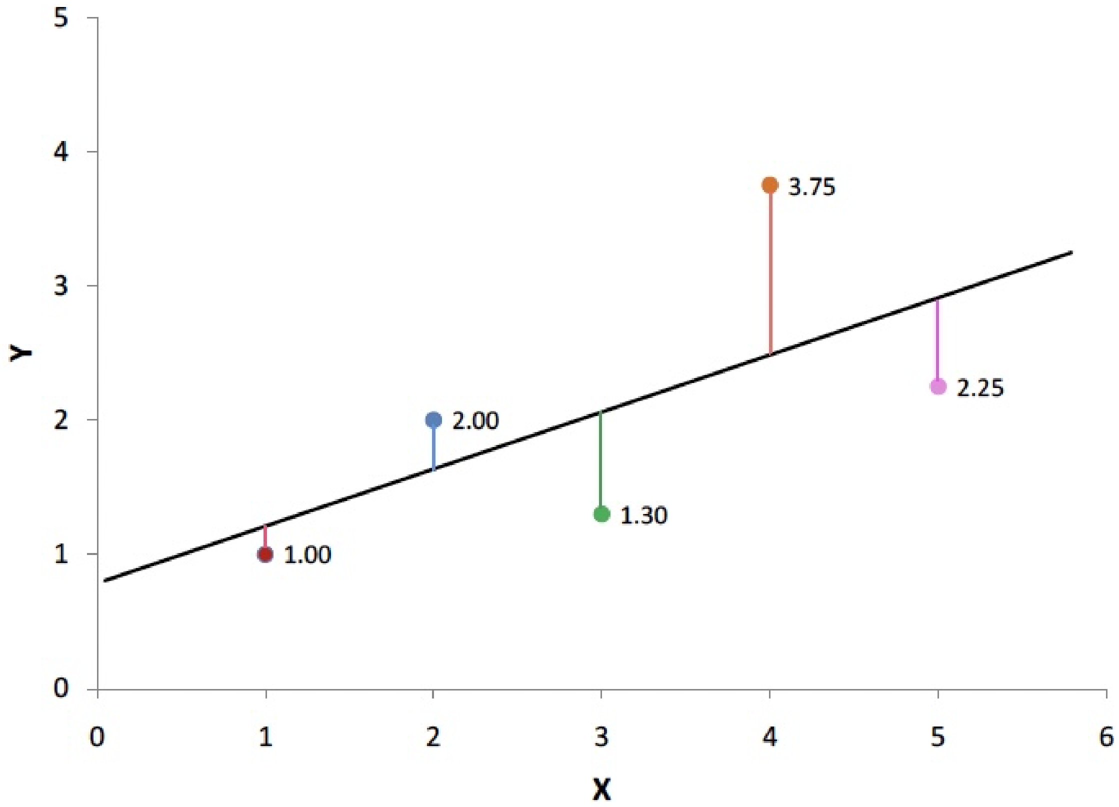
Example data



A scatter plot of the example data

Linear Regression

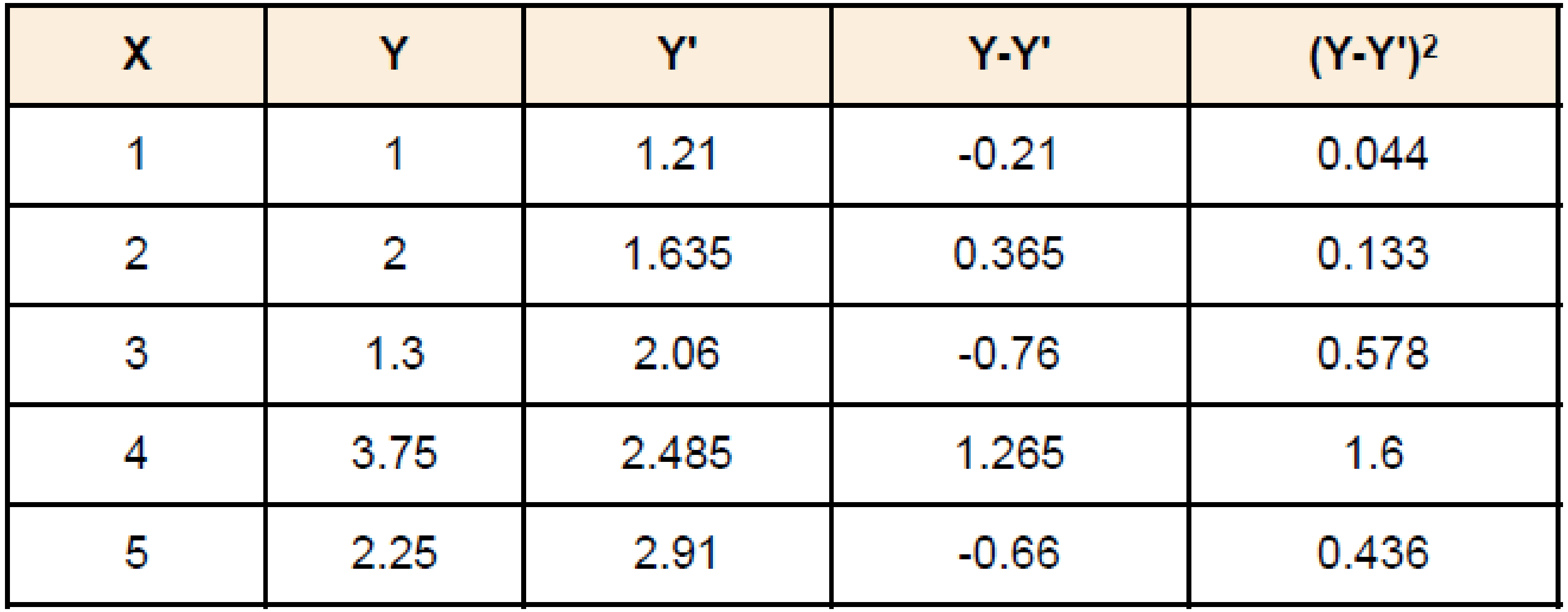
- Linear regression consists of finding the best-fitting straight line through the points.
- The best-fitting line is called a regression line.
- The black diagonal line in the figure is the regression line and consists of the predicted score on Y for each possible value of X.
- The vertical lines from the points to the regression line represent the errors of prediction.
- As you can see, the red point is very near the regression line; its error of prediction is small.
- By contrast, the orange point is much higher than the regression line and therefore its error of prediction is large.



A scatter plot of the example data. The black line consists of the predictions, the points are the actual data, and the vertical lines between the points and the black line represent errors of prediction.

Linear Regression

- The error of prediction for a point is the value of the point minus the predicted value (the value on the line).
- Table below shows the predicted values (Y') and the errors of prediction (Y-Y’).



X	Y	Y'	Y-Y'	$(Y-Y')^2$
1	1	1.21	-0.21	0.044
2	2	1.635	0.365	0.133
3	1.3	2.06	-0.76	0.578
4	3.75	2.485	1.265	1.6
5	2.25	2.91	-0.66	0.436

Linear Regression

- The formula for a regression line is

  $Y' = bX + A$

- where Y' is the predicted score, b is the slope of the line, and A is the Y intercept.
- The equation for the line in previous Figure is

  $Y' = 0.425X + 0.785$

- For X = 1,

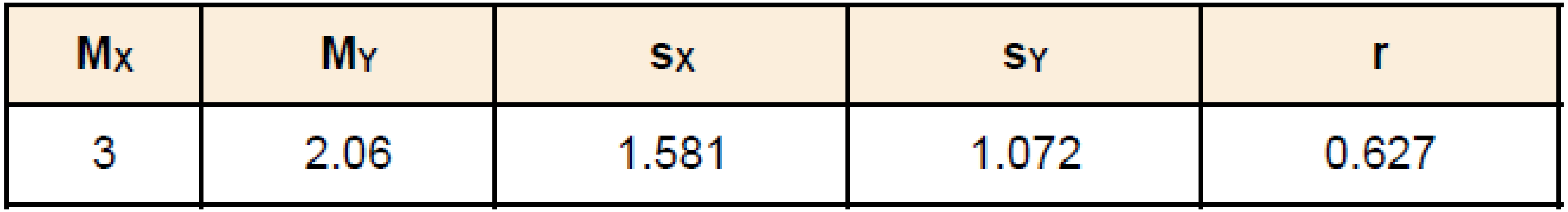
  $Y' = (0.425)(1) + 0.785 = 1.21.$

- For X = 2,

  $Y' = (0.425)(2) + 0.785 = 1.64$

Computing the Regression Line

- The calculations are based on the statistics shown in Table below.
- M_X is the mean of X, M_Y is the mean of Y, s_X is the standard deviation of X, s_Y is the standard deviation of Y, and r is the correlation between X and Y.



M_X	M_Y	s_X	s_Y	r
3	2.06	1.581	1.072	0.627

Computing the Regression Line

- Standard deviation:

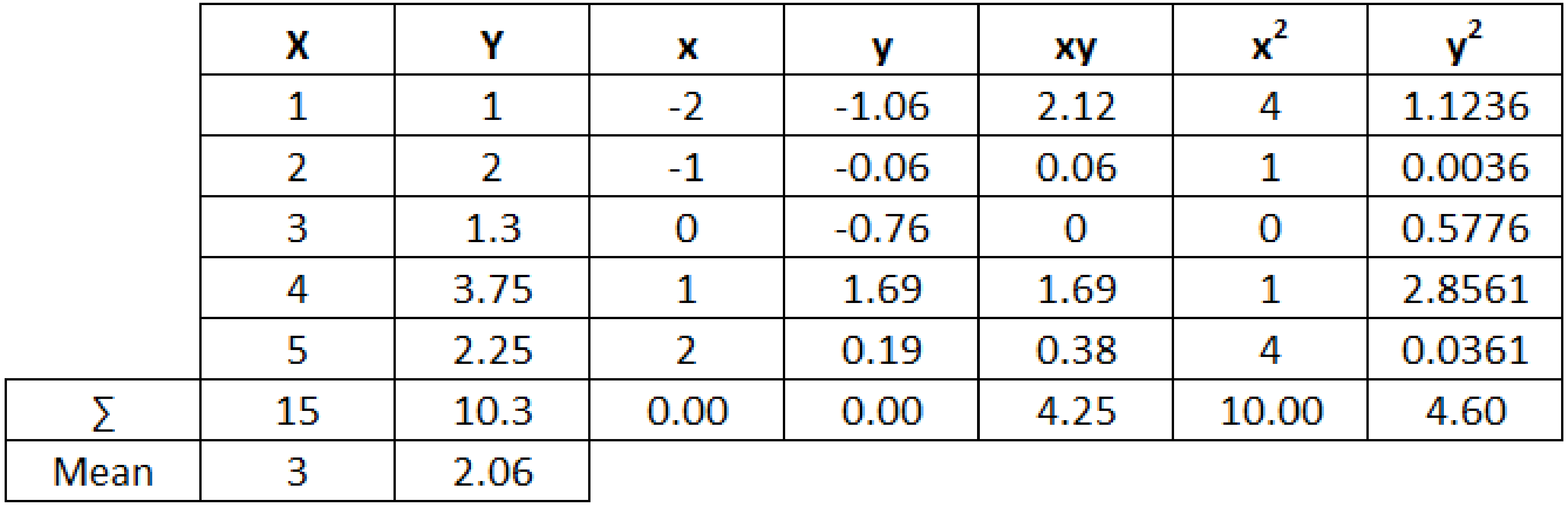
$$s = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \overline{x})^2}{N-1}}$$

- s = sample standard deviation
- N = the number of sample
- x_i = the observed values of a sample item
- $\overline{x}$ = the mean value of the observations

Computing the Regression Line

- r is the correlation between X and Y

$$r = \frac{\sum xy}{\sqrt{\sum x^2 \sum y^2}}$$



	X	Y	x	y	xy	x^2	y^2
	1	1	-2	-1.06	2.12	4	1.1236
	2	2	-1	-0.06	0.06	1	0.0036
	3	1.3	0	-0.76	0	0	0.5776
	4	3.75	1	1.69	1.69	1	2.8561
	5	2.25	2	0.19	0.38	4	0.0361
∑	15	10.3	0.00	0.00	4.25	10.00	4.60
Mean	3	2.06					

Computing the Regression Line

- The slope (b) can be calculated as follows:

$$b = r\frac{s_y}{s_x}$$

$$b = (0.627)\frac{1.072}{1.581} = 0.425$$

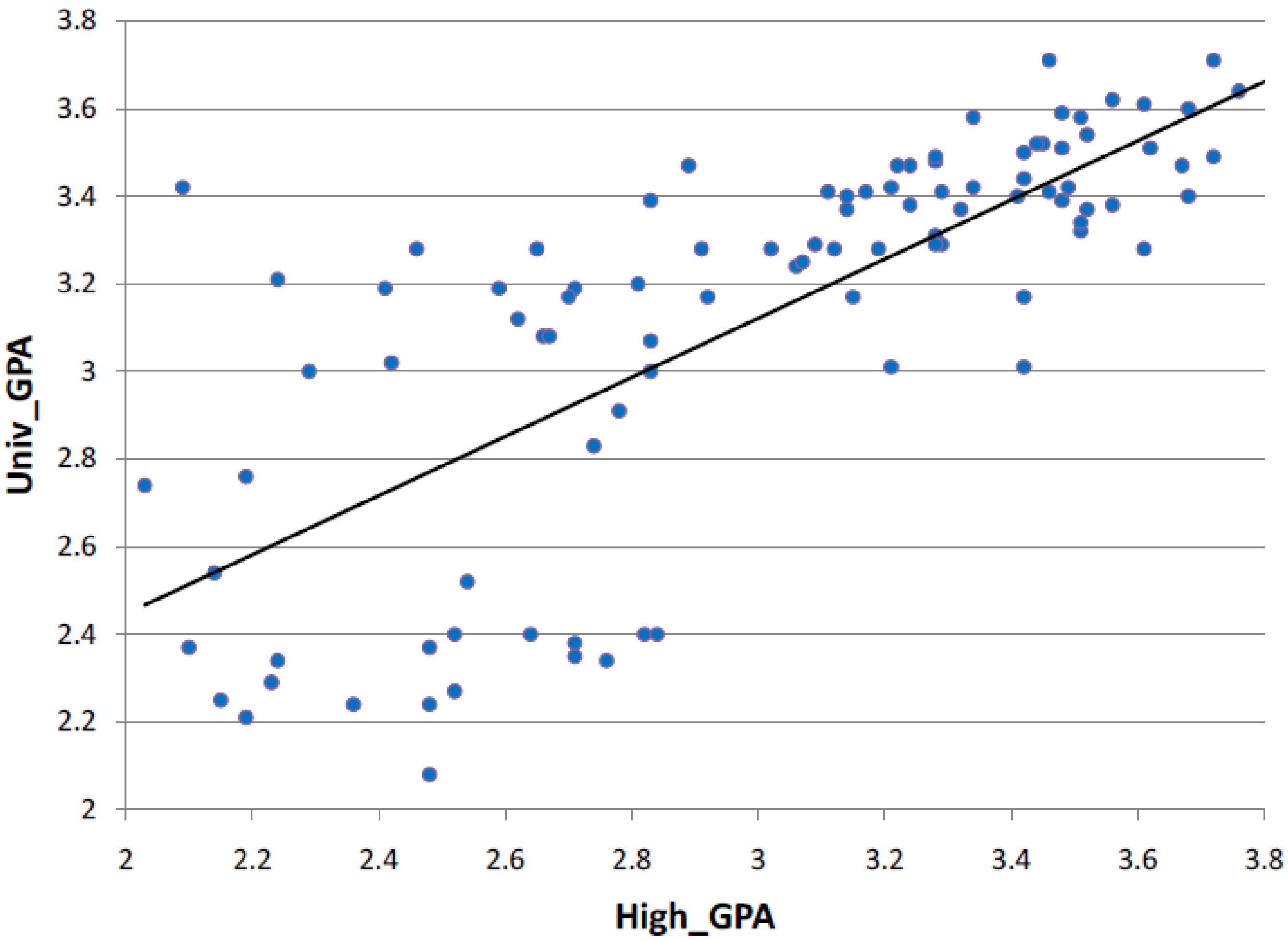
- and the intercept (A) can be calculated as

  A = M_Y - b M_X

  A = 2.06 - (0.425)(3) = 0.785

Real Example

- The figure shows a scatter plot of University GPA as a function of High School GPA.
- You can see from the figure that there is a strong positive relationship



Partitioning the Sums of Squares

Introduction

- One aspect of regression: it can divide the variation in Y into 2 parts: the variation of the predicted scores and the variation in the errors of prediction.
- The variation of Y is called the **sum of squares Y**.
- In the population, the formula is

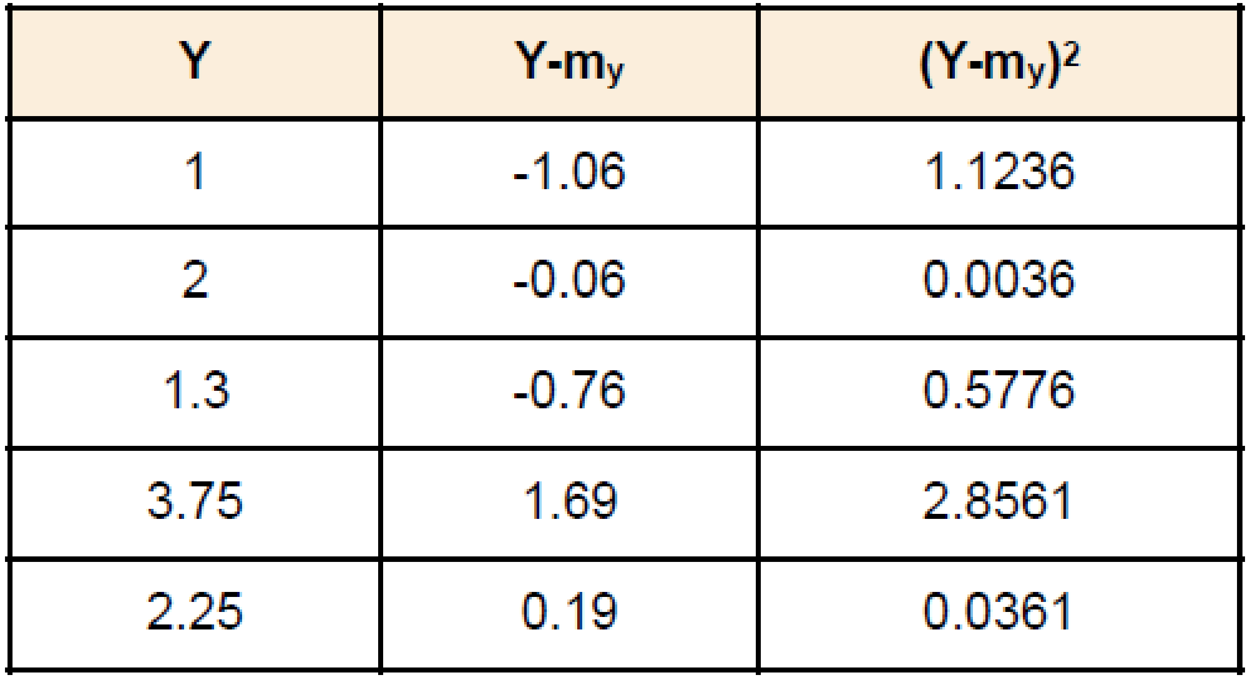
$$SSY = \sum (Y - \mu_Y)^2$$

- When computed in a sample, use the sample mean, M:

$$SSY = \sum (Y - M_Y)^2$$

Example

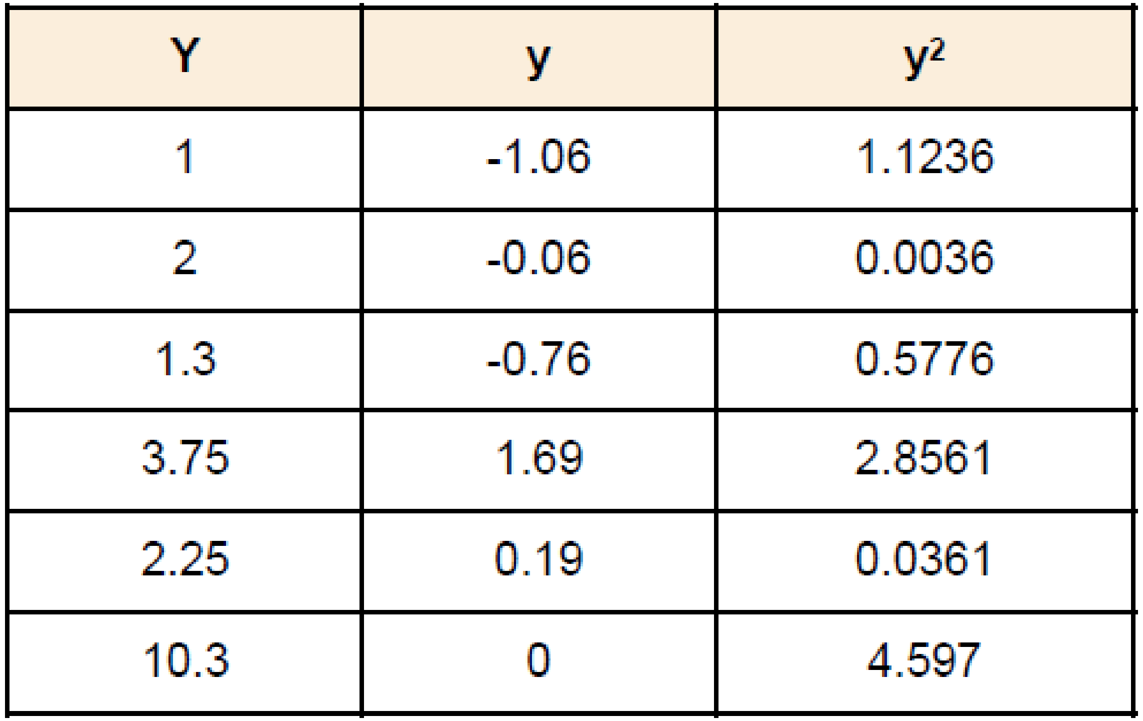
- SSY is the sum of squares Y, Y is an individual value of Y, and M_Y is the mean of Y.
- The mean of Y is 2.06 and SSY is the sum of the values in the third column and is equal to 4.597.



Y	$Y\text{-}m_y$	$(Y\text{-}m_y)^2$
1	-1.06	1.1236
2	-0.06	0.0036
1.3	-0.76	0.5776
3.75	1.69	2.8561
2.25	0.19	0.0361

Deviation Scores

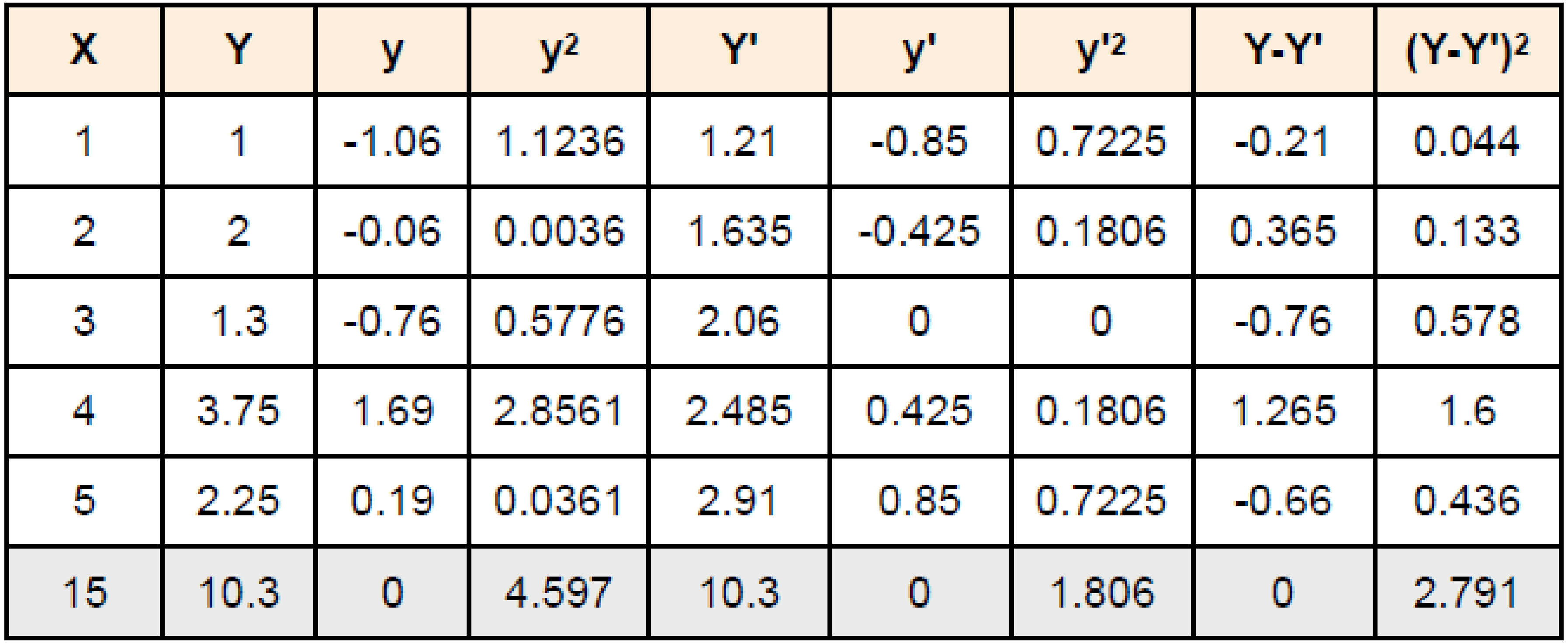
- It is sometimes convenient to use formulas that use deviation scores rather than raw scores.
- Deviation scores are simply deviations from the mean.
- By convention, small letters rather than capitals are used for deviation scores.
- Therefore, the score, y indicates the difference between Y and the mean of Y.



Y	y	y^2
1	-1.06	1.1236
2	-0.06	0.0036
1.3	-0.76	0.5776
3.75	1.69	2.8561
2.25	0.19	0.0361
10.3	0	4.597

How the SSY is partitioned

- The data in this Table are reproduced from the introductory section.
- The column X has the values of the predictor variable, and the column Y has the criterion variable.



X	Y	y	y^2	Y'	y'	y'^2	Y-Y'	$(Y-Y')^2$
1	1	-1.06	1.1236	1.21	-0.85	0.7225	-0.21	0.044
2	2	-0.06	0.0036	1.635	-0.425	0.1806	0.365	0.133
3	1.3	-0.76	0.5776	2.06	0	0	-0.76	0.578
4	3.75	1.69	2.8561	2.485	0.425	0.1806	1.265	1.6
5	2.25	0.19	0.0361	2.91	0.85	0.7225	-0.66	0.436
15	10.3	0	4.597	10.3	0	1.806	0	2.791

How the SSY is partitioned

- The 3rd column, y, the differences between the column Y and the mean of Y.
- The 4th column, y^2, is simply the square of the y column.
- The column Y’ contains the predicted values of Y.

  Y' = 0.425X + 0.785
- The column y' contains deviations of Y' from the mean of Y' and y'^2 is the square of this column.
- The next to last column, Y-Y', contains the actual scores (Y) minus the predicted scores (Y').
- The last column contains the squares of these errors of prediction.

How the SSY is partitioned

- Recall that SSY is the sum of the squared deviations from the mean, therefore the sum of the y^2 column (4.597).
- SSY can be partitioned into two parts: the sum of squares predicted (SSY') and the sum of squares error (SSE).
- SSY' is the sum of the squared deviations of the predicted scores from the mean predicted score → the sum of y'^2 column (1.806).
- SSE is the sum of the squared errors of prediction, therefore, the sum of the $(Y-Y')^2$ column (2.791).
- This can be summed up as:

$$SSY = SSY' + SSE$$

$$4.597 = 1.806 + 2.791$$

How the SSY is partitioned

- The SSY is the total variation, the SSY' is the variation explained, and the SSE is the variation unexplained.

- Therefore, the proportion of variation explained can be computed as:

$$\text{Proportion explained} = \frac{\text{SSY'}}{\text{SSY}}$$

- Similarly, the proportion not explained is:

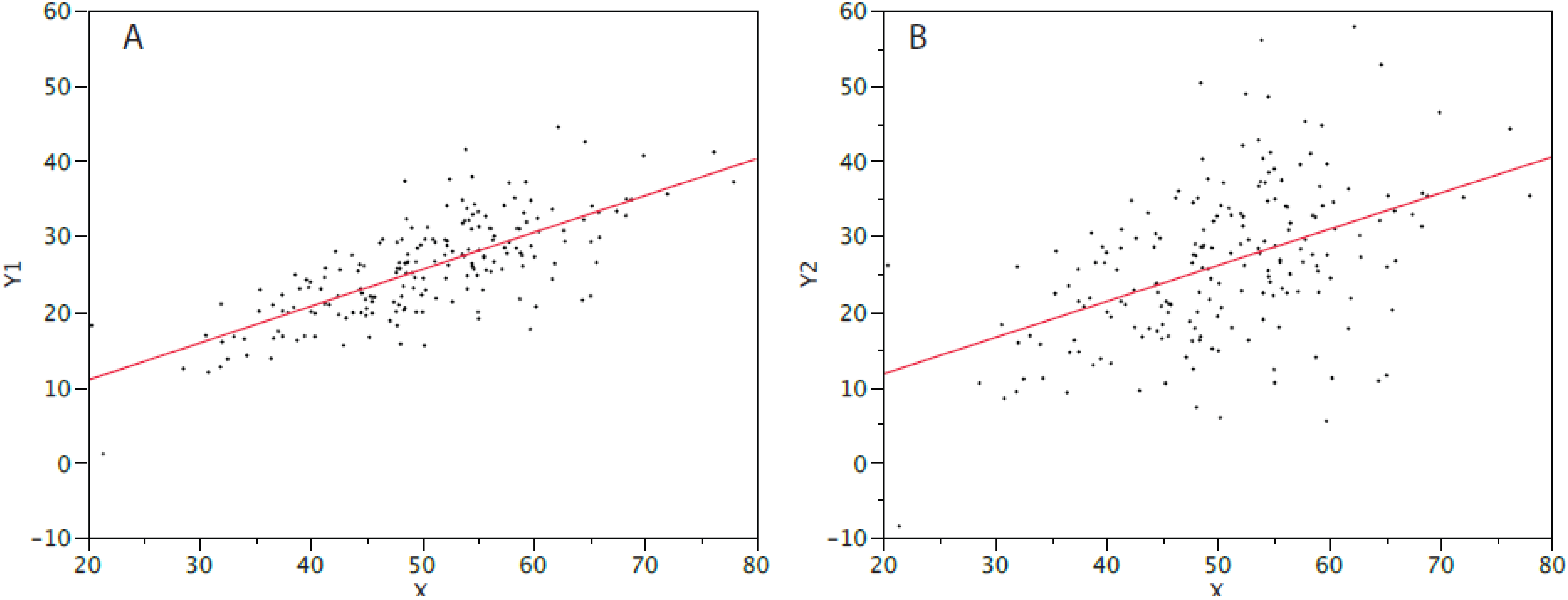
$$\text{Proportion not explained} = \frac{\text{SSE}}{\text{SSY}}$$

Pearson’s Correlation

- There is an important relationship between the proportion of variation explained and Pearson's correlation: **r^2 is the proportion of variation explained**.
- Therefore, if r = 1, then, naturally, the proportion of variation explained is 1; if r = 0, then the proportion explained is 0.
- Other example: for r = 0.4, the proportion of variation explained is 0.16

Standard Error of the Estimate

Introduction



Introduction

- In Graph A, the points are closer to the line than they are in Graph B.
- Therefore, the predictions in Graph A are more accurate than in Graph B.
- The standard error of the estimate is a measure of the accuracy of predictions.
- Recall that the regression line is the line that minimizes the sum of squared deviations of prediction (also called the sum of squares error).

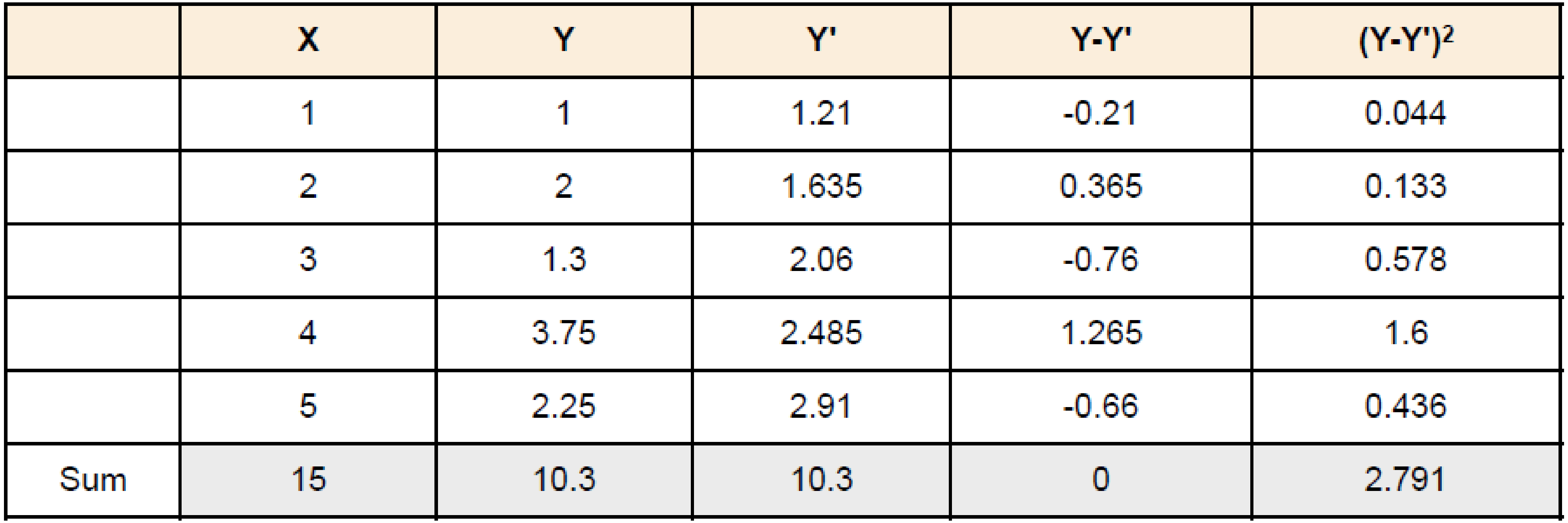
The Standard Error

- The standard error of the estimate is closely related to this quantity and is defined below:

$$\sigma_{est} = \sqrt{\frac{\sum (Y - Y')^2}{N}}$$

- where σ_{est} is the standard error of the estimate, Y is an actual score, Y' is a predicted score, and N is the number of pairs of scores.

Example



	X	Y	Y'	Y-Y'	(Y-Y')^2
	1	1	1.21	-0.21	0.044
	2	2	1.635	0.365	0.133
	3	1.3	2.06	-0.76	0.578
	4	3.75	2.485	1.265	1.6
	5	2.25	2.91	-0.66	0.436
Sum	15	10.3	10.3	0	2.791

- The standard error of the estimate is

$$\sigma_{est} = \sqrt{\frac{2.791}{5}} = 0.747$$

Pearson’s Correlation

- There is a version of the formula for the standard error in terms of Pearson’s correlation:

$$\sigma_{est} = \sqrt{\frac{(1-\rho^2)SSY}{N}}$$

- where ρ is the population value of Pearson's correlation and SSY is

$$SSY = \sum (Y - \mu_Y)^2$$

Pearson’s Correlation

- For the data in Table 1, my = 10.30, SSY = 4.597 and r = 0.6268. Therefore,

$$\sigma_{est} = \sqrt{\frac{(1-0.6268^2)(4.597)}{5}} = \sqrt{\frac{2.791}{5}} = 0.747$$

The Standard Error - Sample

- Similar formulas are used when the standard error of the estimate is computed from a sample rather than a population.
- The only difference is that the denominator is N-2 rather than N.
- The reason N-2 is used rather than N-1 is that two parameters (the slope and the intercept) were estimated in order to estimate the sum of squares.

$$s_{est} = \sqrt{\frac{\sum(Y - Y')^2}{N - 2}}$$

$$s_{est} = \sqrt{\frac{2.791}{3}} = 0.964$$

$$s_{est} = \sqrt{\frac{(1 - r^2)SSY}{N - 2}}$$

Inferential Statistics for b and r

Introduction

- This section shows how to conduct significance tests and compute confidence intervals for the regression slope and Pearson's correlation.
- If the regression slope is significantly different from zero, then the correlation coefficient is also significantly different from zero.

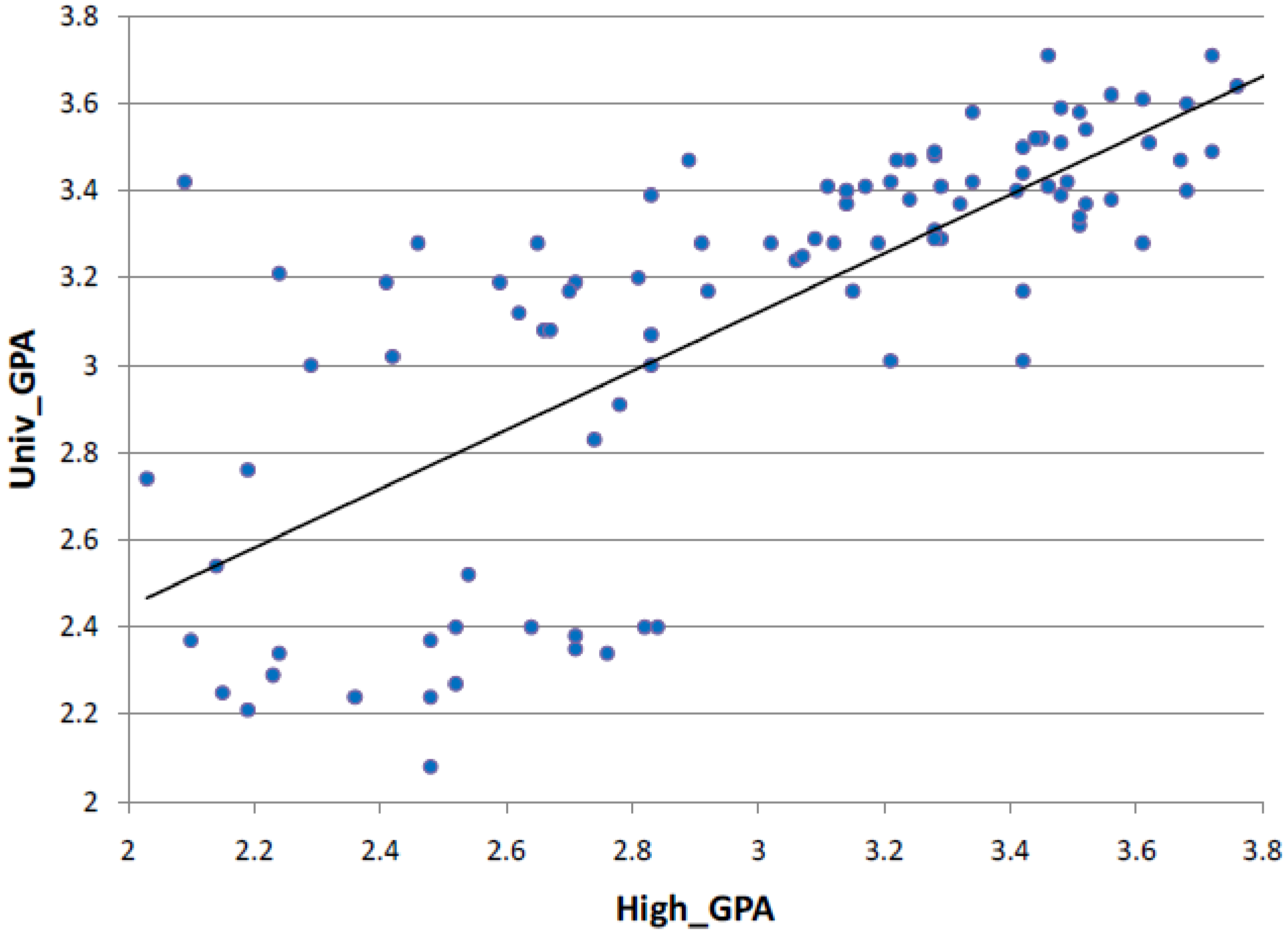
Assumptions

- No assumptions were needed to determine the best-fitting straight line, assumptions are made in the calculation of inferential statistics.
- Naturally, these assumptions refer to the population, not the sample.

1. Linearity: The relationship between the two variables is linear.
2. Homoscedasticity: The variance around the regression line is the same for all values of X. A clear violation of this assumption is shown in next Figure.
3. The errors of prediction are distributed normally. This means that the distributions of deviations from the regression line are normally distributed.

Example

- The predictions for students with high high-school GPAs are very good, whereas the predictions for students with low high-school GPAs aren't very good.
- The points for students with high high-school GPAs are close to the regression line, whereas the points for low high-school GPA students aren't.



Significance Test for the Slope (b)

- Recall the general formula for a t test:

$$t = \frac{statistics - hypothesized\ value}{estimated\ standard\ error\ of\ the\ statistic}$$

- The degrees of freedom for this test are:

df = N-2

- where N is the number of pairs of scores.

Significance Test for the Slope (b)

- The estimated standard error of b is computed using the following formula:

$$s_b = \frac{s_{est}}{\sqrt{SSX}}$$

- where s_b is the estimated standard error of b, s_{est} is the standard error of the estimate, and SSX is the sum of squared deviations of X from the mean of X.

Significance Test for the Slope (b)

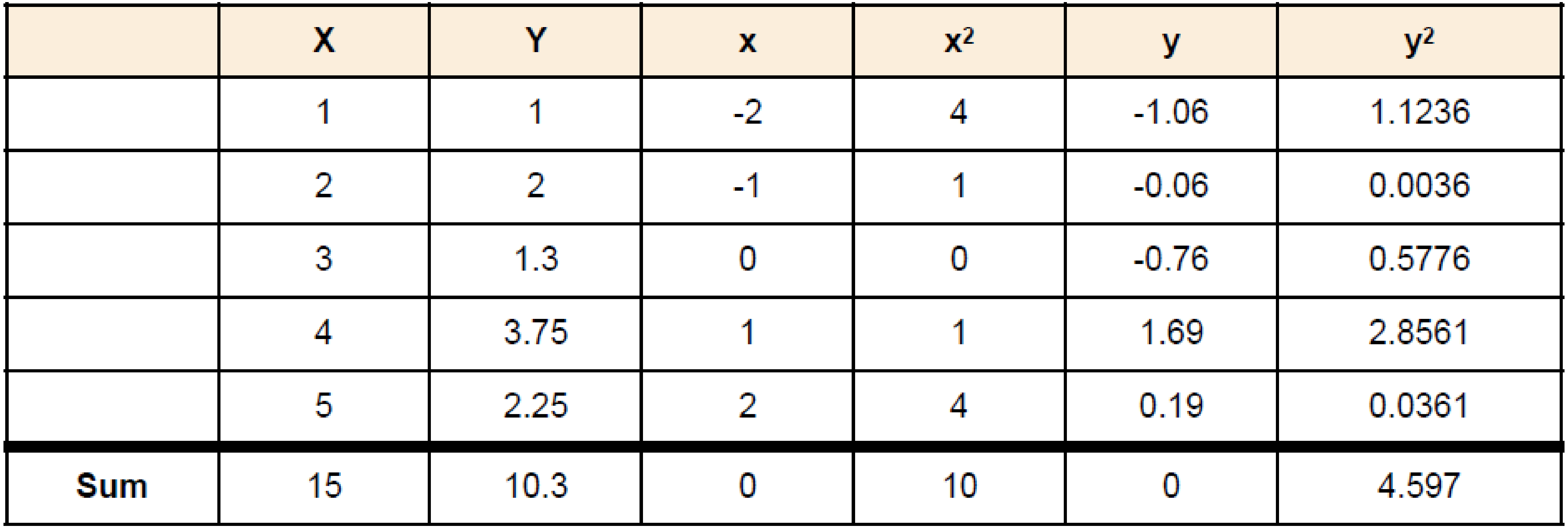
- SSX is calculated as

$$SSX = \sum (X - M_x)^2$$

- where Mx is the mean of X.
- The standard error of the estimate can be calculated as

$$s_{est} = \sqrt{\frac{(1 - r^2)SSY}{N - 2}}$$

Example



	X	Y	x	x^2	y	y^2
	1	1	-2	4	-1.06	1.1236
	2	2	-1	1	-0.06	0.0036
	3	1.3	0	0	-0.76	0.5776
	4	3.75	1	1	1.69	2.8561
	5	2.25	2	4	0.19	0.0361
Sum	15	10.3	0	10	0	4.597

Example

- The column X has the values of the predictor variable, and the column Y has the values of the criterion variable.
- The third column, x, contains the differences between the values of column X and the mean of X.
- The fourth column, x^2, is the square of the x column.
- The fifth column, y, contains the differences between the values of column Y and the mean of Y.
- The last column, y^2, is simply the square of the y column.

Example

- The computation of the standard error of the estimate (s_{est}) for these data is shown in the section on the standard error of the estimate.

$$s_{est} = 0.964$$

- SSX is the sum of squared deviations from the mean of X.
- It is, therefore, equal to the sum of the x^2 column.

$$SSX = 10.00$$

- The standard error of b:

$$s_b = \frac{0.964}{\sqrt{10}} = 0.305$$

Example

- As shown previously, the slope (b) is 0.425.
- Therefore:

$$t = \frac{0.425}{0.305} = 1.39$$

df = N-2 = 5-2 = 3.

- The p value for a two-tailed t test is 0.26 → the slope is not significantly different from 0

References

- Witte, R.S.&Witte, J.S. (2017). Statistics (11th ed.). Wiley. ISBN: 978-1119386056.
- Lane, D.M., Scott, D., Hebl, M., Guerra, R., Osherson, D.& Zimmer, H. (2003). Introduction to Statistics. Online edition at https://open.umn.edu/opentextbooks/textbooks/459
- Levine, D.M., Stephan, D.F. & Szabat, K.A. (2017). Statistics for Managers Using Microsoft Excel (8th ed.). Pearson. ISBN: 978-0134566672

Thank you