FUNDAMENTALS OF MACHINE LEARNING IN DATA SCIENCE

CSIS 3290

CLASSIFICATION (4)

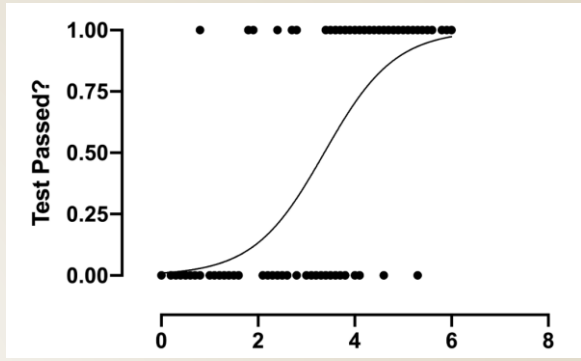
REGRESSION

FATEMEH AHMADI

Regression

- In studying relationships between two variables, collect the data and then construct a scatter plot. The purpose of the scatter plot, as indicated previously, is to determine the nature of the relationship between the variables.
- The possibilities include a positive linear relationship, a negative linear relationship, a curvilinear relationship, or no discernible relationship. After the scatter plot is drawn and a linear relationship is determined, the next steps are to compute the value of the correlation coefficient and test the significance of the relationship.
- If the value of the correlation coefficient is significant, the next step is to determine the **equation of the regression line**, which is the dateline of best fit.
- **Note**: Determining the regression line when *r* is not significant and then making predictions using the regression line are meaningless. The purpose of the regression line is to enable the researcher to see the trend and make predictions on the basis of the data.

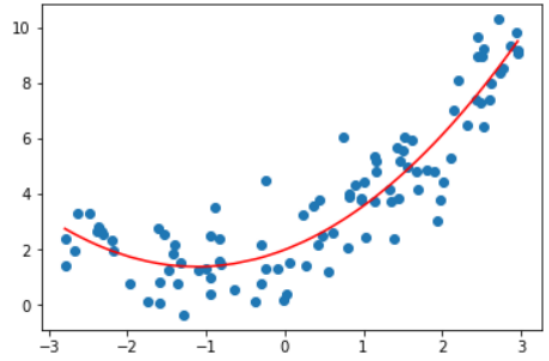
Different Types of Regression Models



- **Linear Regression** (When the dependent variable is continues and numeric with normal distribution)

- **Logistic Regression** (When the dependent variable is categorical or discrete, the logistic regression technique is applicable. In other words, this technique is used to compute the probability of mutually exclusive occurrences such as pass/fail, true/false, 0/1, and so forth. Thus, the target variable can take on only one of two values, and a sigmoid curve represents its connection to the independent variable. It is also applied to multi-class problems.)

- **Polynomial Regression** (The technique of polynomial regression analysis is used to represent a non-linear relationship between dependent and independent variables.)

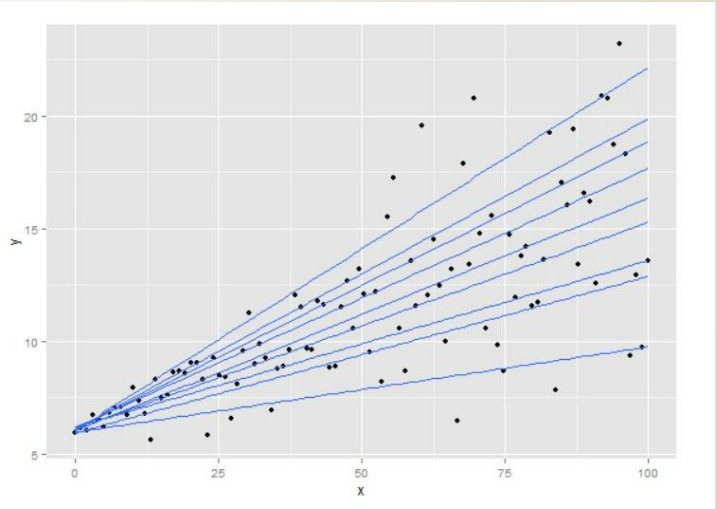


IMAGE

https://www.analyticsvidhya.com/blog/2022/01/different-types-of-regression-models/

Different Regression Models

- **Ridge Regression**
- **Lasso Regression**
- **Quantile Regression** (It is employed when the linear regression requirements are not met or when the data contains outliers)



https://www.analyticsvidhya.com/blog/2022/01/different-types-of-regression-models/

Linear Vs. Logistic Regression

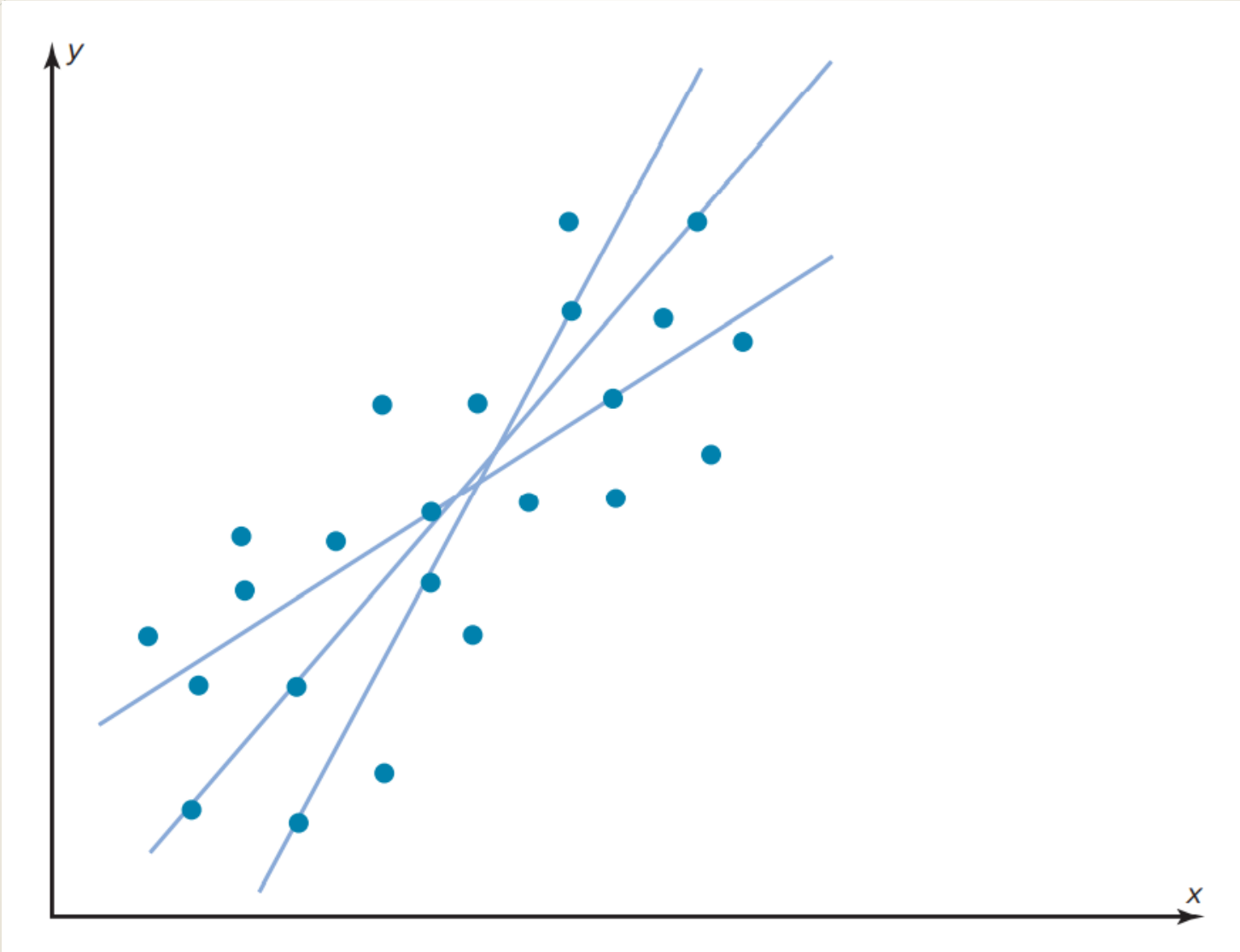
	Linear Regression	Logistic Regression
Response Variable	Continuous (e.g. price, age, height, distance)	Categorical (yes/no, male/female, win/not win)
Equation Used	$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots$	$p(Y) = e^{(\beta 0 + \beta 1 X1 + \beta 2 X2 + \ldots)} / (1 + e^{(\beta 0 + \beta 1 X1 + \beta 2 X2 + \ldots)})$
Method Used to Fit Equation	Ordinary Least Squares	Maximum Likelihood Estimation
Output to Predict	Continuous value ($150, 40 years, 10 feet, etc.)	Probability (0.741, 0.122, 0.345, etc.)

https://www.statology.org/logistic-regression-vs-linear-regression/

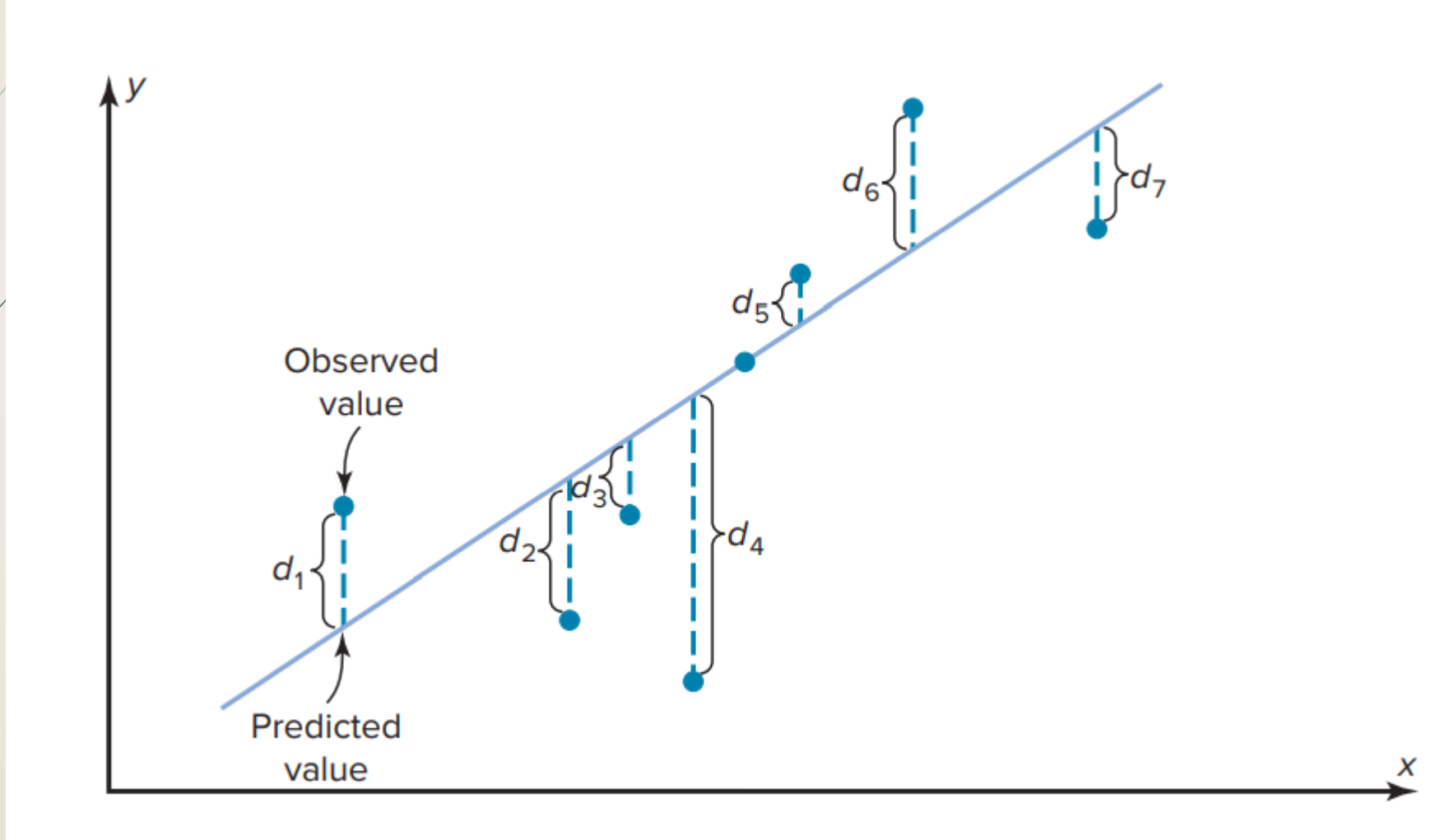
Line of Best Fit

- Next figure shows a scatter plot for the data of two variables. It shows that several lines can be drawn on the graph near the points.
- Given a scatter plot, you must be able to draw the line of best fit. Best fit means that the sum of the squares of the vertical distances from each point to the line is at a minimum.
- The difference between the actual value y and the predicted value y' (that is, the vertical distance) is called a *residual or a predicted error*. Residuals are used to determine the line that best describes the relationship between the two variables.
- The method used for making the residuals as small as possible is called **the method of least squares**. As a result of this method, **the regression line is also called the least squares regression line.**
- The reason you need a line of best fit is that the values of y will be predicted from the values of x; hence, the closer the points are to the line, the better the fit and the prediction will be. When r is positive, the line slopes upward and to the right. When r is negative, the line slopes downward from left to right.

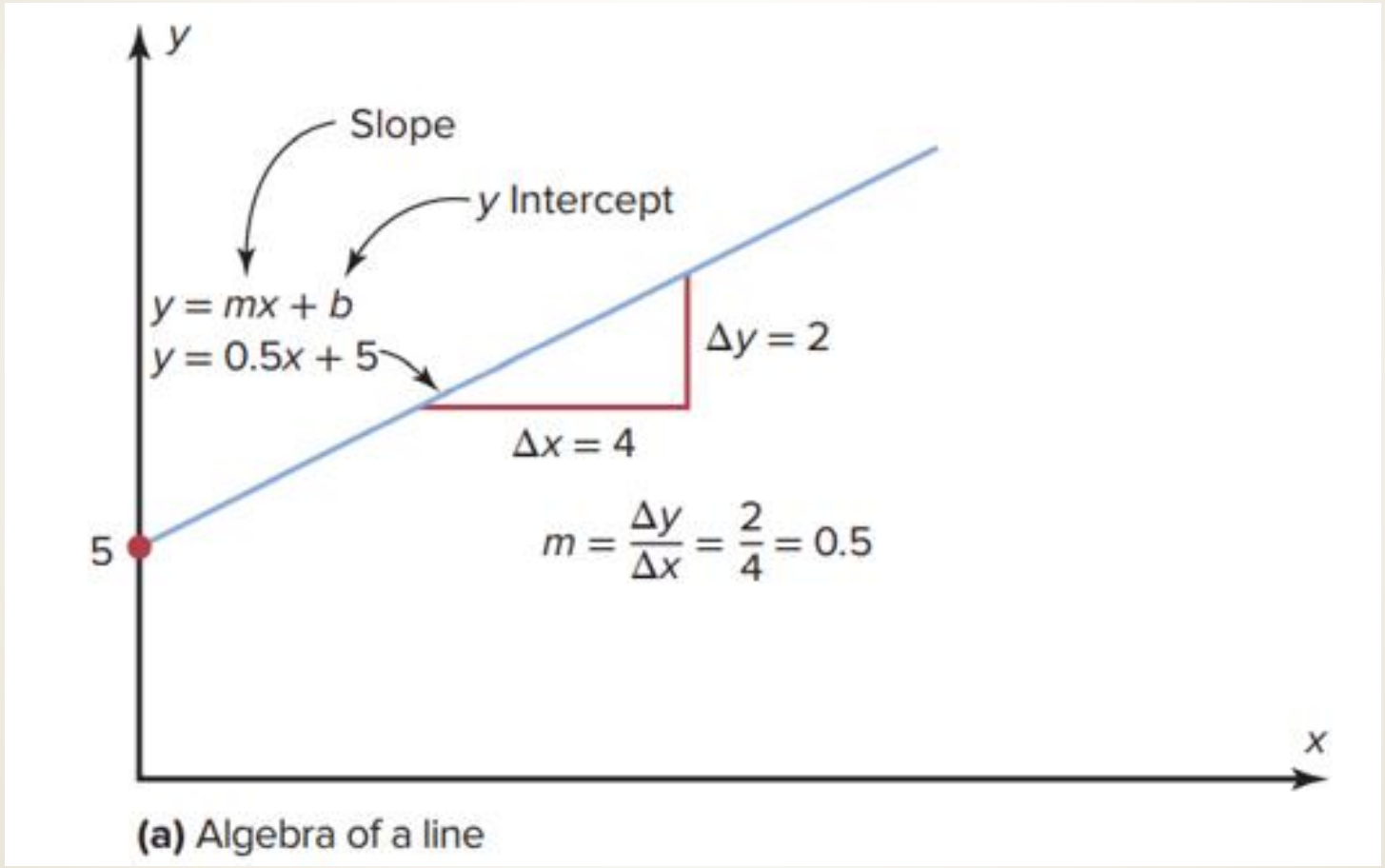
Line of Best Fit



Line of Best Fit



Line of Best Fit



(a) Algebra of a line

Determination of the Regression Line Equation

Formulas for the Regression Line $y' = a + bx$

$$a = \frac{(\Sigma y)(\Sigma x^2) - (\Sigma x)(\Sigma xy)}{n(\Sigma x^2) - (\Sigma x)^2}$$

$$b = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{n(\Sigma x^2) - (\Sigma x)^2}$$

where a is the y' intercept and b is the slope of the line.

Rounding Rule for the Intercept and Slope

Procedure Table

Finding the Regression Line Equation

Step 1 Make a table, as shown in step 2.

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

x	y	xy	x^2	y^2
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
$\Sigma x =$	$\Sigma y =$	$\Sigma xy =$	$\Sigma x^2 =$	$\Sigma y^2 =$

Step 3 When r is significant, substitute in the formulas to find the values of a and b for the regression line equation $y' = a + bx$.

$$a = \frac{(\Sigma y)(\Sigma x^2) - (\Sigma x)(\Sigma xy)}{n(\Sigma x^2) - (\Sigma x)^2} \qquad b = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{n(\Sigma x^2) - (\Sigma x)^2}$$

Example 1- Car Rental Company

Company	Cars x (in ten thousands)	Revenue y (in billions)	xy	x^2	y^2
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			

Example 1- Car Rental Company

Company	Cars x (in 10,000s)	Revenue y (in billions of dollars)	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
	$\Sigma x = 153.8$	$\Sigma y = 18.7$	$\Sigma xy = 682.77$	$\Sigma x^2 = 5859.26$	$\Sigma y^2 = 80.67$

Step 3 Substitute in the formula and solve for r.

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

$$= \frac{(6)(682.77) - (153.8)(18.7)}{\sqrt{[(6)(5859.26) - (153.8)^2][(6)(80.67) - (18.7)^2]}} = 0.982$$

Example 1- Car Rental Company

Find the equation of the regression line for the data in this example, and graph the line on the scatter plot of the data.

SOLUTION

The values needed for the equation are $n = 6$, $\Sigma x = 153.8$, $\Sigma y = 18.7$, $\Sigma xy = 682.77$, and $\Sigma x^2 = 5859.26$. Substituting in the formulas, you get

$$a = \frac{(\Sigma y)(\Sigma x^2) - (\Sigma x)(\Sigma xy)}{n(\Sigma x^2) - (\Sigma x)^2} = \frac{(18.7)(5859.26) - (153.8)(682.77)}{(6)(5859.26) - (153.8)^2} = 0.396$$

$$b = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{n(\Sigma x^2) - (\Sigma x)^2} = \frac{6(682.77) - (153.8)(18.7)}{(6)(5859.26) - (153.8)^2} = 0.106$$

Hence, the equation of the regression line $y' = a + bx$ is

$$y' = 0.396 + 0.106x$$

Example1- Car Rental Company

To graph the line, select any two points for x and find the corresponding values for y. Use any x values between 10 and 60. For example, let $x = 15$. Substitute in the equation and find the corresponding y' value.

$$\begin{aligned} y' &= 0.396 + 0.106x \\ &= 0.396 + 0.106(15) \\ &= 1.986 \end{aligned}$$

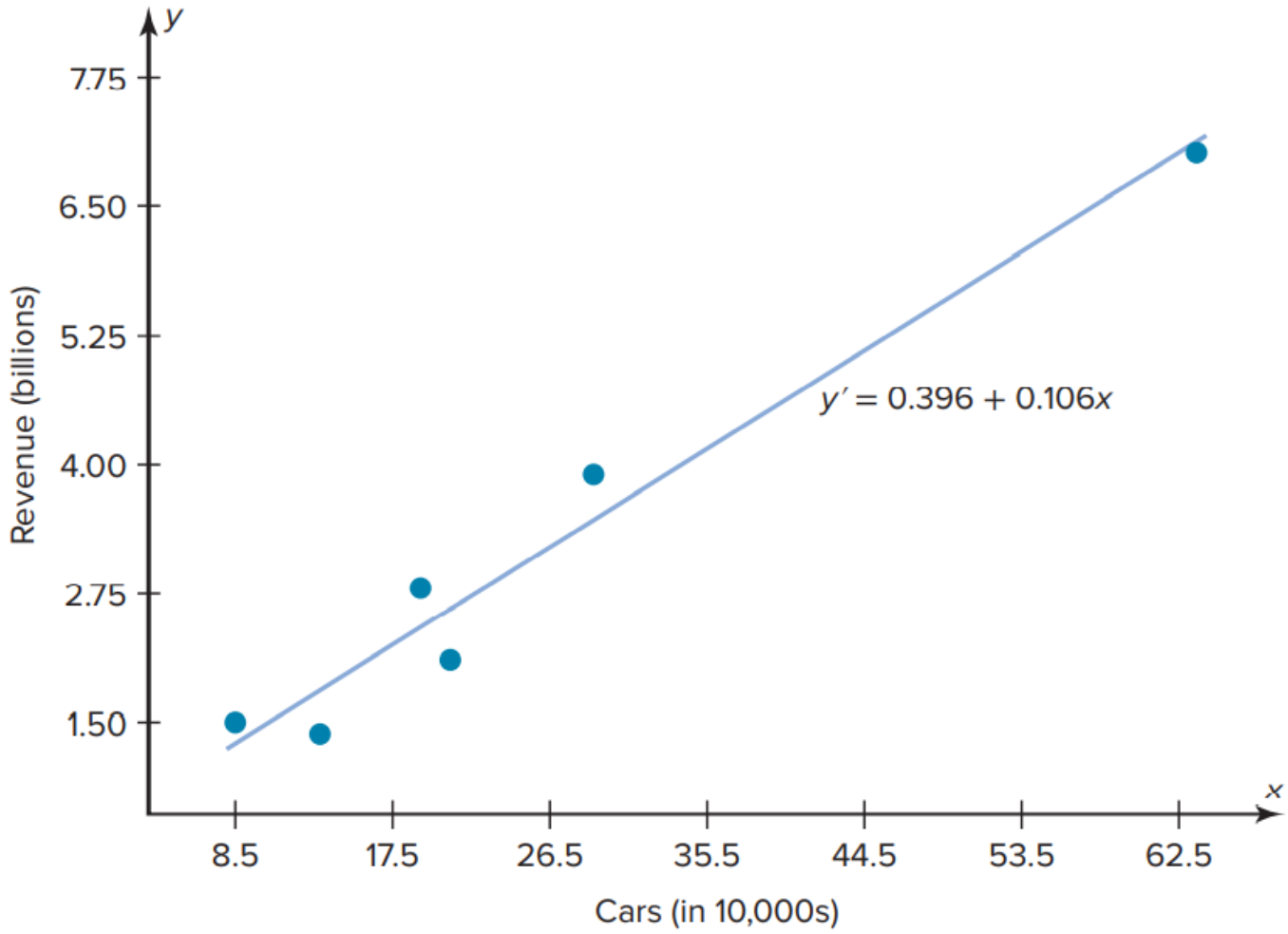
Let $x = 40$; then

$$\begin{aligned} y' &= 0.396 + 0.106x \\ &= 0.396 + 0.106(40) \\ &= 4.636 \end{aligned}$$

Then plot the two points (15, 1.986) and (40, 4.636) and draw a line connecting the two points. See Figure 10–14.

Example1- Car Rental Company

FIGURE 10–14 Regression Line for Example 10–9



Example 2- Absences and Final Grades

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
	$\Sigma x = 57$	$\Sigma y = 511$	$\Sigma xy = 3745$	$\Sigma x^2 = 579$	$\Sigma y^2 = 38{,}993$

Step 3 Substitute in the formula and solve for r.

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

$$= \frac{(7)(3745) - (57)(511)}{\sqrt{[(7)(579) - (57)^2][(7)(38{,}993) - (511)^2]}} = -0.944$$

Example 2- Absences and Final Grades

Find the equation of the regression line for the data in Example 10–5, and graph the line on the scatter plot.

SOLUTION

The values needed for the equation are $n = 7$, $\Sigma x = 57$, $\Sigma y = 511$, $\Sigma xy = 3745$, and $\Sigma x^2 = 579$. Substituting in the formulas, you get

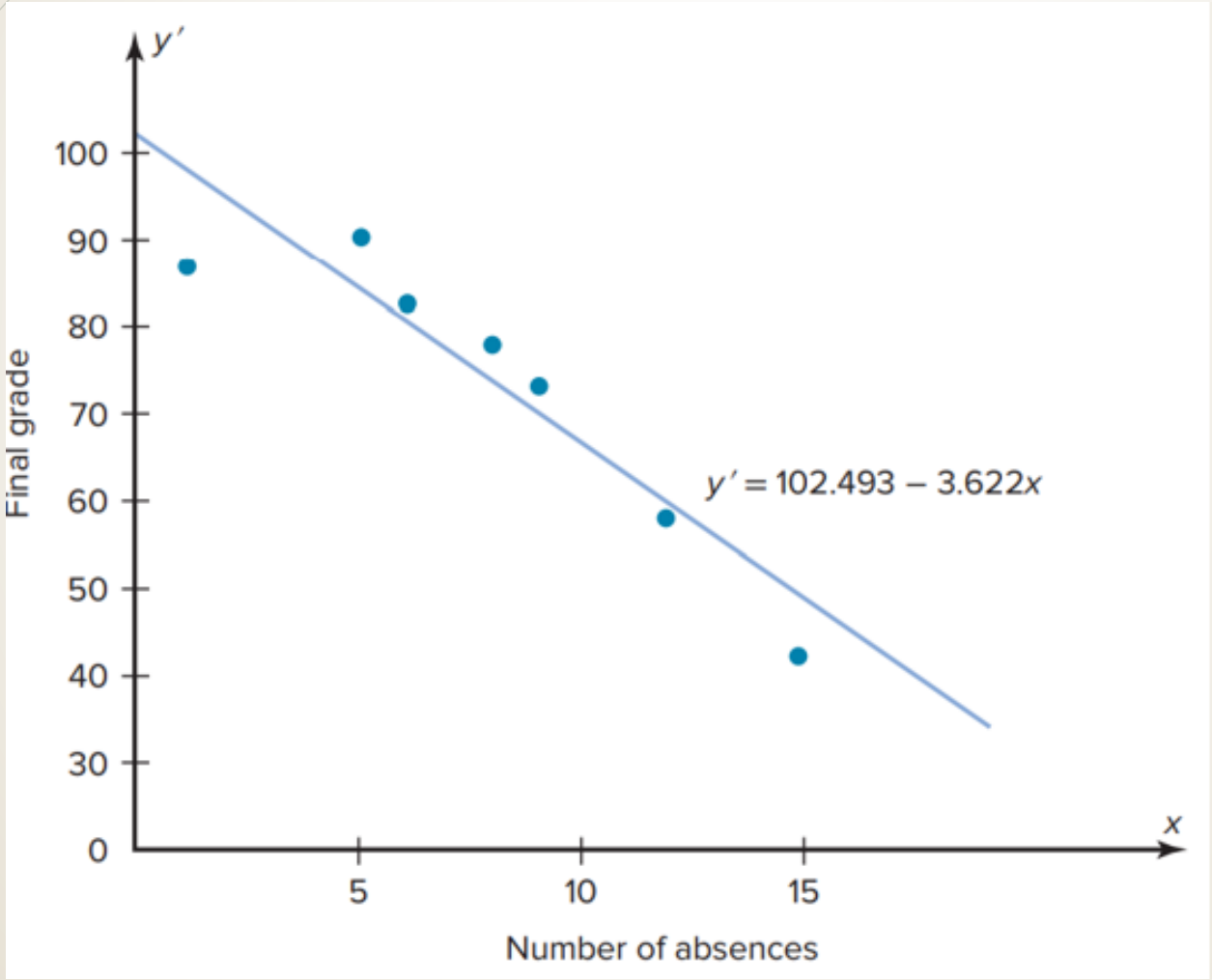
$$a = \frac{(\Sigma y)(\Sigma x^2) - (\Sigma x)(\Sigma xy)}{n(\Sigma x^2) - (\Sigma x)^2} = \frac{(511)(579) - (57)(3745)}{(7)(579) - (57)^2} = 102.493$$

$$b = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{n(\Sigma x^2) - (\Sigma x)^2} = \frac{(7)(3745) - (57)(511)}{(7)(579) - (57)^2} = -3.622$$

Hence, the equation of the regression line $y' = a + bx$ is

$$y' = 102.493 - 3.622x$$

Example 2- Absences and Final Grades



Example 2- Absences and Final Grades

- The sign of the correlation coefficient and the sign of the slope of the regression line will always be the same. That is, if *r* is positive, then *b* will be positive; if *r* is negative, then *b* will be negative.
- The reason is that the numerators of the formulas are the same and determine the signs of *r* and *b*, and the denominators are always positive.
- The regression line will always pass through the point whose *x* coordinate is the mean of the *x* values and whose *y* coordinate is the mean of the *y* values.
- The regression line can be used to make predictions for the dependent variable.

Example2- Absences and Final Grades

Use the equation of the regression line in Example 10–10 to predict the final grade for a student who missed 4 classes.

SOLUTION

Substitute 4 for x in the regression line equation $y' = 102.493 - 3.622x$.

$$\begin{aligned} y' &= 102.493 - 3.622x \\ &= 102.493 - 3.622(4) \\ &= 88.005 \\ &= 88 \text{ (rounded)} \end{aligned}$$

Hence, when a student misses 4 classes, the student's grade on the final exam is predicted to be about 88.

Outliers

- A scatter plot should be checked for outliers. An outlier is a point that seems out of place when compared with the other points.
- Some of these points can affect the equation of the regression line. When this happens, the points are called **influential points** or **influential observations**. When a point on the scatter plot appears to be an outlier, it should be checked to see if it is an influential point.
- An influential point tends to "**pull**" the regression line toward the point itself. To check for an influential point, the regression line should be graphed with the point included in the data set.
- Then a second regression line should be graphed that **excludes the point** from the data set. If the position of the second line is changed considerably, the point is said to be an influential point. Points that are outliers in the *x* direction tend to be influential points.

Extrapolation

- Extrapolation, or making predictions beyond the bounds of the data, must be interpreted cautiously.
- For example, in 1979, some experts predicted that the United States would run out of oil by the year 2003. This prediction was based on the current consumption and on known oil reserves at that time. However, since then, the automobile industry has produced many new fuel-efficient vehicles.
- Also, there are many as yet undiscovered oil fields. Finally, science may someday discover a way to run a car on something as unlikely but as common as peanut oil. In addition, the price of a gallon of gasoline was predicted to reach $10 a few years later.
- Fortunately this has not come to pass. 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.

References

- ✓ Data Mining, Concepts and Techniques, Jiawei Han, Micheline Kamber, Jian Pei. MK. Chapter 9.
- ✓ Elementary Statistics: A Step-by-Step Approach, Allen Bluman, 10th Edition, McGraw Hill, 2017, ISBN 13: 978-1-259-755330, Chapter 10.