

# Regression Analysis

## OVERVIEW

---

In this lab, you will work with the statistical method of linear **regression**. This lab is designed to extend your basic knowledge of linear regression. You will practice linear regression and use it to enhance and expand the method of statistical correlation.

### OBJECTIVES

By the end of the laboratory, you will be able to

- Understand the general concepts of linear regression.
- Draw a line of best fit from a scatterplot.
- Find the slope-intercept form of the equation of a line.
- Use this line to predict Y values from X values.
- Find and interpret the coefficient of determination ( $r^2$ )
- Understand linear regression as it applies to psychology.

### EQUIPMENT

- PC with *Minitab*
- Printer
- Disk with project entitled *correlationdata*
- Hard copies of Correlation scatterplots

## BACKGROUND MATERIAL

---

### Statistical Terms and Topics

- Prediction
- Linear regression line / Line of best fit
- Slope
- Intercept
- Method of least squares
- Coefficient of determination ( $r^2$ )

### Formulas

Slope-intercept form of the equation for the linear regression prediction equation is

$$y = bX + a$$

Where:

- $\hat{Y}$  = predicted score
- $b$  = slope of the line
- $a$  =  $Y$  intercept

## Scenario 1

If you remember from the previous lab, correlation lets us know the **direction** and the **strength** of a relationship between two variables, and this relationship is quantified in the form of a Pearson's Correlation Coefficient. (Find the printed copies of Scatterplots 1 – 4. If you do not have the printed scatterplots with you, open the project entitled *correlationdata* and print scatterplots 1 – 4).

*Scatterplot3* is the first scatterplot you will work with. I have chosen this scatterplot and data set of the four because this one has the strongest relationship ( $r = .235$ ). The relationship of the data will also be displayed in the ***fitted line plot*** you will create for this data set.



## COMPUTER EXERCISE

1. For *scatterplot3*, draw **by-hand** an estimated line of best fit onto the printed scatterplot. Visualize the line which "averages out" the  $y$ -values. There should be approximately the same number of points above the line you draw as below the line. The line you draw does not have to pass through any of the data points, but it can.
2. Use *Minitab* to find  $r^2$  and the regression line for *scatterplot3*.
  - Go to **STAT>REGRESSION>FITTED LINE PLOT**.
  - Choose *siblings* for  $x$  and *number of pets* for  $y$ .
  - Click **OK**.
  - **Print** the graph.
3. Compare *Minitab*'s regression line with the line of best fit you drew by hand. Are the two lines similar? \_\_\_\_\_
4. What is the equation of the line that *Minitab* computed? \_\_\_\_\_

5. What is the  $r^2$  value? \_\_\_\_\_

Now that you know how to generate a regression line and the  $r^2$  value, let's discuss how to interpret and utilize this information. The  $r$ -value (i.e. the correlation coefficient that you found in the Correlation Laboratory found prior to this lab in this workbook) will tell us the direction of the relationship and how strong it is. On the other hand,  $r^2$  tells us the percent of variability in  $y$  that is explained by  $x$ . By computing a least squares regression (line of best fit) we can predict or estimate a  $y$  value based upon a given  $x$  value. In this case, we could predict number of pets ( $y$ ) based upon number of siblings ( $x$ ).

To predict a value for  $y$  is actually quite simple. Using the regression equation, simply substitute the  $x$  value into the equation and solve for  $y$ . For example, to predict how many pets a person having four siblings would have, plug four into the equation for  $x$  and solve for  $y$ .

$$y = 0.94 + 0.50x$$

$$y = 0.94 + 0.50(4)$$

$$y = 0.94 + 2$$

$$y = 2.94 / \text{ or approximately 3 pets}$$

1. Based upon the regression equation, how many **pets** would a person with **2 siblings** have? Show your work below.

$y =$  predicted number of pets = \_\_\_\_\_

2. Based upon the regression equation, how many **pets** would a person with **12 siblings** have? Show your work below.

$y =$  predicted number of pets = \_\_\_\_\_

## Scenario 2

1. Now you choose the next scatterplot (of the three remaining) with the strongest relationship. *Hint: the Pearson's  $r$ -values can help you determine this in addition to the graph.* If you do not already have the scatterplots printed, then do so now.

2. For the scatterplot you have chosen, draw an estimated line of best fit onto the printed scatterplot.
3. Use *Minitab* to find  $r^2$  and the regression line for the scatterplot you have chosen
  - Go to **Stat>Linear Regression>Fitted Line Plot**.
  - Choose columns for  $x$  and  $y$  according to the scatterplot you have chosen.
  - Click **OK**.
  - **Print** the graph.
4. Compare *Minitab*'s regression line with the line of best fit you drew by hand. Are the two lines similar?
5. What is the equation of the line that *Minitab* computed? \_\_\_\_\_
6. What is the  $r^2$  value? \_\_\_\_\_
7. How would you interpret the  $r^2$  value in terms of  $x$  and  $y$ ? \_\_\_\_\_  
\_\_\_\_\_
8. For the scatterplot you have chosen, what predictions could you make using the regression equation? \_\_\_\_\_  
\_\_\_\_\_

Again, choose the scatterplot with the next greatest correlation and complete steps 1-8 from section 2 above . You should have a total of three scatterplots with regression lines and equations when finishing this exercise.

Regression equation \_\_\_\_\_,  $r^2$  \_\_\_\_\_, interpretation \_\_\_\_\_  
 Prediction \_\_\_\_\_

## Application

If we want to know if SAT scores are related to college GPA, we can correlate the two variables using a Pearson's Correlation Coefficient. However, if you wanted to predict someone's college GPA based on their SAT scores, you could use the techniques of linear regression that you have practiced in this exercise.

### Ethics Application

Be careful to remember that correlation does not mean causation. Correlation and linear regression do not imply that changes in  $x$  cause changes in  $y$ . The researcher must still interpret the statistical information in the context of the situation.