SHS

AIRs - LM in

Statistics and Probability
Quarter 4: Week 8- Module 16
Regression Analysis

GOVERNMENT PROPERTY
NOT FOR SALE

Statistics and Probability
Grade 11 Quarter 4: Week 8 - Module 16: Regression Analysis
First Edition, 2021

Development Team of the Module

Author: Sherlyn A. De la Peña

Editor: SDO La Union, Learning Resource Quality Assurance Team

Illustrator: Ernesto F. Ramos Jr., *P II*

Management Team:

Atty. Donato D. Balderas, Jr.
Schools Division Superintendent

Vivian Luz S. Pagatpatan, PHD
Assistant Schools Division Superintendent

German E. Flora, PHD, *CID Chief*

Virgilio C. Boado, PHD, *EPS in Charge of LRMS*

Erlinda M. Dela Peña, EDD, *EPS in Charge of Mathematics*

Michael Jason D. Morales, *PDO II*

Claire P. Toluyen, *Librarian II*

Target

In this lesson, we will take a deeper look at the trend line. We will go to its more accurate analysis by getting the mathematical equation and how it is used in prediction.

.

After going through this lesson, you are expected to:

1. Identifies the independent and dependent variables. **M11/12SP-IVi-1**
2. Calculates the slope and y-intercept of the regression line. **M11/12SP-IVi-3**
3. Interprets the calculated slope and y-intercept of the regression line. **M11/12SP-IVi-4**
4. Predicts the value of the dependent variable given the value of the independent variable. **M11/12SP-IVj-1**
5. Solves problems involving regression analysis. **M11/12SP-IVj-2**

Subtasks:

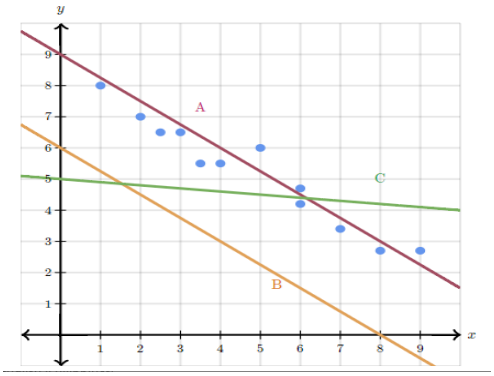
1. Find the regression line

Before going on, check how much you know about this topic. Answer the pretest on the next page in a separate sheet of paper.

Pre-test

Directions: Choose the letter of the correct answer. Write your answer on a separate sheet of paper.

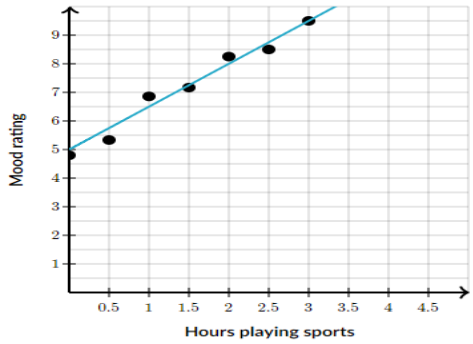
1. Which line fits the data graphed below?



A. Line A
B. Line B
C. Line C
D. None of the lines fit the data.

2. Schon distributed a survey to his fellow students asking them how many hours they'd spent playing sports in the past day. He also asked them to rate their mood on a scale from 0 to 10, with 10 being the happiest. A line was fit

to the data to model the relationship. Which of these linear equations best describes the given model?



A. Y = 5X + 1.5
B. Y = 1.5X + 5
C. Y = -1.5X+ 5
D. Y = -5X + 1.5

3. Refer to the model in number 2, estimate the mood rating for a student that spent 2.5 hours playing sports.
   A. 10 B. 9.5 C. 8.75 D. 1.25

4. If two variables, x and y, have a very strong linear relationship, then
   A. none of these alternatives is correct
   B. there is evidence that x causes a change in y
   C. there is evidence that y causes a change in x
   D. there might not be any causal relationship between x and y

5. Regression analysis was applied between sales (y) and advertising (x) across all the branches of a major national corporation. The following regression function was obtained, y = 5000 + 7.25x. If the advertising budgets of two branches of the corporation differ by 30,000, then what will be the predicted difference in their sales?
   A. 217,500 B. 222,500 C. 5000 D. 7.25

6. Which of the following problems could be studied through the use of regression analysis?
   A. Which runner has been on a track team the longest?
   B. Which runner on a track team has the fastest average speed?
   C. Which runner on a track team has the fastest recorded speed?
   D. How the speeds of runners on a track team relate to how long they each train?

7. In regression analysis the variable that is being predicted is the
   A. Dependent variable B. Independent variable
   C. Intervening variable D. Is usually X

8. In regression analysis, which of the following variable that is used to explain the change in the outcome of an experiment, or some natural process?
   A. The x-variable B. The independent variable
   C. The predictor variable D. The explanatory variable

9. Regression modeling is a statistical framework for developing a mathematical equation, what is being describe in the model?

A. several explanatory and several response variables response are related
B. one explanatory and one or more response variables are related
C. one response and one or more explanatory variables are related
D. All of these are correct.

10. You studied the impact of the dose of a new drug treatment for high blood pressure. You think that the drug might be more effective in people with very high blood pressure. Because you expect a bigger change in those patients who start the treatment with high blood pressure, you use regression to analyze the relationship between the initial blood pressure of a patient (x) and the change in blood pressure after treatment with the new drug (y). If you find a very strong positive association between these variables, what would be your conclusion?
    A. There is evidence that the higher the patients initial blood pressure, the bigger the impact of the new drug.
    B. There is evidence that the higher the patients initial blood pressure, the smaller the impact of the new drug.
    C. There is evidence for an association of some kind between the patient's initial blood pressure and the impact of the new drug on the patient's blood pressure
    D. None of these are correct, this is a case of regression fallacy.

11. The relationship between number of beers consumed (x) and blood alcohol content (y) was studied in 16 male college students by using least squares regression. The following regression equation was obtained from this study: Y' = -0.0127 + 0.0180x. What does the equation imply?
    A. Each beer consumed increases blood alcohol by 1.27%.
    B. Each beer consumed increases blood alcohol by exactly 0.018.
    C. On the average it takes 1.8 beers to increase blood alcohol content by 1%.
    D. Each beer consumed increases blood alcohol by an average amount of 1.8%

12. In regression analysis, if the independent variable is measured in kilograms, what would be the unit of the dependent variable?
    A. Must be also in kilograms
    B. Must be in some unit of weight
    C. Cannot be in kilograms
    D. Can be any units

13. In the case of an algebraic model for a straight line, if a value for the x variable is specified, what would happen to the value for y?
    A. The computed response to the independent value will always give a minimal residual variable.
    B. The computed value of y will always be the best estimate of the mean response.
    C. The exact value of the response variable can be computed.
    D. None of these alternatives is correct.

14. A regression analysis between sales (in P1000) and price (in peso) resulted in the following equation: Y' = 50,000 - 8X, what is the implication of the given equation?
    A. increase of P 1 in price is associated with a decrease of P 8 in sales
    B. increase of P 1 in price is associated with a decrease of P 8000 in sales
    C. increase of P 8 in price is associated with an increase of P 8,000 in sales
    D. increase of P 1 in price is associated with a decrease of P 42,000 in sales

15. Regression analysis was applied to return rates of sparrow hawk colonies. Regression analysis was used to study the relationship between return rate (x: % of birds that return to the colony in a given year) and immigration rate (y: % of new adults that join the colony per year). The following regression equation was obtained. Y' = 31.9 – 0.34x. Based on the above estimated regression equation, if the return rate were to decrease by 10%, what would be the immigration rate of the colony?

A. Increase by 34%
B. Increase by 3.4%
C. Decrease by 0.34%
D. Decrease by 3.4%

Jumpstart

For you to understand the lesson well, do the following activities. Have fun and good luck!

Activity 1: Find my Equation!

Given the following ordered pair in tabular values, find the slope and y-intercept and determine the equation of the line.

X	0	1	2	3	4	5
Y	4	6	8	10	12	14

Activity 2. Dependent or Independent?

Directions. Identify the dependent and independent variables in each bivariate data. Place your answer in the succeeding table.

1. Altitude and acceleration due to gravity
2. Price of goods and the demand
3. Monthly salary and annual income of a worker
4. IQ and academic performance of a student
5. Temperature and volume of air in a balloon
6.

	Dependent	Independent
1		
2		
3		
4		
5		

A. Place each variable on the blank below

1. _____depends upon _____
   (dependent variable) (independent variable)
2.
3.
4.
5.

B. Using the letter X and Y, which one is normally assigned as Y? Assigned as X?

Discover

In a scatterplot, we can draw the trend line if there is an evident correlation between the bivariate data.

When the trend line is drawn, we observe that some points are on the line while others are below or above the line. In other words, we say that the points in the scatterplot regress with reference to the line. If the average y distances of the points from this line is the least, then we call this line the ***regression line*** or the line that "best fit" in the scatterplot. The regression line is the same as the trend line.

To find the regression line, like the equation of a line in Algebra, we write the equation of the regression line using the "point-slope-form".

The Regression Line (The Line of Best Fit)

$$Y' = bX + a$$

where:

a = y-intercept

Formula for the y-intercept $(a) = \frac{(\Sigma Y)(\Sigma X^2)-(\Sigma X)(\Sigma XY)}{n(\Sigma X^2)-(\Sigma X)^2}$

b = slope of the regression line

Formula for the slope of the regression line $(b) = \frac{n(\Sigma XY)-(\Sigma X)(\Sigma Y)}{n(\Sigma X^2)-(\Sigma X)^2}$

n = number of cases

The regression line Y' = bX + a is also called the line prediction equation because we use to predict Y if X is known. Since in the analysis, only the Y distance was considered, the line cannot be used to predict X from Y.

To determine the regression line or do a regression analysis, we go through the following steps:

1. Identify the dependent and independent variable.
2. Find the value of the correlation coefficient *(r*).
3. Test the significance of *r*. If *r* is significant proceed to regression analysis. If r is not significant proceed to Step 4. (regression analysis cannot be done-STOP)
4. Find the value of *a* and *b*.
5. Plug in the value of *a* and *b* in the regression line Y' = bX + a.

Example:

The following data pertains to the height of fathers and to their eldest sons in inches. If there is a significant relationship between the two variables, predict the height of the son if the height of his father is 78 inches.

Height of the Father	Height of the son
71	71
69	69
69	71
65	68
66	68
63	66
68	70
70	72
60	65
58	60

STEPS	SOLUTION
1. Identify the dependent and independent variables	Dependent variable (Y) – height of the son Independent variable (X) - height of the father
2. Compute the correlation coefficient using the formula $r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[n\Sigma x^2 - (\Sigma x)^2][n\Sigma y^2 - (\Sigma y)^2]}}$	<table><tr><th>X</th><th>Y</th><th>X^2</th><th>Y^2</th><th>XY</th></tr><tr><td>71</td><td>71</td><td>5041</td><td>5041</td><td>5041</td></tr><tr><td>69</td><td>69</td><td>4761</td><td>4761</td><td>4761</td></tr><tr><td>69</td><td>71</td><td>4761</td><td>5041</td><td>4899</td></tr><tr><td>65</td><td>68</td><td>4225</td><td>4624</td><td>4420</td></tr><tr><td>66</td><td>68</td><td>4356</td><td>4624</td><td>4488</td></tr><tr><td>63</td><td>66</td><td>3969</td><td>4356</td><td>4158</td></tr><tr><td>68</td><td>70</td><td>4624</td><td>4900</td><td>4760</td></tr><tr><td>70</td><td>72</td><td>4900</td><td>5184</td><td>5040</td></tr><tr><td>60</td><td>65</td><td>3600</td><td>4225</td><td>3900</td></tr><tr><td>58</td><td>60</td><td>3364</td><td>3600</td><td>3480</td></tr><tr><td>ΣX= 659</td><td>ΣY= 680</td><td>ΣX^2= 43601</td><td>ΣY^2= 46356</td><td>ΣXY= 44947</td></tr></table> $r = \frac{10(44947)-(659)(680)}{\sqrt{[10(43601)-(659)^2][10(46356)-(680)^2]}}$ r = 0.95
3. Test the significance using the formula $t = r\sqrt{\frac{n-2}{1-r^2}}$	n =10 and r = 0.95 $t = 0.63\sqrt{\frac{10-2}{1-(0.95)^2}}$ t = 8.61
4. Compare the computed t-value to the critical t-value	Using *df*= n-2 =10 - 2 = 7, level of significance is 0.05 two-tailed test, we find from the table that the critical value of t is 2.306.
5. Make a decision	Since the computed t = 8.61 is greater than the critical t = 2.306, we reject the null hypothesis. So, there is significant relationship between the two variables.

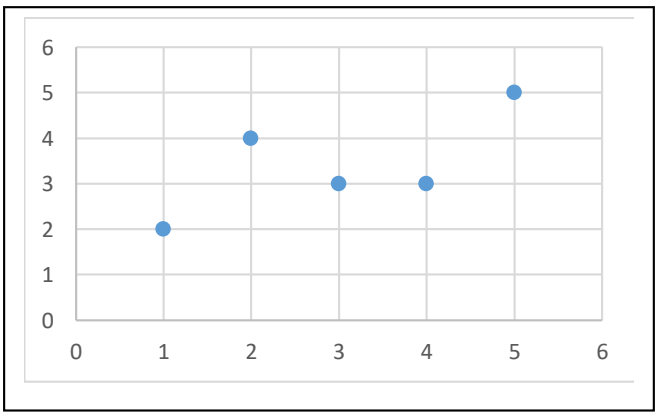
6. Summarize	There is a sufficient evidence to conclude that there is a significant relationship the number of height of the father and height of the son. Thus, we will proceed to regression analysis.
7. Compute the value of a and b in the regression equation using the formula $a = \frac{(\Sigma Y)(\Sigma X^2)-(\Sigma X)(\Sigma XY)}{n(\Sigma X^2)-(\Sigma X)^2}$ $b = \frac{n(\Sigma XY)-(\Sigma X)(\Sigma Y)}{n(\Sigma X^2)-(\Sigma X)^2}$	$a = \frac{(680)(43601)-(659)(44947)}{10(43601)-(659)^2}$ $a = 16.55$ $b = \frac{10(44947)-(659)(680)}{10(43601)-(659)^2}$ $b = 0.78$
8. Form the regression equation	Plug in the value of a and b in equation $Y' = bX + a$ $Y' = 0.78X + 16.55$ The regression equation for predicting the height of the son given the height of the father is $Y' = 0.78X + 16.55$
9. Predict the height of the son if the height of the father is 78 inches.	$X = 78$ $Y' = 0.78X + 16.55$ $Y' = 0.78(78) + 16.55$ $= 77.39$ or 77 inches So, the predicted height of the son whose father's height is 78 inches is 77 inches.

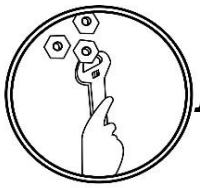
Explore

Activity 3: Line Test!

For the scatterplot determine the regression line. Plot the line to test whether this line is closed to the points. Skip testing the significance of r and do the following:

a. Compute r.
b. Find the slope (b) and the y-intercept (a).
c. Find the regression line.
d. Plot the regression line. Is the line closest to the points?
e. Predict value of Y when X =2.5. Show this in the graph.





Deepen

The following data show the age of a car and the average mileage/liter.

Age(in years)	0	1	2	3	4	5	6
Mileage per liter (in km)	20.6	18.1	16.3	15.5	14.1	13.9	11.2

a. Find the regression line that will predict the average mileage/liter of the car.
b. Find the average mileage of the car at age 10 years.

Gauge

Directions: Choose the letter of the correct answer. Write your answer on a separate sheet of paper.

1. What is the relationship between two sets of variables used to describe or predict information?
   A. Correlation
   B. Linear regression
   C. Regression line
   D. Regression analysis

2. What do we call a prediction where a variable (y) is dependent on a second variable (x) based on the regression equation of a given set of data?
   A. Correlation
   B. Regression Line
   C. Regression analysis
   D. Slope-intercept form

3. Which of the following is the same as the point-slope form equation of a line in algebra?
   A. Correlation
   B. Regression Line
   C. Regression analysis
   D. Slope-intercept form

4. Which of the following problems could be studied through the use of regression analysis?
   A. Which runner has been on a track team the longest?
   B. Which runner on a track team has the fastest average speed?
   C. Which runner on a track team has the fastest recorded speed?
   D. How the speeds of runners on a track team relate to how long they each train?

5. In regression analysis the variable that is being predicted is the
   A. The x-variable
   B. The independent variable
   C. The predictor variable
   D. The explanatory variable

6. The difference between regression analysis and correlation analysis is
   A. Regression enables prediction of the independent variable
   B. Regression estimates the line of best fit through the data
   C. Regression provides measures of association in units of the variable being measured
   D. All of the above

7. The number of degrees of freedom associated with the standard error of estimate is

A. n-1 since only the slope is estimated from sample data
B. n-1 since only the intercept is estimated from sample data
C. n-1 since only the predicted value of y is estimated from sample data
D. n-2 since the slope and the intercept are estimated from sample data

8. Regression modeling is a statistical framework for developing a mathematical equation, what is being describe in the model?
   A. several explanatory and several response variables response are related
   B. one explanatory and one or more response variables are related
   C. one response and one or more explanatory variables are related
   D. All of these are correct.

9. Research shows that the emotional intelligence of a person is related to his/her academic performance. Likewise, academic performance is related to job performance. Which ordered pair of variables corresponds to the ordered pair (dependent, independent)?
   I. (job performance, academic performance)
   II. (academic performance, intelligence)

   A. I only
   B. Both I and II
   C. II only
   D. Neither

10. In regression analysis, if the independent variable is measured in kilograms, what would be the unit of the dependent variable?
   A. Must be also in kilograms
   B. Must be in some unit of weight
   C. Cannot be in kilograms
   D. Can be any units

11. A regression analysis between sales (in P1000) and price (in peso) resulted in the following equation: Y' = 50,000 - 8X, what is the implication of the given equation?
   A. increase of P 1 in price is associated with a decrease of P 8 in sales
   B. increase of P 1 in price is associated with a decrease of P 8000 in sales
   C. increase of P 8 in price is associated with an increase of P 8,000 in sales
   D. increase of P 1 in price is associated with a decrease of P 42,000 in sales

For numbers 12-13, use the situation below:

A student conducted a regression analysis between the math grades of his classmates and number of times they were absent in the subject. He found that the

regression line that will predict the grade (y) if the number of absences (x) is known $y = 97.732 - 2.61x$.

12. What is the predicted grade of a student who has no absences?
    A. 93 B. 98 C. 97 D. 95
13. What is the predicted grade of a student who has 5 absences?
    A. 79 B. 82 C. 85 D. 90

14. For the regression line $y = 2.6x + 0.56$, what will be the value of y if x = 3.5?
    A. 7.66 B. 8.66 C. 9.66 D. 10.66

15. The relationship between number of beers consumed (x) and blood alcohol content (y) was studied in 16 male college students by using least squares regression. The following regression equation was obtained from this study: Y' = -0.0127 + 0.0180x. What does the equation imply?
    A. Each beer consumed increases blood alcohol by 1.27%.
    B. Each beer consumed increases blood alcohol by exactly 0.018.
    C. On the average it takes 1.8 beers to increase blood alcohol content by 1%.
    D. Each beer consumed increases blood alcohol by an average amount of 1.8%

References

Books

Luis Allan B. Melosantos, Janice F. Antonio, Josephine R. Sacluti, Ryan M. Bruce. 2016. *Math Connections in the Digital Age: Statistics and Probability*. Manila. Sibs Publishing Inc.

Rene R. Belecina, Elisa S. Baccay, Efren B. Mateo. 2016. *Statistics and Probability*. Manila. Rex Bookstore, Inc

Websites

Linear regression multiple choice questions from https://www.mindcoral.com/question/library/1499

Quiz: Simple Linear Regression from https://cliffnotes.com/study-guides/statistics/bivariate-relationships/quiz-simple-linear-regression