*STATISTICS IS THE GRAMMAR OF SCIENCE*

# PROBABILITY AND STATISTICS

# LECTURE – 21

## REGRESSION ANALYSIS
### CORRELATION ANALYSIS

PREPARED BY
**HAZBER SAMSON**
FAST NUCES ISLAMABAD

# REGRESSION ANALYSIS

***INTRODUCION:*** At the beginning of the nineteenth century, **Legendre and Gauss** published papers on the *method of least squares*, which implemented the earliest form of what is now known as *linear regression*. The term *regression* was first introduced by **Francis Galton**. In a famous paper, Galton found that, although there was a tendency for tall parents to have tall children and for short parents to have short children, the average height of children born of parents of a given height tended to move or "regress" toward the average height in the population as a whole.1 In other words, the height of the children of unusually tall or unusually short parents tends to move toward the average height of the population. Galton's *law of universal regression* was confirmed by his friend Karl Pearson, who collected more than a thousand records of heights of members of family groups.2 He found that the average height of sons of a group of tall fathers was less than their fathers' height and the average height of sons of a group of short fathers was greater than their fathers' height, thus "regressing" tall and short sons alike toward the average height of all men. In the words of Galton, this was "regression to mediocrity. The modern interpretations of regression is, as follows however,

***DEFINITION:*** Regression analysis is a statistical technique for investigating and modeling the relationship between variables.

Applications of regression are numerous and occur in almost every field, including engineering, the physical and chemical sciences, economics, management, life and biological sciences, and the social sciences. In fact, regression analysis may be the most widely used statistical technique.

Usually, Researchers want more than just finding the best linear approximation of one variable given a set of others. They want valid statistical relationships. They want to draw conclusions about what happens if one of the variables actually changes. That is: they want to say something about things that are not observed (yet). In this case, they want to find a fundamental relationship. To do this it is assumed that there is a general relationship that is valid for all possible observations from a well-defined population. Restricting attention to linear relationships, we specify a statistical model as linear regression model.
Some examples of analyses using regression models include the following:

• Estimating weight gain by the addition to children's diet of different amounts of various dietary supplements
• Predicting scholastic success (grade point ratio) based on students' scores on an aptitude or entrance test
• Estimating amounts of sales associated with levels of expenditures for various types of advertising
• Predicting fuel consumption for home heating based on daily temperatures and other weather factors
• Estimating changes in interest rates associated with the amount of deficit spending

**Regression analysis** is used to predict the value of one variable on the basis of the other variables. Regression analysis is generally classified into two kinds, **simple** and **multiple**. The regression will be **simple** if it involves only one independent variable and it will be **multiple** if it involves more than one independent variables.

# CORRELATION ANALYSIS

The primary interest of a regression analysis is to make inferences about the dependent variable using information from the independent variable. However, this is not always the case. For example, suppose that we have measurements on the height and weight of a sample of adult males. In this particular study, instead of wanting to estimate weight as a function of height (or vice versa), we simply want an indicator of the strength of the relationship between these measurements.

*A group of techniques used to measure the strength of linear association between two variables is called correlation analysis.* Correlation is the degree of co-variation between the variables. The simple correlation measures the strength or closeness of linear relationship between two variables.

***EXAMPLES***

1- Correlation between height and weights of players of cricket
2- Correlation between marks of students in economics and statistics

A correlation model describes the strength of the relationship between two variables. In a correlation model, both variables are **random variables**, and the model specifies a joint distribution of both variables instead of the conditional distribution of y for a fixed value of x.

As correlation and regression are related concepts, they are often confused, and it useful to repeat the basic definitions of the two concepts:

***REGRESSION MODEL*** The regression model describes a linear relationship where an independent or factor variable is used to estimate or explain the behavior of the dependent or response variable. In this analysis, one of the variables, x, is "**fixed**", or chosen at particular values. The other, y, is the only variable subject to a random error.

***CORRELATION MODEL*** The correlation Model describes the strength of linear relationship between two variables, where both are **random variables**.

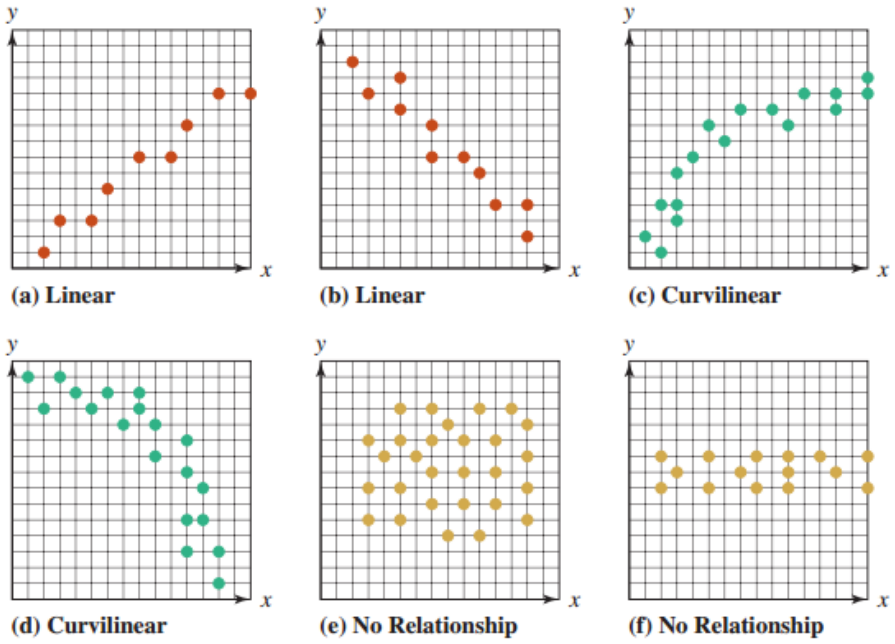
The basic idea of correlation analysis is to report the relationship between two variables. The usual first step is to plot the data in a **scatter diagram**.

***SCATTER PLOT*** A two-dimensional plot showing the values for the joint occurrence of two quantitative variables. The scatter plot may be used to graphically represent the relationship between two variables. It is also known as a scatter diagram.

Scatter Plots are easy to understand and they give us quick picture of linear relationship. The linear relationship can be either positive (as the x variable increases, the y variable also increases) or negative (as the x variable increases, the y variable decreases).

Scatter Plots give us basic idea but for formal discussion we need a quantitative measure of the strength of the linear relationship between two variables. So let's study the quantitative measure of correlation which is known as correlation coefficient.

## SCATTER PLOTS SHOWING DIFFERENT RELATIONSHIPS



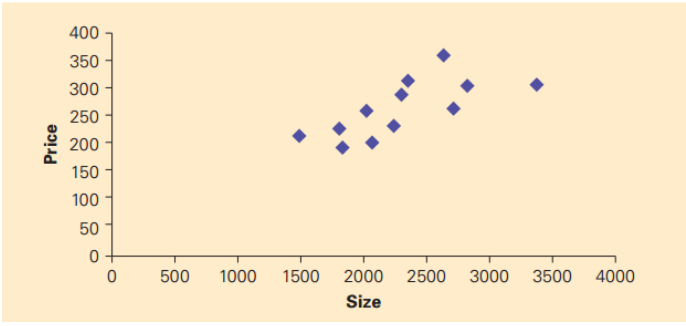

## EXAMPLE-2.15 ANALYZING THE RELATIONSHIP BETWEEN PRICE AND SIZE OF HOUSE

A real estate agent wanted to know to what extent the selling price of a home is related to its size. To acquire this information, he took a sample of 12 homes that had recently sold, recording the price in thousands of dollars and the size in square feet. These data are listed in the accompanying table. Use a graphical technique to describe the relationship between size and price.

| Size ($ft^2$) | Price ($1,000) |
|---|---|
| 2,354 | 315 |
| 1,807 | 229 |
| 2,637 | 355 |
| 2,024 | 261 |
| 2,241 | 234 |
| 1,489 | 216 |
| 3,377 | 308 |
| 2,825 | 306 |
| 2,302 | 289 |
| 2,068 | 204 |
| 2,715 | 265 |
| 1,833 | 195 |

***SOLUTION*** We apply Procedure discussed above

**FIGURE – 1 SCATTER PLOT OF SIZE VS PRICE**




**Interpretation**: The scatter diagram reveals that, in general, the greater the size of the house, the greater the price. However, there are other variables that determine price.

# CORRELATION COEFFICIENT

Several statistics can be used to measure the correlation between two quantitative variables. The statistic most commonly used is the linear correlation coefficient, r, which is also called the **Pearson product moment correlation coefficient (PPMC)** in honor of its developer, Karl Pearson. He originated the correlation coefficient about 1900.

Correlation coefficient describes the strength of the relationship between two sets of interval-scaled or ratio-scaled variables. Designated r, it is often referred to as Pearson's rand as the Pearson product-moment correlation coefficient. It can assume any value from −1.00 to +1.00 inclusive. A correlation coefficient of −1.00 or +1.00 indicates perfect correlation. How the scatter diagram would appear if the relationship between the two variables were linear and perfect is shown in figure below.

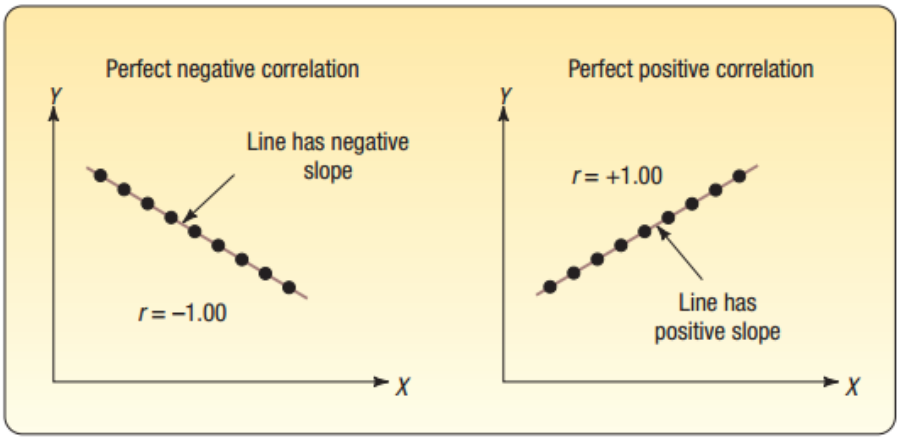
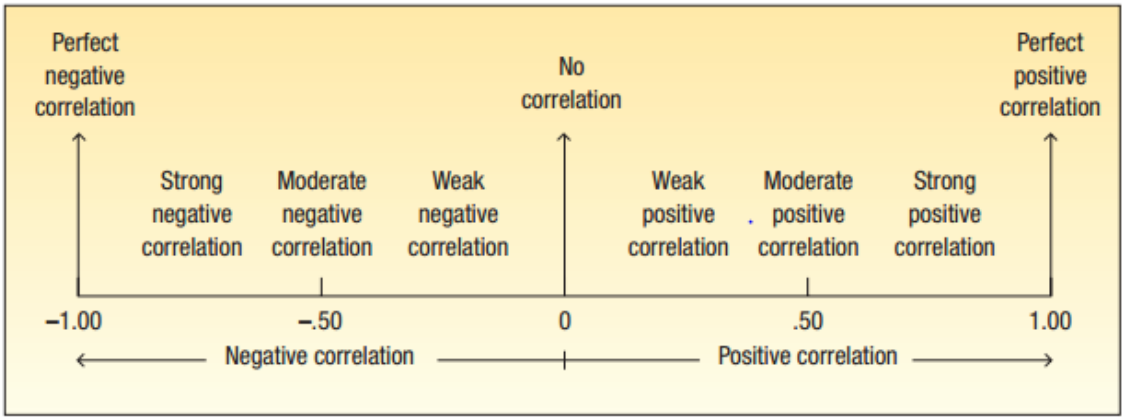



**CHART 13–2** Scatter Diagrams Showing Perfect Negative Correlation and Perfect Positive Correlation

**CORRELATION COEFFICIENT** A measure of the strength of the linear relationship between two variables.

The following drawing summarizes the strength and direction of the correlation coefficient.



## COEFFICIENT OF CORRELATION

A numerical measure of strength of linear relationship between two variables is called correlation coefficient or coefficient of correlation. It is denote by $r$ .

If $y = a + b\,x$ than $r$ is given by

$$r = \frac{n\left(\sum xy\right) - \left(\sum x\right)\left(\sum y\right)}{\sqrt{[n\sum x^2 - \left(\sum x\right)^2][n\sum y^2 - \left(\sum y\right)^2]}}$$

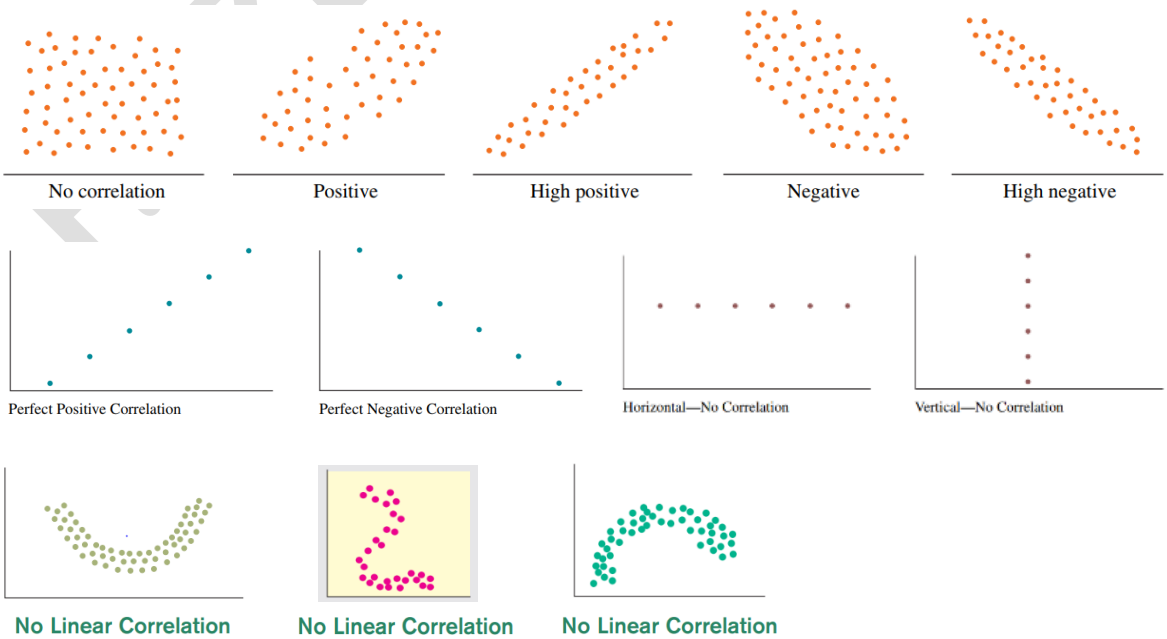
$$r = \frac{\sum\left(x-\bar{x}\right)\left(y-\bar{y}\right)}{\sqrt{\sum\left(x-\bar{x}\right)^2 . \sum\left(y-\bar{y}\right)^2}}$$

$$r = \frac{S_{xy}}{\sqrt{S_x^2 . S_y^2}} = \frac{S_{xy}}{S_x . S_y} \quad where\ S_{xy} = \frac{\sum\left(x-\bar{x}\right)\left(y-\bar{y}\right)}{n-1},\ S_x^2 = \frac{\sum\left(x-\bar{x}\right)^2}{n-1},\ S_y^2 = \frac{\sum\left(y-\bar{y}\right)^2}{n-1}$$

### CHARACTERISTICS OF THE CORRELATION COEFFICIENT

1. The sample correlation coefficient is identified by the lowercase letter *r*.
2. It shows the direction and strength of the linear relationship between two interval- or ratio-scale variables.
3. It ranges from −1 up to and including +1.
4. A value near 0 indicates there is little linear relationship between the variables.
5. A value near 1 indicates a direct or positive linear relationship between the variables.
6. A value near −1 indicates an inverse or negative linear relationship between the variables.

## SCATTER PLOTS AND CORRELATION



## COEFFICIENT OF DETERMINATION

Note that we cannot precisely interpret the meaning of coefficient of correlation except for -1, 0, and +1. We can judge the coefficient of correlation in relation to its proximity to only -1, 0, and +1. Fortunately, we have another measure that can be precisely interpreted. It is the **coefficient of determination**, which is calculated by squaring the coefficient of correlation. For this reason, we denote it $R^2$.

***DEFINITION*** The coefficient of determination measures the amount of variation in the dependent variable that is explained by the variation in the independent variable.

For example, if the coefficient of correlation is -1 or +1, a scatter diagram would display all the points lining up in a straight line. The coefficient of determination is 1, which we interpret to mean that 100% of the variation in the dependent variable *Y* is explained by the variation in the independent variable *X*. If the coefficient of correlation is 0, then there is no linear relationship between the two variables, $R^2 = 0$ and none of the variation in *Y* is explained by the variation in *X* .

***FORMULA*** Coefficient of Determination = Square of the Coefficient of Correlation

$$R^2 = r^2$$

For Example if the coefficient of correlation between two variables X and Y was calculated to be *r* = 0.8711. Then the coefficient of determination will $R^2 = (0.8711)^2 = 0.7588$. This tells us that 75.88% of the variation in Y is explained by X. The remaining 24.12% is due to other factors.

***PROPERTIES OF COEFFICIENT OF DETERMINATION***

$1 - Coefficient\ of\ \det ermination\ cannot\ be\ negative.\ i,e.\ R^2 \geq 0$

$2 - Coefficient\ of\ \det ermination\ lies\ between\ 0\ and\ 1.\ i,e.\ 0 \leq R^2 \leq 1$

***EXAMPLE-1*** consider the data given below

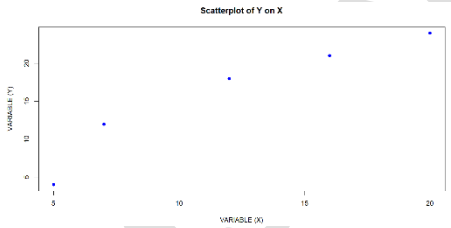
| | | | | | |
|---|---|---|---|---|---|
| X | 5 | 7 | 12 | 16 | 20 |
| Y | 4 | 12 | 18 | 21 | 24 |

(a) Construct a Scatter Plot.
(b) Determine the Coefficient of Correlation and interpret it.
(c) Determine the Coefficient of Determination and interpret it.

***SOLUTION*** First of all we shall construct the scatter plot and analyze.

**(A) *SCATTER PLOT***

Scatter plot shows that there exists a positive relation between X and Y.




**(B) *COEFFICIENT OF CORRELATION***

Correlation Coefficient is given by $r = \dfrac{n\left(\sum xy\right) - \left(\sum x\right)\left(\sum y\right)}{\sqrt{[n\sum x^2 - \left(\sum x\right)^2][n\sum y^2 - \left(\sum y\right)^2]}}$

Let's make table for the values used in formulas

| $X$ | $Y$ | $XY$ | $X^2$ | $Y^2$ |
|---|---|---|---|---|
| 5 | 4 | 20 | 25 | 16 |
| 7 | 12 | 84 | 49 | 144 |
| 12 | 18 | 216 | 144 | 324 |
| 16 | 21 | 336 | 256 | 441 |
| 20 | 24 | 480 | 400 | 576 |
| **60** | **79** | **1136** | **874** | **1501** |

So $r = \dfrac{n\left(\sum xy\right) - \left(\sum x\right)\left(\sum y\right)}{\sqrt{[n\sum x^2 - \left(\sum x\right)^2][n\sum y^2 - \left(\sum y\right)^2]}}$

$$r = \frac{(5)(1136) - (60)(79)}{\sqrt{[(5)(874) - (60)^2][(5)(1501) - (79)^2]}} = \frac{940}{986.55} = 0.95$$

**Interpretation:** Since $r = 0.95 > 0.80$ which shows that that there is a strong positive correlation between variables x and y.

**(C) *COEFFICIENT OF DETERMINATION***

$R^2 = (0.95)^2 = 0.90$

**Interpretation:** $R^2 = 0.90$ shows that 90% variation in Y is due to X and remaining 10% is due to other factors.

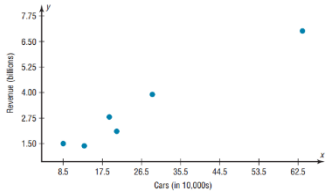
***EXAMPLE-2*** Consider the data shows for car rental companies in the United States for a recent year.

| Company | Cars (in ten thousands) | Revenue (in billions) |
|---|---|---|
| A | 63.0 | $7.0 |
| B | 29.0 | 3.9 |
| C | 20.8 | 2.1 |
| D | 19.1 | 2.8 |
| E | 13.4 | 1.4 |
| F | 8.5 | 1.5 |

Source: *Auto Rental News.*

(a) Construct a Scatter Plot for the data.
(b) Determine the Coefficient of Correlation and interpret it.
(c) Determine the Coefficient of Determination and interpret it.

***SOLUTION (A) SCATTER PLOT***




Scatter Plot shows that there is a positive correlation between number of cars and Revenue.

**(B) *CORRELATION COEFFICIENT***

Correlation Coefficient is given by $r = \dfrac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$

Find the values of $xy$, $x^2$, and $y^2$ and place these values in the corresponding columns of the table.

| Company | Cars $x$ (in 10,000s) | Revenue $y$ (in billions) | $xy$ | $x^2$ | $y^2$ |
|---|---|---|---|---|---|
| A | 63.0 | 7.0 | 441.00 | 3969.00 | 49.00 |
| B | 29.0 | 3.9 | 113.10 | 841.00 | 15.21 |
| C | 20.8 | 2.1 | 43.68 | 432.64 | 4.41 |
| D | 19.1 | 2.8 | 53.48 | 364.81 | 7.84 |
| E | 13.4 | 1.4 | 18.76 | 179.56 | 1.96 |
| F | 8.5 | 1.5 | 12.75 | 72.25 | 2.25 |
| | $\Sigma x = 153.8$ | $\Sigma y = 18.7$ | $\Sigma xy = 682.77$ | $\Sigma x^2 = 5859.26$ | $\Sigma y^2 = 80.67$ |

$$So\ r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}} = \frac{(6)(682.77) - (153.8)(18.7)}{\sqrt{[(6)(5859.26) - (153.8)^2][(6)(80.67) - (18.7)^2]}} = 0.9819$$

**Interpretation:** $r = 0.9819$ shows that there is a strong positive relationship between the number of cars a rental agency has and its annual revenue.

**(C) *COEFFICIENT OF DETERMINATION*** $R^2 = r^2 = (0.9819)^2 = 0.96$

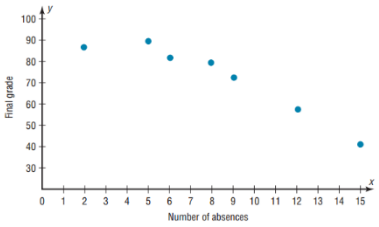
**Interpretation:** $R^2 = 0.96$ shows that 96% variation in revenue is due to number of cars and remaining 9% is due to other factors.

***EXAMPLE-3*** Consider the data obtained in a study on the number of absences and the final grades of seven randomly selected students from a statistics class.

| Student | Number of absences $x$ | Final grade $y$ (%) |
|---|---|---|
| A | 6 | 82 |
| B | 2 | 86 |
| C | 15 | 43 |
| D | 9 | 74 |
| E | 12 | 58 |
| F | 5 | 90 |
| G | 8 | 78 |

(a) Construct a Scatter Plot for the data.
(b) Determine the Coefficient of Correlation and interpret it.
(c) Determine the Coefficient of Determination and interpret it.

***SOLUTION* (A) *SCATTER PLOT***




Scatter Plot shows that there is a negative correlation between number of absences and Final Grade in Statistics Course.

**(B) *CORRELATION COEFFICIENT***

Correlation Coefficient is given by $r = \dfrac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$

Find the values of $xy$, $x^2$, and $y^2$; place these values in the corresponding columns of the table.

| Student | Number of absences $x$ | Final grade $y$ (%) | $xy$ | $x^2$ | $y^2$ |
|---|---|---|---|---|---|
| A | 6 | 82 | 492 | 36 | 6,724 |
| B | 2 | 86 | 172 | 4 | 7,396 |
| C | 15 | 43 | 645 | 225 | 1,849 |
| D | 9 | 74 | 666 | 81 | 5,476 |
| E | 12 | 58 | 696 | 144 | 3,364 |
| F | 5 | 90 | 450 | 25 | 8,100 |
| G | 8 | 78 | 624 | 64 | 6,084 |
| | $\Sigma x = 57$ | $\Sigma y = 511$ | $\Sigma xy = 3745$ | $\Sigma x^2 = 579$ | $\Sigma y^2 = 38{,}993$ |

$$So\ r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}} = \frac{(7)(3745) - (57)(511)}{\sqrt{[(7)(579) - (57)^2][(7)(38993) - (511)^2]}} = -0.944$$

**Interpretation:** The value of r indicates a strong negative relationship between number of absences and final grades.

**(C) *COEFFICIENT OF DETERMINATION*** $R^2 = r^2 = (-0.944)^2 = 0.89$

**Interpretation:** $R^2 = 0.89$ shows that 89% variation in final grades is due to Number of absences and remaining 11% is due to other factors.

# EXERCISE – 5.1

## *CORRELATION ANALYSIS*

**1.** **Crimes** The number of murders and robberies per100,000 population for a random selection of states is shown. Is there a linear relationship between the variables?

| **Murders** | 2.4 | 2.7 | 5.6 | 2.6 | 2.1 | 3.3 | 6.6 | 5.7 |
|---|---|---|---|---|---|---|---|---|
| **Robberies** | 25.3 | 14.3 | 151.6 | 91.1 | 80 | 49 | 173 | 95.8 |

**(a)** Construct Scatter Plot for the data and interpret.
**(b)** Find Correlation Coefficient and Coefficient of Determination also interpret the result.
$(Ans: r = 0.804,\ R^2 = 0.65)$

**2.** **State Debt and Per Capita Tax** An economics student wishes to see if there is a relationship between the amount of state debt per capita and the amount of tax per capita at the state level. Based on the following data, can she or he conclude that per capita state debt and per capita state taxes are related?

| **Per capita debt *x*** | 1924 | 907 | 1445 | 1608 | 661 |
|---|---|---|---|---|---|
| **Per capita tax *y*** | 1685 | 1838 | 1734 | 1842 | 1317 |

**(a)** Construct Scatter Plot for the data and interpret.
**(b)** Find Correlation Coefficient and Coefficient of Determination also interpret the result.
$(Ans: r = 0.518,\ R^2 = 0.27)$

**3.** **Faculty and Students** The number of faculty and the number of students are shown for a random selection of small colleges. Is there a significant relationship between the two variables? Switch *x* and *y* and repeat the process. Which do you think is really the independent variable?

| **Faculty** | 99 | 110 | 113 | 116 | 138 | 174 | 220 |
|---|---|---|---|---|---|---|---|
| **Students** | 1353 | 1290 | 1091 | 1213 | 1384 | 1283 | 2075 |

**(a)** Construct Scatter Plot for the data and interpret.
**(b)** Find Correlation Coefficient and Coefficient of Determination also interpret the result.
$(Ans: r = 0.812,\ R^2 = 0.66)$

**4.** **Class Size and Grades** School administrators wondered whether class size and grade achievement (in percent) were related. A random sample of classes revealed the following data. Are the variables linearly related?

| **No. of students** | 15 | 10 | 8 | 20 | 18 | 6 |
|---|---|---|---|---|---|---|
| **Avg. grade (%)** | 85 | 90 | 82 | 80 | 84 | 92 |

**(a)** Construct Scatter Plot for the data and interpret.
**(b)** Find Correlation Coefficient and Coefficient of Determination also interpret the result.
$(Ans: r = -0.673,\ R^2 = 0.45)$

**5.** **Alumni Contributions** The director of an alumni association for a small college wants to determine whether there is any type of relationship between the amount of an alumnus's contribution (in dollars) and the number of years the alumnus has been out of school. The data follow.

| **Years *x*** | 1 | 5 | 3 | 10 | 7 | 6 |
|---|---|---|---|---|---|---|
| **Contribution *y*** | 500 | 100 | 300 | 50 | 75 | 80 |

**(a)** Construct Scatter Plot for the data and interpret.
**(b)** Find Correlation Coefficient and Coefficient of Determination also interpret the result.
$(Ans: r = -0.883,\ R^2 = 0.78)$