Chapter 3

Descriptive Analysis and Presentation of Bivariate Data

Chapter 3 Overview

- 3-1 Bivariate Data
- 3-2 Linear Correlation
- 3-3 Linear Regression

3-1 Bivariate Data

- Bivariate Data is data that has two different variables obtained from the same population.
  - Similar to x vs y data in algebra.
- Possible Types of Data:
  - Two Qualitative Variables
  - One Qualitative Variable & One Quantitative Variable
  - Two Quantitative Variables

Two Qualitative Variables

When bivariate data result from two qualitative (attribute or categorical) variables, the data are often arranged on a **cross-tabulation** or **contingency table**. Let’s look at an example.

Example 1 – *Constructing Cross-Tabulation Tables*

Thirty students from our college were randomly identified and classified according to two variables: gender (M/F) and major (liberal arts, business administration, technology), as shown in Table 3.1.

Name	Gender	Major	Name	Gender	Major	Name	Gender	Major
Adams	M	LA	Feeney	M	T	McGowan	M	BA
Argento	F	BA	Flanigan	M	LA	Mowers	F	BA
Baker	M	LA	Hodge	F	LA	Ornt	M	T
Bennett	F	LA	Holmes	M	T	Palmer	F	LA
Brand	M	T	Jopson	F	T	Pullen	M	T
Brock	M	BA	Kee	M	BA	Rattan	M	BA
Chun	F	LA	Kleeberg	M	LA	Sherman	F	LA
Crain	M	T	Light	M	BA	Small	F	T
Cross	F	BA	Linton	F	LA	Tate	M	BA
Ellis	F	BA	Lopez	M	T	Yamamoto	M	LA

Genders and Majors of 30 College Students **[TA03-01]**

Table 3.1

Example 1 – *Constructing Cross-Tabulation Tables*

cont'd

- These 30 bivariate data can be summarized on a 2×3 cross-tabulation table, where the two rows represent the two genders, male and female, and the three columns represent the three major categories of liberal arts (LA), business administration (BA), and technology (T).

- The entry in each cell is found by determining how many students fit into each category. Adams is male (M) and liberal arts (LA) and is classified in the cell in the first row, first column.

Example 1 – *Constructing Cross-Tabulation Tables*

cont'd

- See the red tally mark in Table 3.2.

Gender	Major					
	LA		BA		T	
M	||||	(5)	|||| |	(6)	|||| ||	(7)
F	|||| |	(6)	|||	(4)	||	(2)

Cross-Tabulation of Gender and Major (tallied)

Table 3.2

- The other 29 students are classified (tallied, shown in black) in a similar fashion.

Example 1 – *Constructing Cross-Tabulation Tables* cont'd

- The resulting 2×3 cross-tabulation (contingency) table, Table 3.3, shows the frequency for each cross-category of the two variables along with the row and column totals, called *marginal totals* (or *marginals*). The total of the marginal totals is the grand total and is equal to *n*, the sample size.

	Major			
Gender	LA	BA	T	Row Total
M	5	6	7	18
F	6	4	2	12
Col. Total	11	10	9	30

Cross-Tabulation of Gender and Major (frequencies)

Table 3.3

Example 1 – *Constructing Cross-Tabulation Tables* cont'd

- Contingency tables often show percentages (relative frequencies). These percentages can be based on the entire sample or on the subsample (row or column) classifications.

- Percentages Based on the Grand Total (Entire Sample)

- The frequencies in the contingency table shown in Table 3.3 can easily be converted to percentages of the grand total by dividing each frequency by the grand total and multiplying the result by 100.

Example 1 – *Constructing Cross-Tabulation Tables* cont'd

- For example, 6 becomes 20% → $\left[\left(\frac{6}{30}\right) \times 100 = 20\right]$

See Table 3.4.

	Major			
Gender	LA	BA	T	Row Total
M	17%	20%	23%	60%
F	20%	13%	7%	40%
Col. Total	37%	33%	30%	100%

Cross-Tabulation of Gender and Major (relative frequencies; % of grand total)

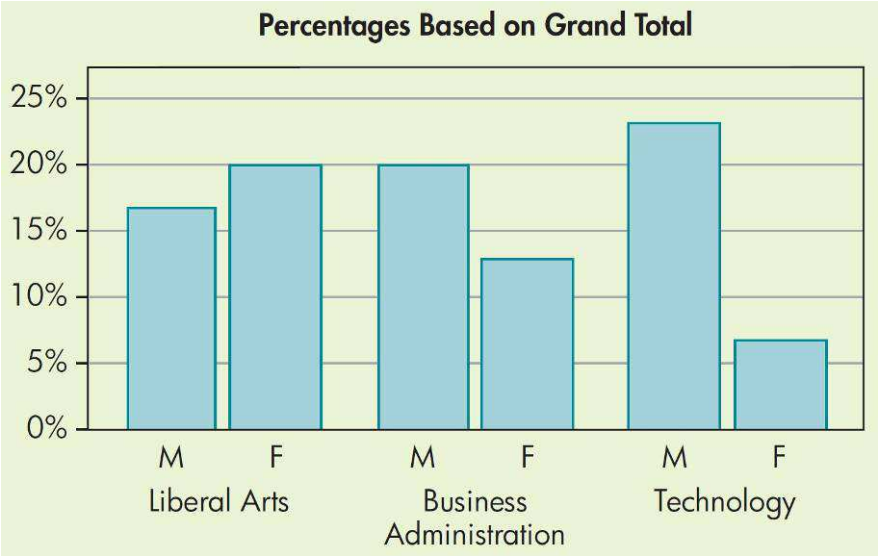
Table 3.4

- From the table of percentages of the grand total, we can easily see that 60% of the sample are male, 40% are female, 30% are technology majors, and so on.

Example 1 – *Constructing Cross-Tabulation Tables*

cont'd

- These same statistics (numerical values describing sample results) can be shown in a bar graph (see Figure 3.1).



Bar Graph

Figure 3.1

3-1 Contingency Table

- The following table represents the percentages of a sample of 100 doctors and nurses and whether or not they smoke.

Smoking?	Doctor	Nurse
Yes	2.88	9.21
No	24.46	63.45

- What percentage of the sample smokes?
- What percentage of the sample was nurses?
- How many people in the sample smoke?

One Qualitative and One Quantitative Variable

When bivariate data result from one qualitative and one quantitative variable, the quantitative values are viewed as separate samples, each set identified by levels of the qualitative variable.

Example 2 – *Constructing Side-by-side Comparisons*

- The distance required to stop a 3000-pound automobile on wet pavement was measured to compare the stopping capabilities of three tire tread designs (see Table 3.7).

Design A ($n = 6$)			Design B ($n = 6$)			Design C ($n = 6$)		
37	36	38	33	35	38	40	39	40
34	40	32	34	42	34	41	41	43

Stopping Distances (in feet) for Three Tread Designs **[TA03-07]**

Table 3.7

- Tires of each design were tested repeatedly on the same automobile on a controlled wet pavement.

Example 2 – *Constructing Side-by-side Comparisons*

cont'd

- The design of the tread is a qualitative variable with three levels of response, and the stopping distance is a quantitative variable.

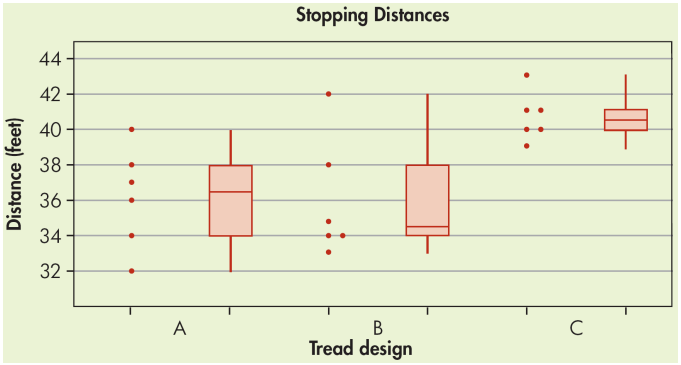
- The distribution of the stopping distances for tread design A is to be compared with the distribution of stopping distances for each of the other tread designs.

- This comparison may be made with both numerical and graphic techniques.

Example 2 – *Constructing Side-by-side Comparisons*

cont'd

- Some of the available options are shown in Figure 3.4, Table 3.8, and Table 3.9.



Dotplot and Box-and-Whiskers Display Using a Common Scale
Figure 3.4

	Design A	Design B	Design C
High	40	42	43
Q_3	38	38	41
Median	36.5	34.5	40.5
Q_1	34	34	40
Low	32	33	39

5-Number Summary for Each Design
Table 3.8

	Design A	Design B	Design C
Mean	36.2	36.0	40.7
Standard deviation	2.9	3.4	1.4

Mean and Standard Deviation for Each Design
Table 3.9

Two Quantitative Variables

- When the bivariate data are the result of two quantitative variables, it is customary to express the data mathematically as **ordered pairs** (x, y), where x is the **input variable** (sometimes called the **independent variable**) and y is the **output variable** (sometimes called the **dependent variable**).

- The data are said to be *ordered* because one value, x, is always written first.

They are called *paired* because for each x value, there is a corresponding y value from the same source.

Two Quantitative Variables

- For example, if *x* is height and *y* is weight, then a height value and a corresponding weight value are recorded for each person.

The input variable, *x*, is measured or controlled in order to predict the output variable, *y*.

Suppose some research doctors are testing a new drug by prescribing different dosages and observing the lengths of the recovery times of their patients.

Two Quantitative Variables

- The researcher can control the amount of drug prescribed, so the amount of drug is referred to as *x*.

In the case of height and weight, either variable could be treated as input and the other as output, depending on the question being asked. However, different results will be obtained from the regression analysis, depending on the choice made.

- In problems that deal with two quantitative variables, we present the sample data pictorially on a *scatter diagram*.

Two Quantitative Variables

- Scatter diagram A plot of all the ordered pairs of bivariate data on a coordinate axis system. The input variable, *x*, is plotted on the horizontal axis, and the output variable, *y*, is plotted on the vertical axis.

- **Note:** When you construct a scatter diagram, it is convenient to construct scales so that the range of the *y* values along the vertical axis is equal to or slightly shorter than the range of the *x* values along the horizontal axis.

- This creates a “window of data” that is approximately square.

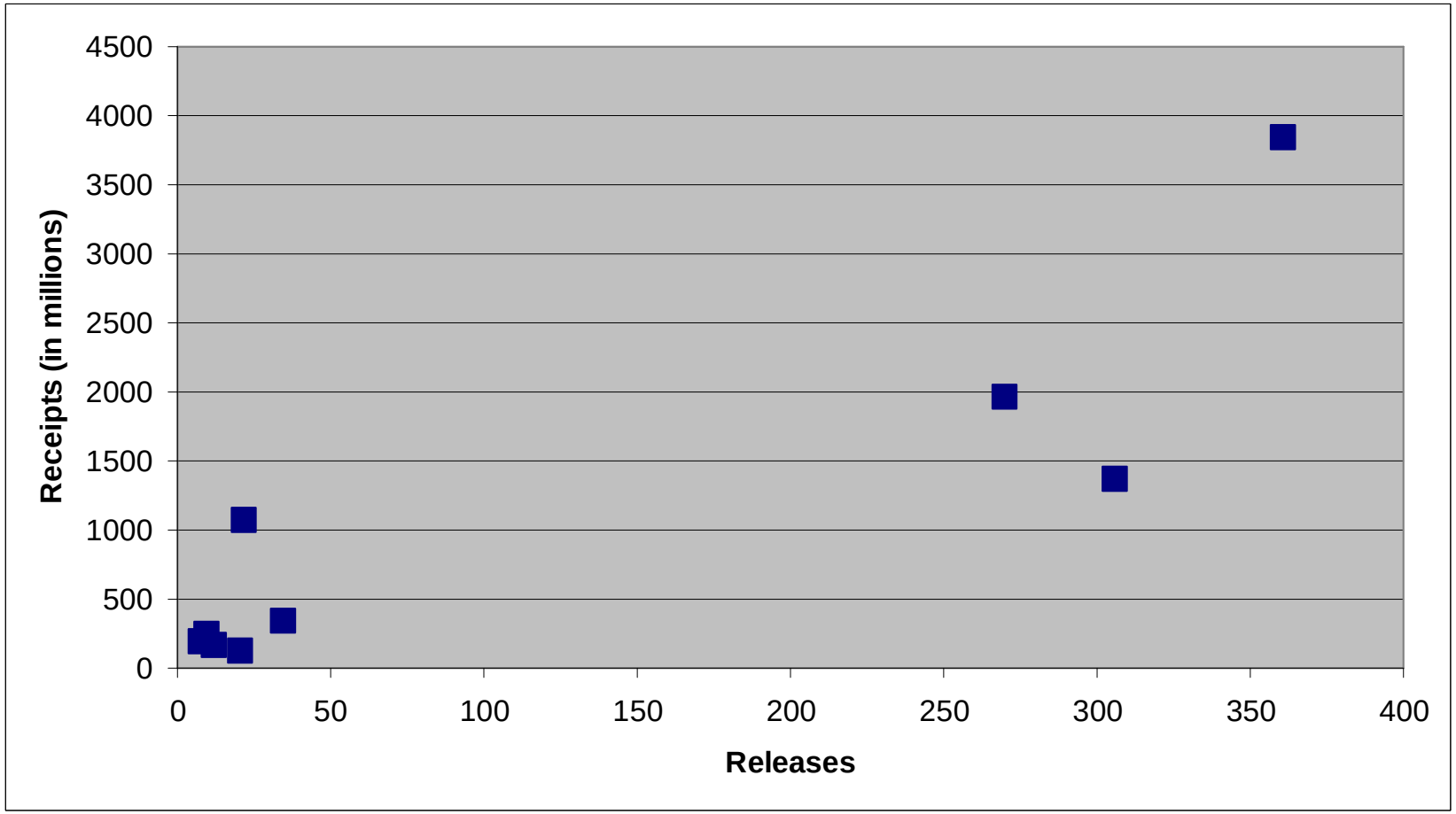
Ex.) 3 – Scatter Plots

Construct a scatter plot for the data shown for the data on gross receipts for movie studios.

No. of releases *x*	361	270	306	22	35	10	8	12	21
Gross receipts *y* (million $)	3844	1962	1371	1064	334	241	188	154	125

Step 1: Draw and label the *x* and *y* axes.
Step 2: Plot each point on the graph.

Ex.) 3 – Scatter Plots



Positive Relationship

Ex.) 4 – Scatter Plots

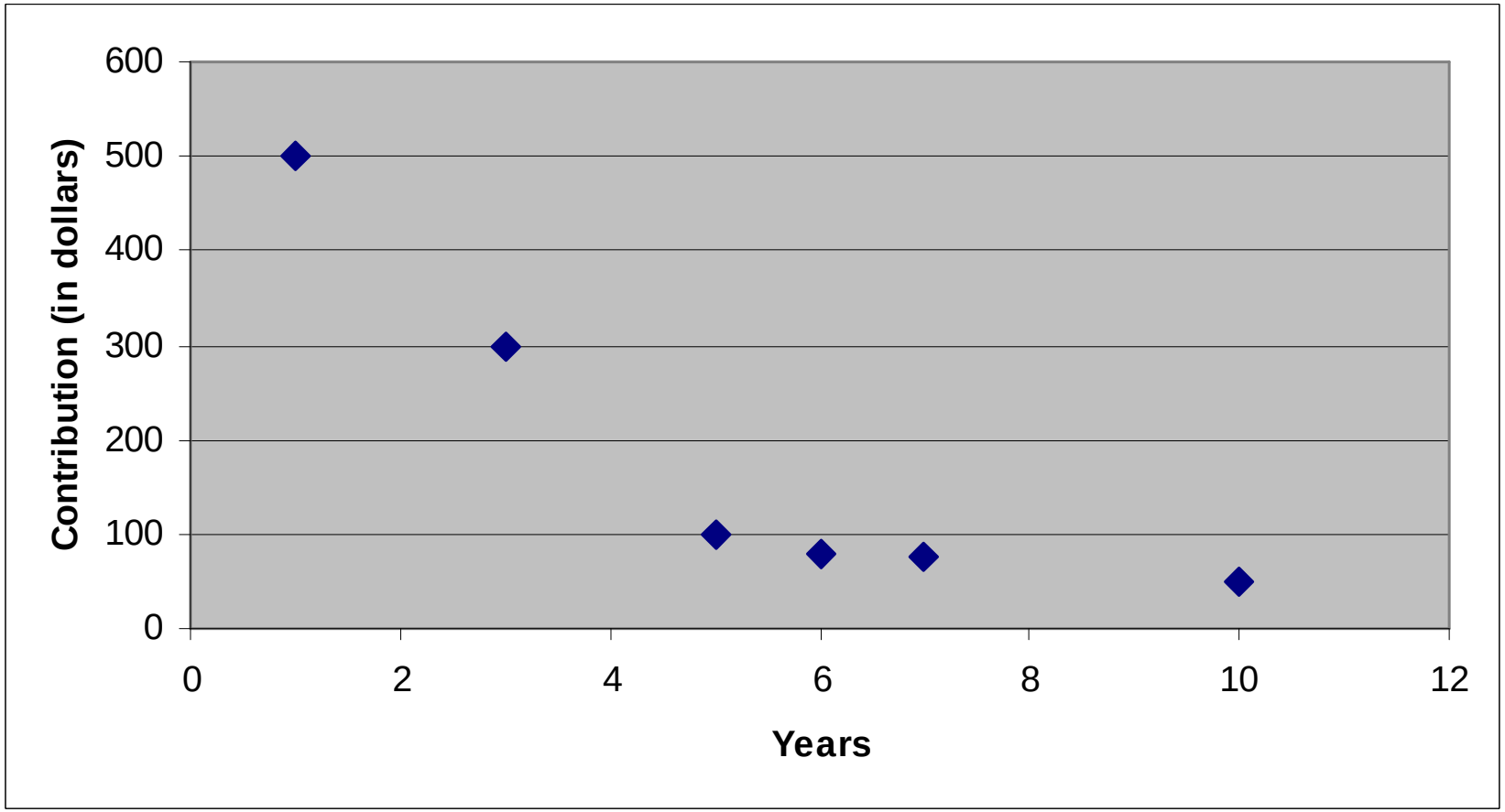
Construct a scatter plot for the data obtained in a study on the number of years an alumnus has been out of school and the amount of money donated.

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

Step 1: Draw and label the x and y axes.

Step 2: Plot each point on the graph.

Ex.) 4 – Scatter Plots



Negative Relationship

Ex.) 5 – Scatter Plots

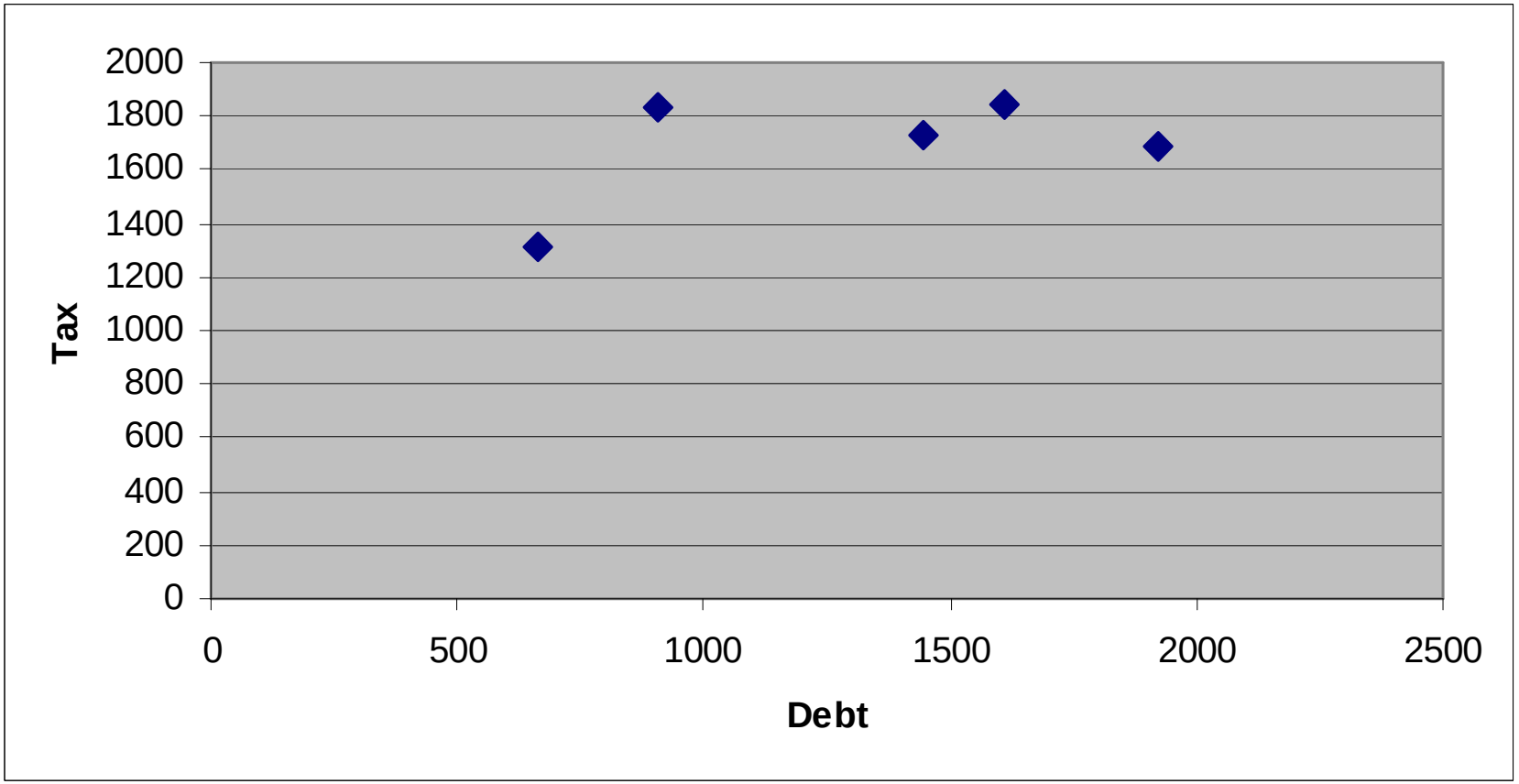
Construct a scatter plot for the data obtained in a study on the amount of state debt per capita and the amount of tax per capita at the state level.

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

Step 1: Draw and label the *x* and *y* axes.

Step 2: Plot each point on the graph.

Ex.) 5 – Scatter Plots



Very Weak Relationship

3-1 Scatter Plots

- For a group of army inductees, the weight, *x,* and exercise capacity, *y*, were recorded for 10 individuals. Draw a scatter plot.

x	180	150	200	155	225	175	130	250	160	190
y	30	25	20	30	15	28	30	20	26	20

Introduction to Section 3-2 & 3-3

- **Correlation** is a statistical method used to determine whether a linear relationship between variables exists.

- **Regression** is a statistical method used to describe the nature of the relationship between variables—that is, positive or negative, linear or nonlinear.

Introduction to Section 3-2 & 3-3

- The purpose of these sections is to answer these questions statistically:
  1. Are two or more variables related?
  2. If so, what is the strength of the relationship?
  3. What type of relationship exists?
  4. What kind of predictions can be made from the relationship?

Introduction to Section 3-2 & 3-3

1. *Are two or more variables related?*
2. *If so, what is the strength of the relationship?*

To answer these two questions, statisticians use the **correlation coefficient**, a numerical measure to determine whether two or more variables are related and to determine the strength of the relationship between or among the variables.

Introduction to Section 3-2 & 3-3

3. What type of relationship exists?

There are two types of relationships: simple and multiple.

In a simple relationship, there are two variables: an **independent variable** (predictor variable) and a **dependent variable** (response variable).

In a multiple relationship, there are two or more independent variables that are used to predict one dependent variable.

Introduction to Section 3-2 & 3-3

4. What kind of predictions can be made from the relationship?

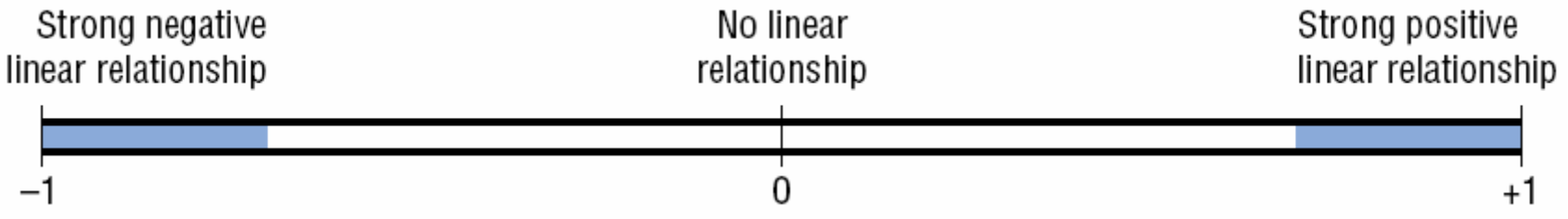
Predictions are made in all areas and daily. Examples include weather forecasting, stock market analyses, sales predictions, crop predictions, gasoline price predictions, and sports predictions. Some predictions are more accurate than others, due to the strength of the relationship. That is, the stronger the relationship is between variables, the more accurate the prediction is.

3-2 Linear Correlation

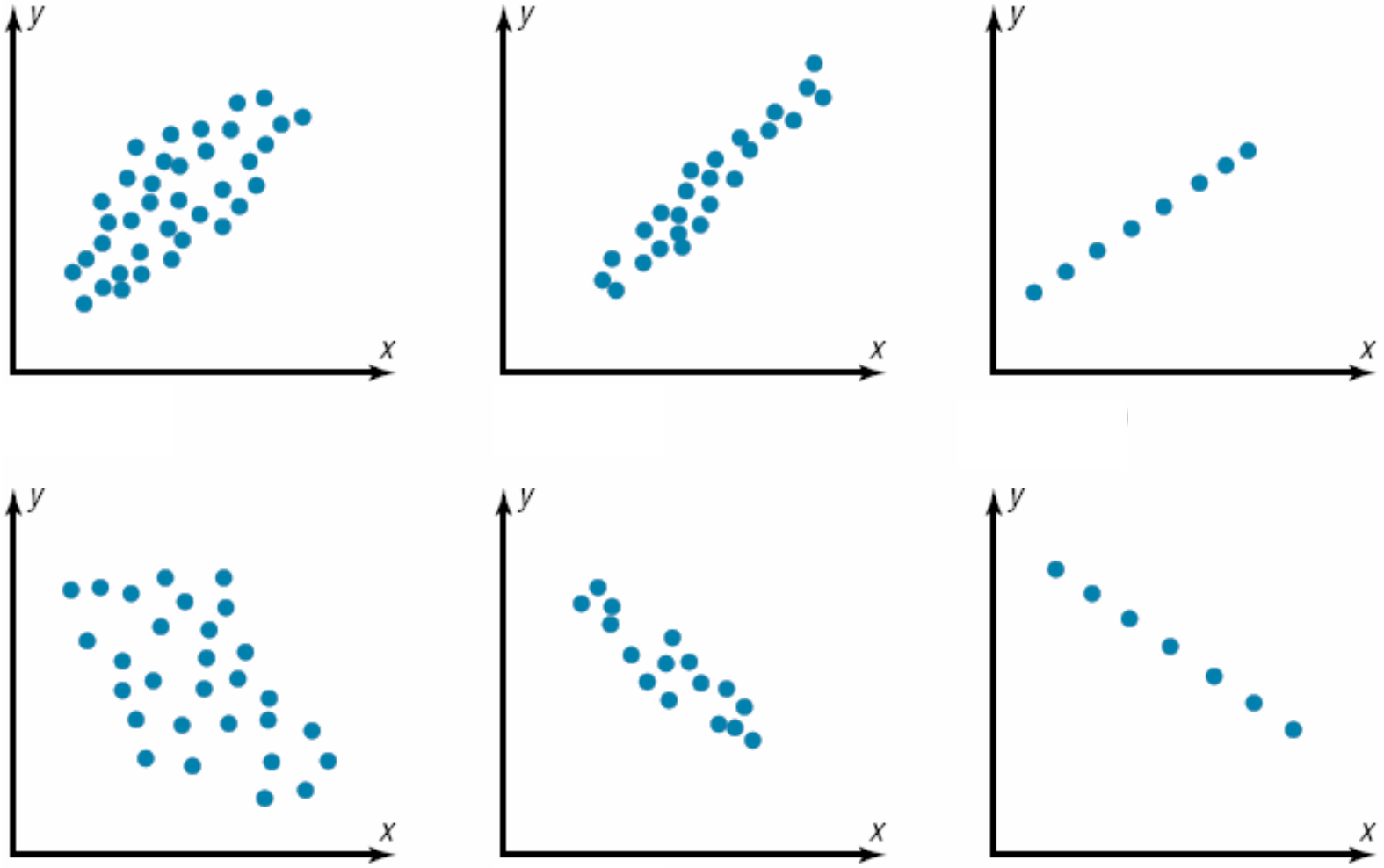
- The **correlation coefficient** computed from the sample data measures the strength and direction of a linear relationship between two variables.
- There are several types of correlation coefficients. The one explained in this section is called the **Pearson product moment correlation coefficient (PPMC)**.
- The symbol for the sample correlation coefficient is r. The symbol for the population correlation coefficient is ρ.

Linear Correlation

- The range of the correlation coefficient is from –1 to +1.
- If there is a **strong positive linear relationship** between the variables, the value of *r* will be close to +1.
- If there is a **strong negative linear relationship** between the variables, the value of *r* will be close to –1.



Linear Correlation



Correlation Coefficient

The formula for the correlation coefficient is

$$r = \frac{s_{xy}}{\sqrt{s_x s_y}}$$

where:

$$s_{xy} = \sum xy - \frac{\sum x \sum y}{n}$$

$$s_x = \sum x^2 - \frac{\left(\sum x\right)^2}{n} \qquad s_y = \sum y^2 - \frac{\left(\sum y\right)^2}{n}$$

where *n* is the number of data pairs.

Rounding Rule: Round to three decimal places.

Ex.) 6 – Correlation Coefficient

Compute the correlation coefficient for the data in Example 3.

	No of releases, x	Gross Receipts, y	xy	x^2	y^2
	361	3844	1387684	130321	14776336
	270	1962	529740	72900	3849444
	306	1371	419526	93636	1879641
	22	1064	23408	484	1132096
	35	334	11690	1225	111556
	10	241	2410	100	58081
	8	188	1504	64	35344
	12	156	1872	144	24336
	21	125	2625	441	15625
sums	1045	9285	2380459	299315	21882459

$\Sigma x = 1045$ $\Sigma y = 9285$

$\Sigma xy = 2380459$ $\Sigma x^2 = 299315$

$\Sigma y^2 = 21882459$

Ex.) 6 – Correlation Coefficient

Compute the correlation coefficient for the data in Example 3.

$\Sigma x = 1045$, $\Sigma y = 9285$, $\Sigma xy = 2380459$,
$\Sigma x^2 = 299315$, $\Sigma y^2 = 21882459$, $n = 9$

1. Calculate s_{xy}

$$s_{xy} = \sum xy - \frac{\sum x \sum y}{n}$$

$$= 2380459 - \frac{(1045)(9285)}{9}$$

$$= 2380459 - 1078091.667$$

$$= 1302367.333$$

Formulas:

$$r = \frac{s_{xy}}{\sqrt{s_x s_y}}$$

$$s_{xy} = \sum xy - \frac{\sum x \sum y}{n}$$

$$s_x = \sum x^2 - \frac{\left(\sum x\right)^2}{n}$$

$$s_y = \sum y^2 - \frac{\left(\sum y\right)^2}{n}$$

Ex.) 6 – Correlation Coefficient

Compute the correlation coefficient for the data in Example 3.

$\Sigma x = 1045$, $\Sigma y = 9285$, $\Sigma xy = 2380459$,
$\Sigma x^2 = 299315$, $\Sigma y^2 = 21882459$, $n = 9$

2. Calculate s_x

$$s_x = \sum x^2 - \frac{\left(\sum x\right)^2}{n}$$

$$= 299315 - \frac{(1045)^2}{9}$$

$$= 299315 - 121336.111$$

$$= 177978.889$$

Formulas:

$$r = \frac{s_{xy}}{\sqrt{s_x s_y}}$$

$$s_{xy} = \sum xy - \frac{\sum x \sum y}{n}$$

$$s_x = \sum x^2 - \frac{\left(\sum x\right)^2}{n}$$

$$s_y = \sum y^2 - \frac{\left(\sum y\right)^2}{n}$$

Ex.) 6 – Correlation Coefficient

Compute the correlation coefficient for the data in Example 3.

$\Sigma x = 1045$, $\Sigma y = 9285$, $\Sigma xy = 2380459$,
$\Sigma x^2 = 299315$, $\Sigma y^2 = 21882459$, $n = 9$

3. Calculate s_y

$$s_y = \sum y^2 - \frac{\left(\sum y\right)^2}{n}$$

$$= 21882459 - \frac{(9285)^2}{9}$$

$$= 21882459 - 9579025$$

$$= 12303434$$

Formulas:

$$r = \frac{s_{xy}}{\sqrt{s_x s_y}}$$

$$s_{xy} = \sum xy - \frac{\sum x \sum y}{n}$$

$$s_x = \sum x^2 - \frac{\left(\sum x\right)^2}{n}$$

$$s_y = \sum y^2 - \frac{\left(\sum y\right)^2}{n}$$

Ex.) 6 – Correlation Coefficient

Compute the correlation coefficient for the data in Example 3.

$s_{xy} = 1302367.333$, $s_x = 177978.889$
$s_y = 12303434$

4. Calculate r

$$r = \frac{s_{xy}}{\sqrt{s_x s_y}}$$

$$= \frac{1302367.333}{\sqrt{(177978.889)(12303434)}}$$

$$= \frac{1302367.333}{\sqrt{2189751514204.826}}$$

$$= \frac{1302367.333}{1479780.901} = \boxed{0.880 \text{ (strong positive relationship)}}$$

Formulas:

$$r = \frac{s_{xy}}{\sqrt{s_x s_y}}$$

$$s_{xy} = \sum xy - \frac{\sum x \sum y}{n}$$

$$s_x = \sum x^2 - \frac{\left(\sum x\right)^2}{n}$$

$$s_y = \sum y^2 - \frac{\left(\sum y\right)^2}{n}$$

Ex.) 7 – Correlation Coefficient

Compute the correlation coefficient for the data in Example 4.

	Years, x	Contribution, y	xy	x^2	y^2
	1	500	500	1	250000
	5	100	500	25	10000
	3	300	900	9	90000
	10	50	500	100	2500
	7	75	525	49	5625
	6	80	480	36	6400
sums	32	1105	3405	220	364525

$$r = -0.883 \text{ (strong negative relationship)}$$

Ex.) 8 – Correlation Coefficient

Compute the correlation coefficient for the data in Example 5.

	Debt, x	Tax, y	xy	x^2	y^2
	1924	1685	3241940	3701776	2839225
	907	1838	1667066	822649	3378244
	1445	1734	2505630	2088025	3006756
	1608	1842	2961936	2585664	3392964
	661	1317	870537	436921	1734489
sums	6545	8416	11247109	9635035	14351678

$r = 0.518$ (slight positive relationship)

Possible Relationships Between Variables

When there is a strong linear correlation between variables, any of the following five possibilities can exist.

1. There is a *direct cause-and-effect* relationship between the variables. That is, *x* causes *y*.
2. There is a *reverse cause-and-effect* relationship between the variables. That is, *y* causes *x*.
3. The relationship between the variables may be *caused by a third variable*.
4. There may be a *complexity of interrelationships* among many variables.
5. The relationship may be *coincidental*.

Possible Relationships Between Variables

1. There is a *direct cause-and-effect* relationship between the variables. That is, *x* causes *y.*

For example,

- ☐ water causes plants to grow
- ☐ poison causes death
- ☐ heat causes ice to melt

Possible Relationships Between Variables

2. There is a *reverse cause-and-effect* relationship between the variables. That is, *y* causes *x.*

For example,

- Suppose a researcher believes excessive coffee consumption causes stress, but the researcher fails to consider that the reverse situation may occur. That is, it may be that an extremely stressed person drinks a lot of coffee to get more accomplished.

Possible Relationships Between Variables

3. The relationship between the variables may be *caused by a third variable*.

For example,

- If a statistician correlated the number of deaths due to drowning and the number of cans of soft drink consumed daily during the summer, he or she would probably find a significant relationship. However, the soft drink is not necessarily responsible for the deaths, since both variables may be related to heat and humidity.

Possible Relationships Between Variables

4. There may be a *complexity of interrelationships* among many variables.

For example,

- A researcher may find a significant relationship between students' high school grades and college grades. But there probably are many other variables involved, such as IQ, hours of study, influence of parents, motivation, age, and instructors.

Possible Relationships Between Variables

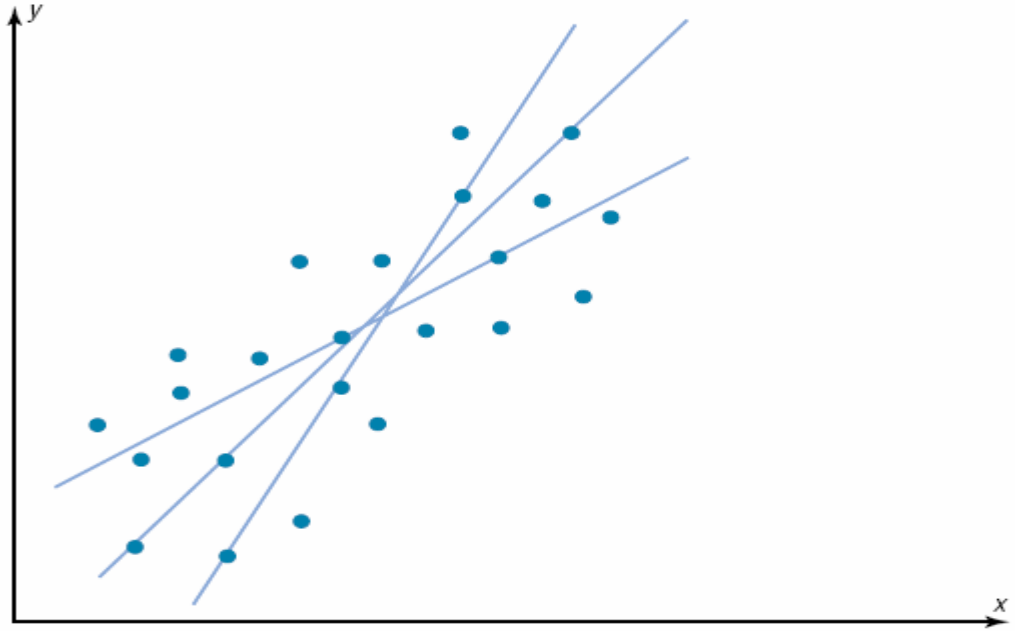
5. The relationship may be *coincidental.*

For example,

- A researcher may be able to find a significant relationship between the increase in the number of people who are exercising and the increase in the number of people who are committing crimes. But common sense dictates that any relationship between these two values must be due to coincidence.

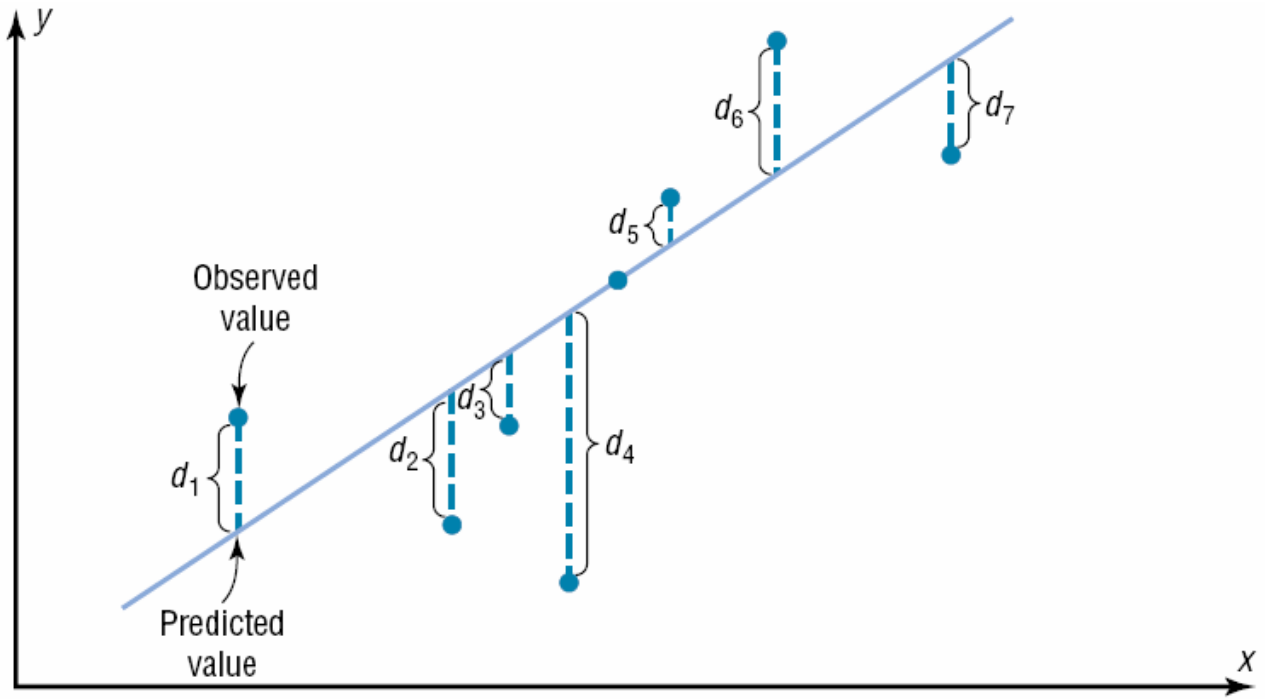
3-3 Linear Regression

- If the value of the correlation coefficient is significant, the next step is to determine the equation of the **regression line** which is the data's line of best fit.



Linear Regression

- **Best fit** means that the sum of the squares of the vertical distance from each point to the line is at a minimum.



Regression Line $\hat{y} = b_0 + b_1 x$

$$b_1 = \frac{SS(xy)}{SS(x)}$$

$$b_0 = \frac{\Sigma y - (b_1 \cdot \Sigma x)}{n}$$

Ex.) 9 – Linear Regression

Find the equation of the regression line for the data in Examples 3 & 6, and graph the line on the scatter plot.

$\Sigma x = 1045$, $\Sigma y = 9285$, $\Sigma xy = 2380459$, $\Sigma x^2 = 299315$, $\Sigma y^2 = 21882459$, $n = 9$, $s_{xy} = 1302367.333$, $s_x = 177978.889$ $s_y = 12303434$

$$b_1 = \frac{SS(xy)}{SS(x)} = \frac{1302367.333}{177978.889} = 7.318$$

$$b_0 = \frac{\Sigma y - (b_1 \cdot \Sigma x)}{n} = \frac{9285 - (7.318 \cdot 1045)}{9} = \frac{1637.69}{9} = 181.966$$

$$\hat{y} = b_0 + b_1 x \rightarrow \hat{y} = 181.966 + 7.318x$$

Ex.) 9 – Linear Regression

Find two points to sketch the graph of the regression line.

Use any *x* values between 8 and 361. For example, let *x* equal 10 and 350. Substitute in the equation and find the corresponding *y* value.

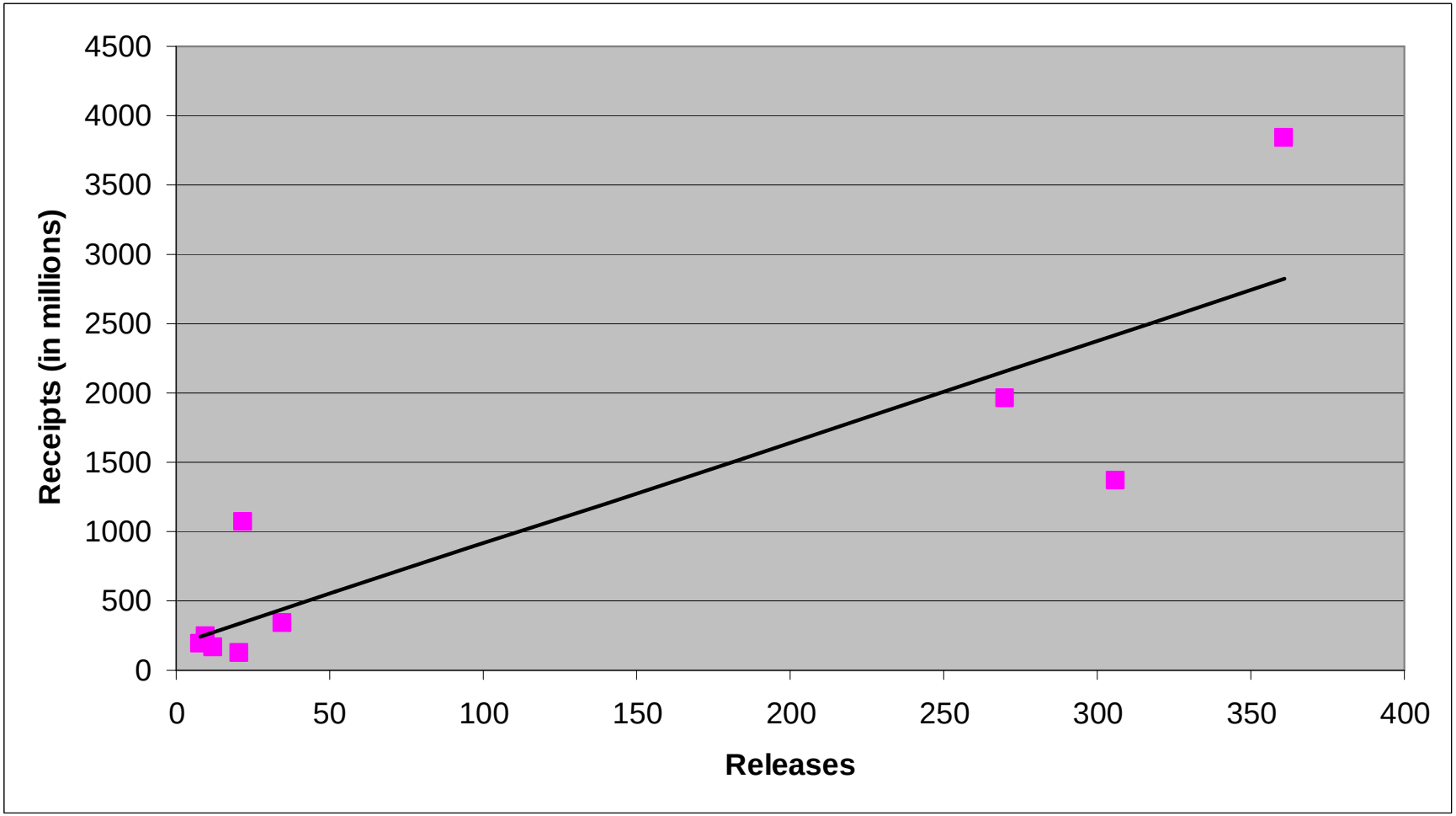
$$\begin{aligned}\hat{y} &= 181.966 + 7.318x \\ &= 181.966 + 7.318(10) \\ &= 255.146\end{aligned}$$

$$\begin{aligned}\hat{y} &= 181.966 + 7.318x \\ &= 181.966 + 7.318(350) \\ &= 2743.266\end{aligned}$$

Plot (10, 255.146) and (350, 2743.266), and sketch the resulting line.

Note: Due to rounding answers might vary slightly but should be very close.

Ex.) 9 – Linear Regression



Ex.) 10 – Equation of the Regression Line

Compute the equation of the regression line for the data in Example 4.

	Years, x	Contribution, y	xy	x^2	y^2
	1	500	500	1	250000
	5	100	500	25	10000
	3	300	900	9	90000
	10	50	500	100	2500
	7	75	525	49	5625
	6	80	480	36	6400
sums	32	1105	3405	220	364525

Ex.) 11 – Equation of the Regression Line

Compute the equation of the regression line for the data in Example 5.

	Debt, x	Tax, y	xy	x^2	y^2
	1924	1685	3241940	3701776	2839225
	907	1838	1667066	822649	3378244
	1445	1734	2505630	2088025	3006756
	1608	1842	2961936	2585664	3392964
	661	1317	870537	436921	1734489
sums	6545	8416	11247109	9635035	14351678

Linear Regression

- The magnitude of the change in one variable when the other variable changes exactly 1 unit is called a **marginal change.** The value of slope b_1 of the regression line equation represents the marginal change.
- For valid predictions, the value of the correlation coefficient must be significant.
- When *r* is not significantly different from 0, the best predictor of *y* is the mean of the data values of *y.*

Assumptions for Valid Predictions

1. For any specific value of the independent variable *x,* the value of the dependent variable *y* must be normally distributed about the regression line.
2. The standard deviation of each of the dependent variables must be the same for each value of the independent variable.

Extrapolations (Future Predictions)

- **Extrapolation**, or making predictions beyond the bounds of the data, must be interpreted cautiously.
- Remember that when predictions are made, they are based on present conditions or on the premise that present trends will continue. This assumption may or may not prove true in the future.

Procedure Table

Step 1: Make a table with subject, x, y, xy, x^2, and y^2 columns.

Step 2: Find the values of xy, x^2, and y^2. Place them in the appropriate columns and sum each column.

Step 3: Substitute in the formula to find the value of r.

Step 4: When r is significant, substitute in the formulas to find the values of a and b for the regression line equation $\hat{y} = b_0 + b_1 x$.