BITS Pilani

Pilani Campus

Linear Regression

Akanksha Bharadwaj
Asst. Professor, CS/IS Department

BITS Pilani
Pilani Campus

SS ZG536, ADV STAT TECHNIQUES FOR ANALYTICS
Contact Session 8

Covariance

- Variables may change in relation to each other
- *Covariance* measures how much the movement in one variable predicts the movement in a corresponding variable
- “Covariance” indicates the **direction** of the linear relationship between variables.

Variance Vs Covariance

Variance:

- Gives information on variability of a single variable.

$$S_x^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}$$

Covariance:

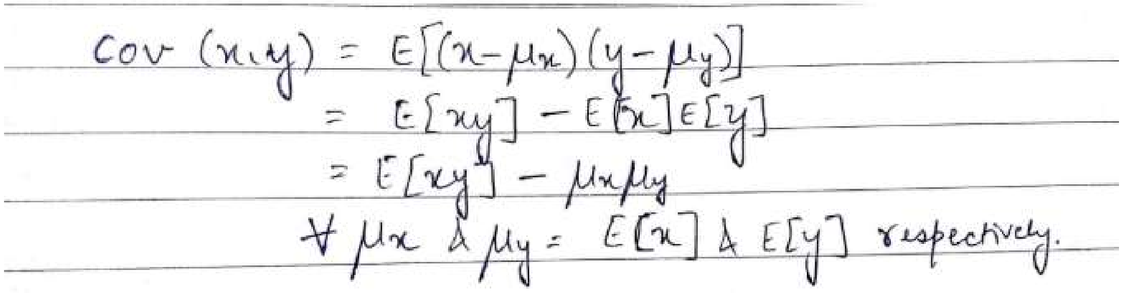
- Gives information on the degree to which two variables vary together.

$$\operatorname{cov}(x, y) = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{n-1}$$

Covariance

$$\operatorname{cov}(x, y) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

- E[x] is the expected value or mean of a sample ‘x’, then cov(x,y) can be represented in the following way:



$$\operatorname{cov}(x,y) = E[(x-\mu_x)(y-\mu_y)]$$

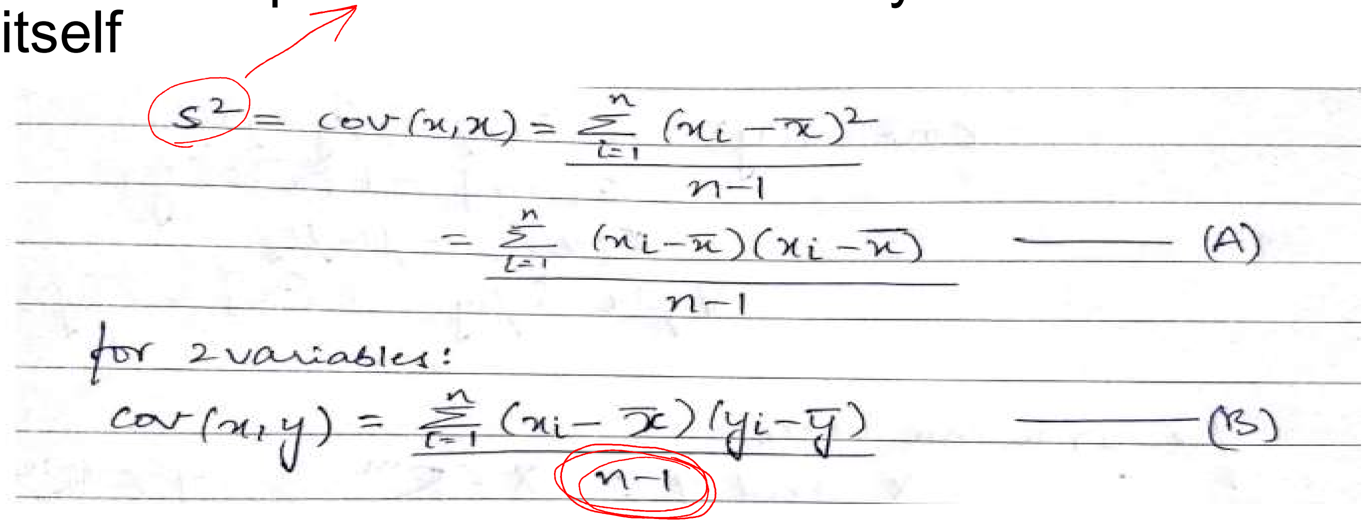
$$= E[xy] - E[x]E[y]$$

$$= E[xy] - \mu_x\mu_y$$

$\forall\ \mu_x$ & μ_y = $E[x]$ & $E[y]$ respectively.

Sampled variance

- ‘s²’ or sampled variance is basically the covariance of a variable with itself



$$\boxed{s^2} = cov(x,x) = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}$$

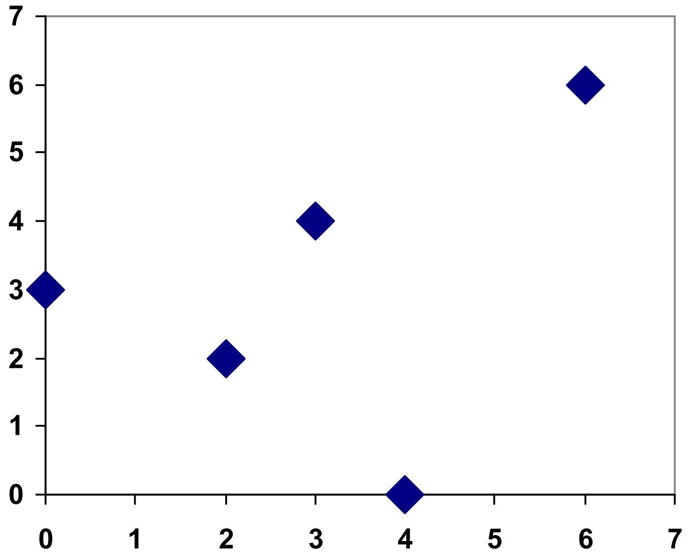
$$= \frac{\sum_{i=1}^{n}(x_i-\bar{x})(x_i-\bar{x})}{n-1} \quad \text{———} \quad (A)$$

for 2 variables:

$$cov(x,y) = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\boxed{n-1}} \quad \text{———} \quad (B)$$

- In the above formula, the numerator of the equation(A) is called the **sum of squared deviations.**
- In equation(B) with two variables x and y, it is called the **sum of cross products**.
- In the above formula, n is the number of samples in the data set. The value (n-1) indicates the degrees of freedom.

Example Covariance



x	y	$x_i-\bar{x}$	$y_i-\bar{y}$	$(x_i-\bar{x})(y_i-\bar{y})$
0	3	-3	0	0
2	2	-1	-1	1
3	4	0	1	0
4	0	1	-3	-3
6	6	3	3	9
$\bar{x}=3$	$\bar{y}=3$			$\sum=7$

$n=5$

$$\text{cov}(x,y)=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}))}{n-1}=\frac{7}{4}=\boxed{1.75}$$

What does this number tell us?

If both variables tend to increase or decrease together, the coefficient is positive.

Problem with Covariance:

- The value obtained by covariance is dependent on the size of the data’s standard deviations: if large, the value will be greater.

- Even if the relationship between x and y is exactly the same in the large versus small standard deviation datasets.

Example of how covariance value relies on variance

	High variance data				Low variance data		
Subject	**x**	**y**	**x error * y error**		**x**	**y**	**X error * y error**
1	101	100	2500		54	53	9
2	81	80	900		53	52	4
3	61	60	100		52	51	1
4	51	50	0		51	50	0
5	41	40	100		50	49	1
6	21	20	900		49	48	4
7	1	0	2500		48	47	9
Mean	51	50			51	50	
Sum of x error * y error :			7000		**Sum of x error * y error :**		28
Covariance:			**1166.67**		**Covariance:**		**4.67**

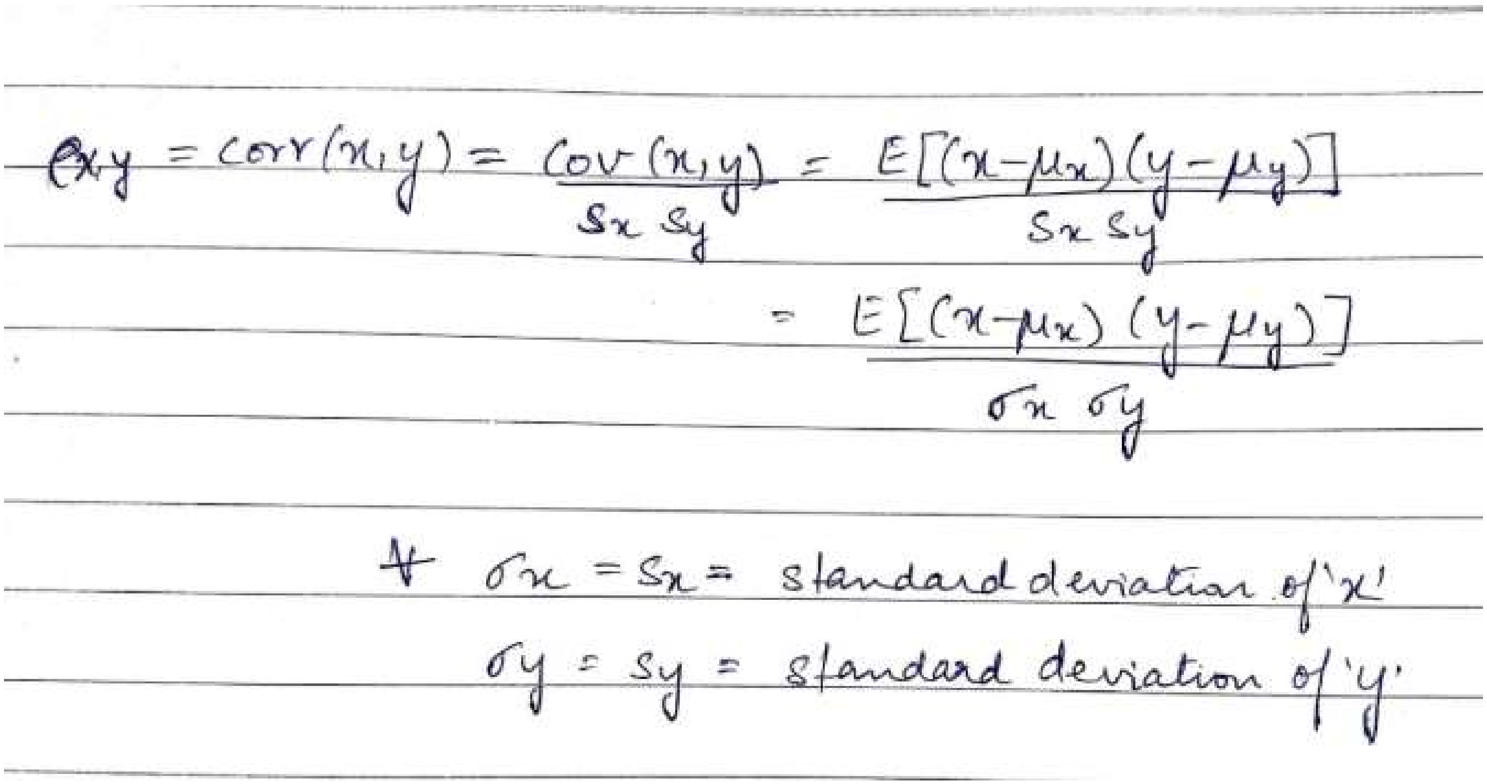
Correlation

- It is *a measure of* ***the degree of relatedness*** *of variables.* It can help a business researcher determine, for example, whether the stocks of two airlines rise and fall in any related manner.

- For a sample of pairs of data, correlation analysis can yield a numerical value that represents the degree of relatedness of the two stock prices over time.

- it is obtained by dividing the covariance of the two variables by the product of their standard deviations

Mathematical representation



$$\rho_{x,y} = \text{corr}(x,y) = \frac{\text{cov}(x,y)}{S_x S_y} = \frac{E[(x-\mu_x)(y-\mu_y)]}{S_x S_y}$$

$$= \frac{E[(x-\mu_x)(y-\mu_y)]}{\sigma_x \sigma_y}$$

$\forall$ $\sigma_x = S_x =$ standard deviation of 'x'

$\sigma_y = S_y =$ standard deviation of 'y'

Questions a Pearson correlation answers

- Is there a statistically significant relationship between age and height?

- Is there a relationship between temperature and ice cream sales?

- Is there a relationship among job satisfaction, productivity, and income?

- Which two variable have the strongest co-relation between age, height, weight, size of family and family income?

Assumptions

- For the Pearson r correlation, both variables should be **normally distributed**.

- There should be **no significant outliers**.

- Each variable should be **continuous**

- The two variables have a **linear relationship**.

- The observations are **paired observations.** That is, for every observation of the independent variable, there must be a corresponding observation of the dependent variable.

Pearson product-moment correlation coefficient

- Researchers virtually always deal with sample data, this section introduces a widely used sample **coefficient of correlation**, *r*.

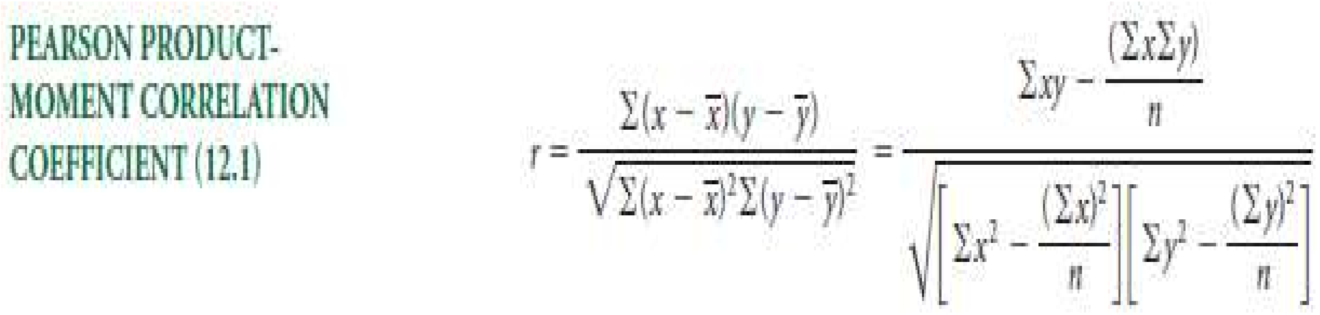
- The term *r* is a *measure of the linear correlation of two variables.* It is a number that ranges from -1 to 0 to +1, representing the strength of the relationship between the variables.

- An *r* value of +1 denotes a perfect positive relationship between two sets of numbers.

- An *r* value of -1 denotes a perfect negative correlation, which indicates an inverse relationship between two variables: as one variable gets larger, the other gets smaller.

- An *r* value of 0 means no linear relationship is present between the two variables.

Formula

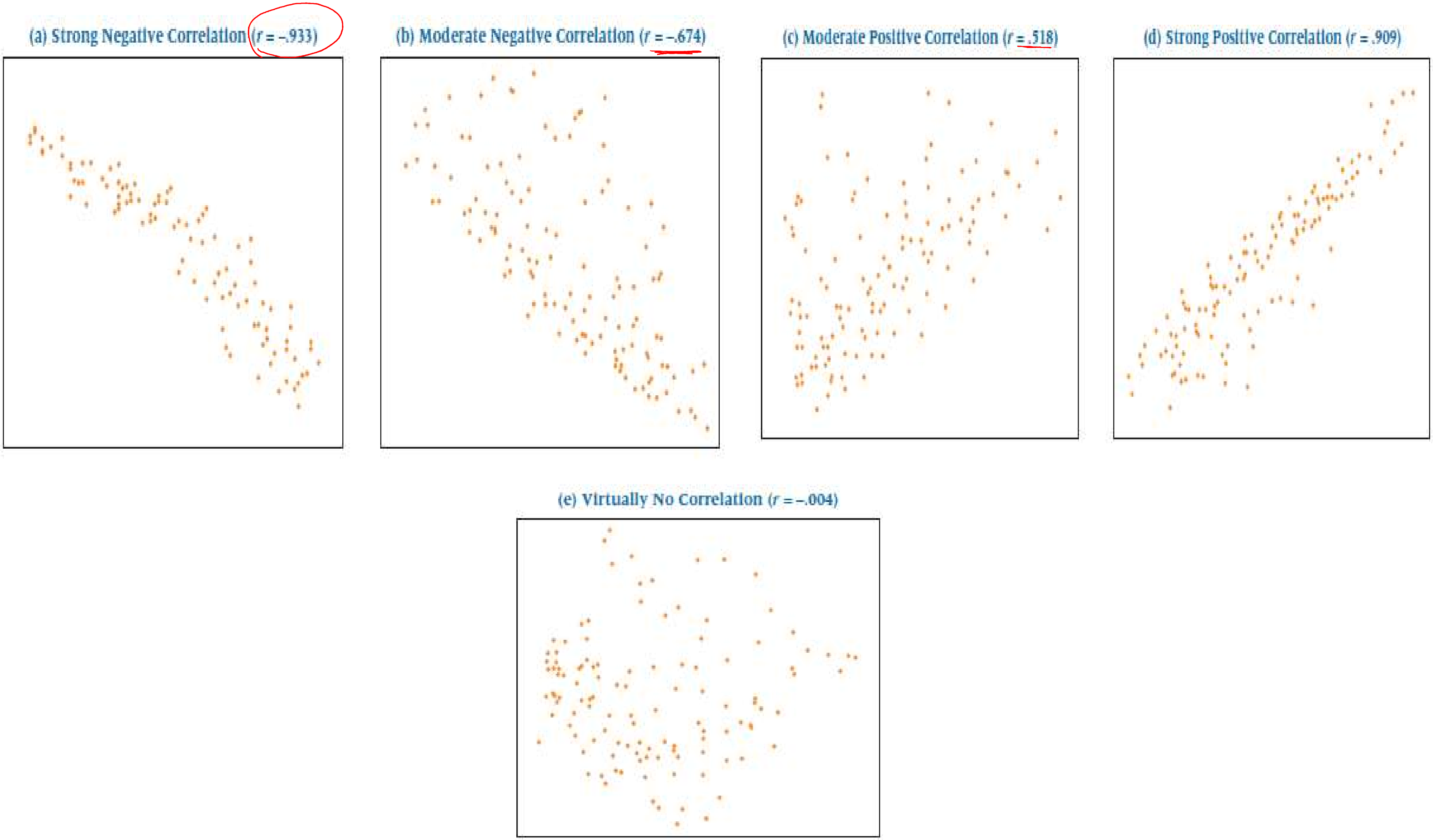


PEARSON PRODUCT-MOMENT CORRELATION COEFFICIENT (12.1)

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

- The closer it is to +1 or -1, the more closely are the two variables are related.

- The positive sign signifies the direction of the correlation i.e. if one of the variables increases, the other variable is also supposed to increase.



(a) Strong Negative Correlation ($r = -.933$)

(b) Moderate Negative Correlation ($r = -.674$)

(c) Moderate Positive Correlation ($r = .518$)

(d) Strong Positive Correlation ($r = .909$)

(e) Virtually No Correlation ($r = -.004$)

Correlation does not have units but Covariance always has units

- As we see from the formula of covariance, it assumes the units from the product of the units of the two variables.
- On the other hand, correlation is dimensionless. It is a unit-free measure of the relationship between variables.
- This is because we divide the value of covariance by the product of standard deviations which have the same units.
- The value of covariance is affected by the change in scale of the variables.
- However, on doing the same, the value of correlation is not influenced by the change in scale of the values.

Advantages of the Correlation Coefficient

The Correlation Coefficient has several advantages over covariance for determining strengths of relationships:

- While correlation coefficients lie between -1 and +1, covariance can take any value between -∞ and +∞.
- Because of it’s numerical limitations, correlation is more useful for determining **how strong** the relationship is between the two variables.
- Correlation isn’t affected by changes in the center (i.e. mean) or scale of the variables

Exercise

What is the measure of correlation between the interest rate of federal funds and the commodities futures index? With data such as those shown in Table 12.1, which represent the values for interest rates of federal funds and commodities futures indexes for a sample of 12 days.

$$\mathrm{cov}(x,y) = \sum_{i=1}^{n} \frac{(x_i-\bar{x})\ (y_i-\bar{y})}{(n-1)}$$

$$\mathrm{cor} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})\ (y_i-\bar{y})}{\sqrt{\Sigma(x-\bar{x})^2\ \Sigma(y-\bar{y})^2}}$$

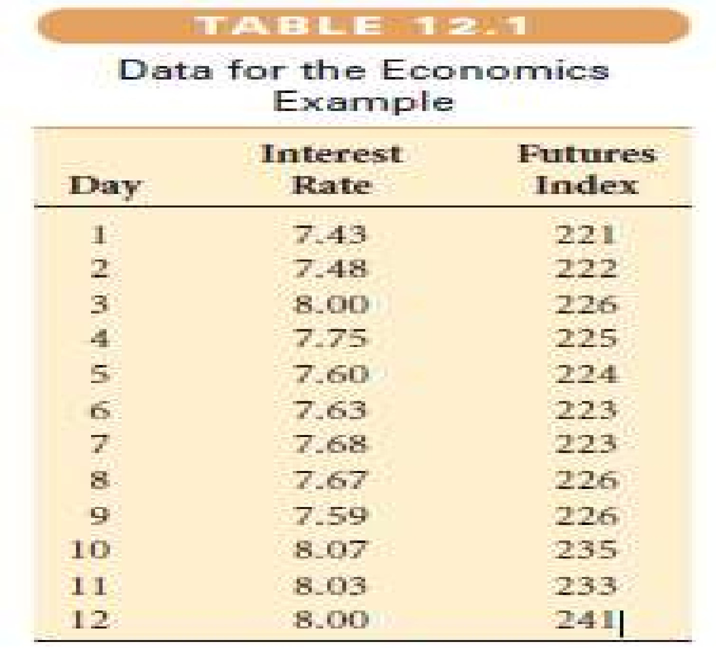
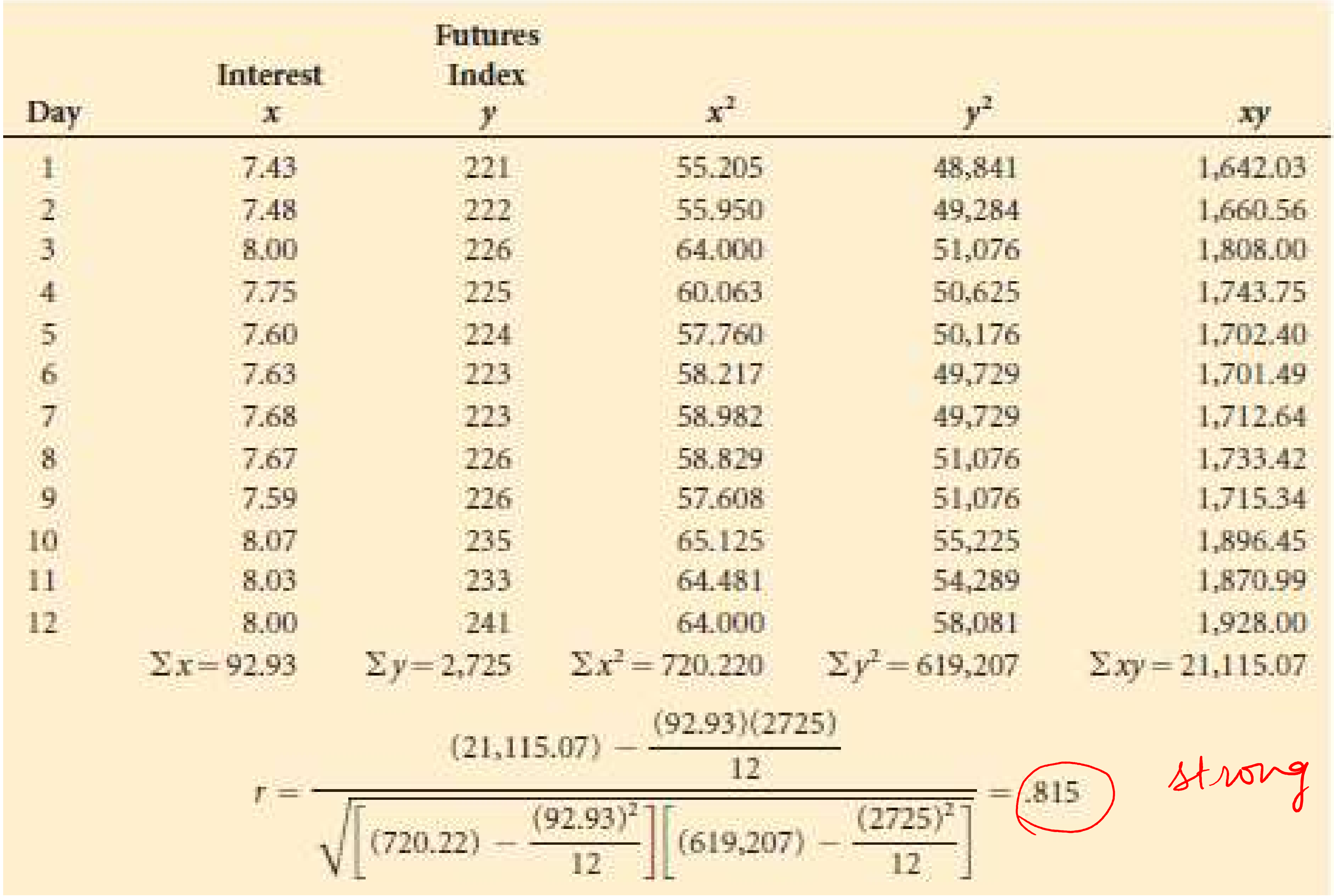


TABLE 12.1

Data for the Economics Example

Day	Interest Rate	Futures Index
1	7.43	221
2	7.48	222
3	8.00	226
4	7.75	225
5	7.60	224
6	7.63	223
7	7.68	223
8	7.67	226
9	7.59	226
10	8.07	235
11	8.03	233
12	8.00	241



Day	Interest x	Futures Index y	x^2	y^2	xy
1	7.43	221	55.205	48,841	1,642.03
2	7.48	222	55.950	49,284	1,660.56
3	8.00	226	64.000	51,076	1,808.00
4	7.75	225	60.063	50,625	1,743.75
5	7.60	224	57.760	50,176	1,702.40
6	7.63	223	58.217	49,729	1,701.49
7	7.68	223	58.982	49,729	1,712.64
8	7.67	226	58.829	51,076	1,733.42
9	7.59	226	57.608	51,076	1,715.34
10	8.07	235	65.125	55,225	1,896.45
11	8.03	233	64.481	54,289	1,870.99
12	8.00	241	64.000	58,081	1,928.00
	$\Sigma x = 92.93$	$\Sigma y = 2{,}725$	$\Sigma x^2 = 720.220$	$\Sigma y^2 = 619{,}207$	$\Sigma xy = 21{,}115.07$

$$r = \frac{(21{,}115.07) - \dfrac{(92.93)(2725)}{12}}{\sqrt{\left[(720.22) - \dfrac{(92.93)^2}{12}\right]\left[(619{,}207) - \dfrac{(2725)^2}{12}\right]}} = .815$$

strong + relationship

Correlation Vs Covariance

- In simple words, both the terms measure the relationship and the dependency between two variables.
- “Covariance” indicates the direction of the linear relationship between variables.
- “Correlation” on the other hand measures both the strength and direction of the linear relationship between two variables.
- Correlation is a function of the covariance.
- **What sets them apart is the fact that correlation values are standardized whereas, covariance values are not.**

Regression

- **Correlation** tells you if there is an association between x and y but it **doesn’t describe the relationship** or allow you to predict one variable from the other.

- To **do this we need REGRESSION**!

Regression Analysis

- **Regression analysis** is *the process of constructing a mathematical model or function that can be used to predict or determine one variable by another variable or other variables*.
- The most elementary regression model is called **simple regression** or **bivariate regression** involving two variables in which one variable is predicted by another variable.

Regression Analysis

- Regression analysis is a way of mathematically sorting out which of those variables does indeed have an impact.

It answers the questions:

- Which factors matter most?
- Which can we ignore?
- How do those factors interact with each other?
- And, perhaps most importantly, how certain are we about all of these factors?

In regression analysis, those **factors** are called **variables**.

Dependent and independent variable

- In simple regression, *the variable to be predicted* is called the **dependent variable** and is designated as *y*.

- The *predictor* is called the **independent variable**, or *explanatory variable*, and is designated as *x*.

- In simple regression analysis, only a straight-line relationship between two variables is examined.

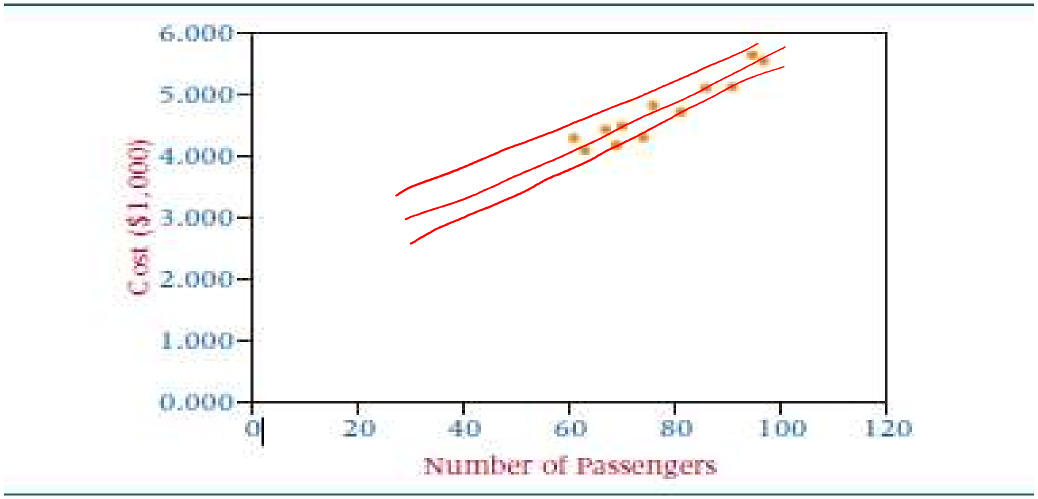
- Nonlinear relationships and regression models with more than one independent variable can be explored by using multiple regression models

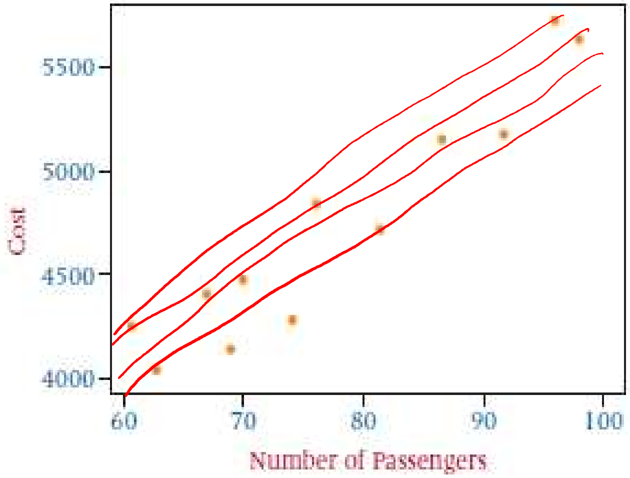
Scatter Plot

- Usually, the first step in simple regression analysis is to construct a **scatter plot**

- Graphing the data in this way yields preliminary information about the shape and spread of the data

Scatter Plot





- Try to imagine a line passing through the points. Is a linear fit possible? Would a curve fit the data better?

innovate achieve lead

BITS Pilani

Pilani Campus

Simple Linear Regression

Equation of regression line

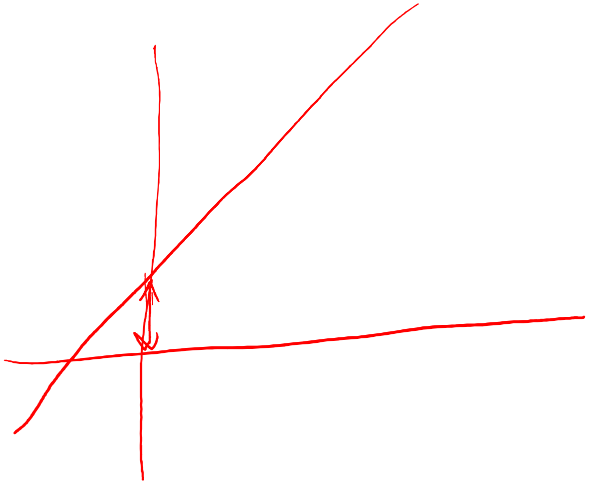
- In math courses, the slope-intercept form of the equation of a line often takes the form

$$y = mx + b$$

where

m = slope of the line

b = y intercept of the line



Equation of regression line

$$\underbrace{Y_i}_{\text{dependent variable}} = \underbrace{\beta_0}_{\text{population y intercept}} + \overbrace{\beta_1}^{\text{population slope}} \underbrace{x_i}_{\text{independent variable}} + \underbrace{\varepsilon_i}_{\text{error of prediction}}$$

- Unless the points being fitted by the regression equation are in perfect alignment, the regression line will miss at least some of the points.

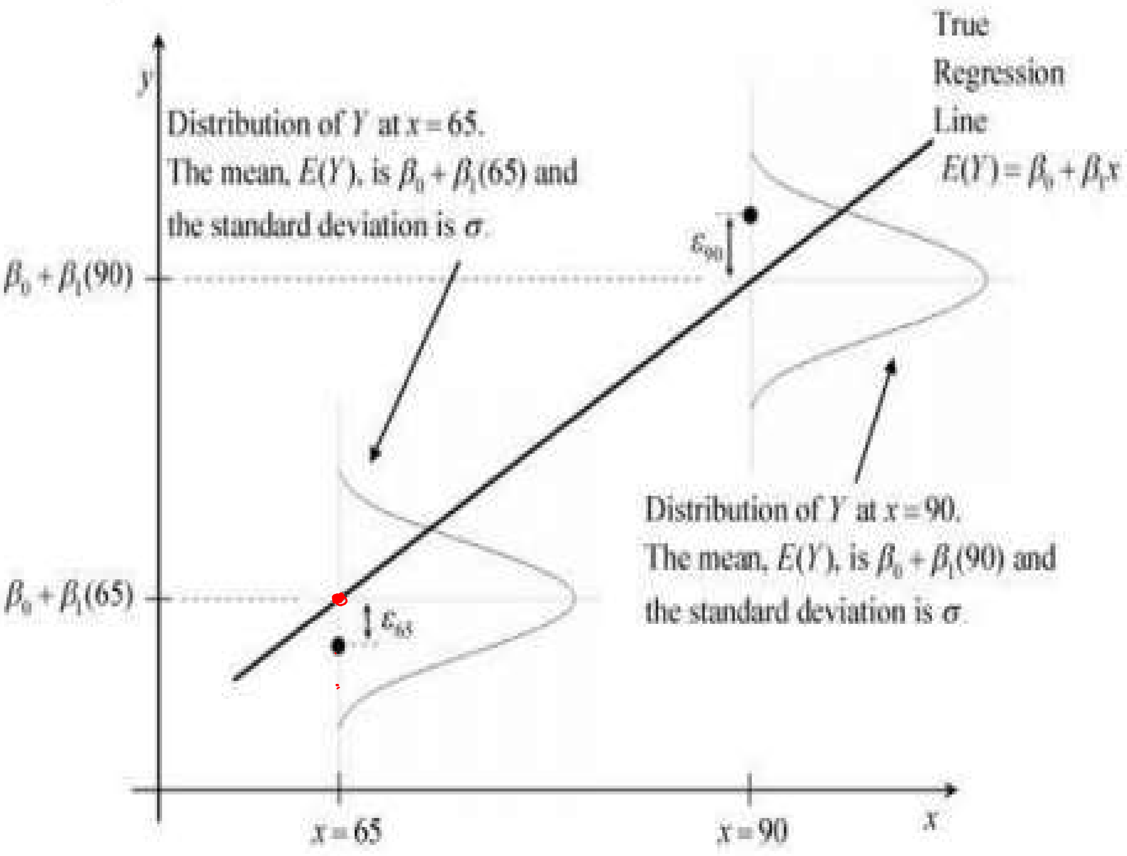
- In the preceding equation, Ɛ represents the error of the regression line in fitting these points. Sum of error $\approx$ zero

- If a point is on the regression line the value of error will be zero

Assumptions about the Error

- *Summation of all the error terms E(*ε_i) almost equals to 0.
- $\sigma(\varepsilon_i) = \sigma_\varepsilon$ where σ_ε is unknown.
- The errors are independent, that is, the error in the *i*th observation is independent of the error observed in the *j*th observation.
- The ε_i are normally distributed (with mean 0 and standard deviation σ_ε).

Assumptions



y
Distribution of Y at $x = 65$.
The mean, $E(Y)$, is $\beta_0 + \beta_1(65)$ and
the standard deviation is σ.
True
Regression
Line
$E(Y) = \beta_0 + \beta_1 x$
$\beta_0 + \beta_1(90)$
ε_{90}
$\beta_0 + \beta_1(65)$
ε_{65}
Distribution of Y at $x = 90$.
The mean, $E(Y)$, is $\beta_0 + \beta_1(90)$ and
the standard deviation is σ.
$x = 65$
$x = 90$
x

Estimated Regression model

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

dependent variable

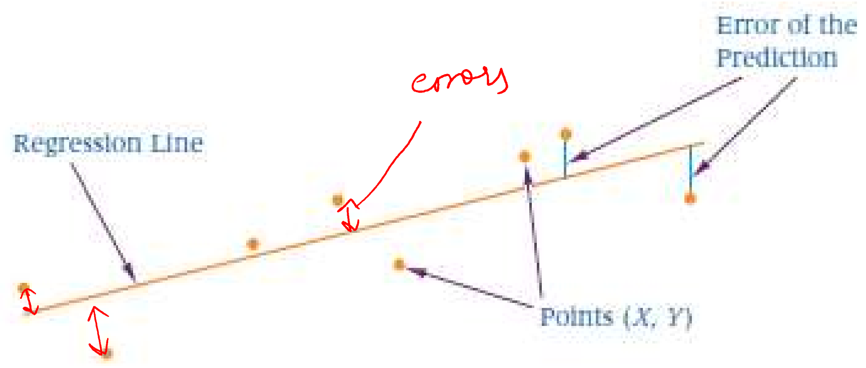
sample intercept

sample slope

- Sample Regression line provides estimate of population regression line

- To determine the equation of the regression line for a sample of data, the researcher must determine the values for b_0 and b_1.

Least Squares regression line



- Observe that the line does not actually pass through any of the points. The vertical distance from each point to the line is the **error** of the prediction.
- In theory, an infinite number of lines could be constructed to pass through these points in some manner.
- The **least squares regression line** is the regression line that results in the smallest sum of errors squared.

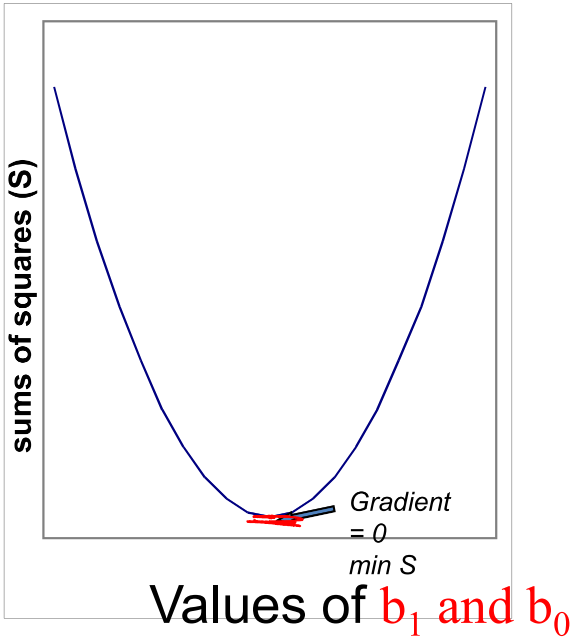
Minimising sums of squares

- Need to minimise $\Sigma(\overset{\text{actual}}{y}–\underset{\text{predicted value}}{\hat{y}})^2$
- $\hat{y} = b_1x + b_0$

- so need to minimise:$\Sigma(y - b_1x + b_0)^2$

- If we plot the sums of squares for all
- different values of b_1 and b_0 we get a parabola,
- because it is a squared term

- So the min sum of squares is at the
- bottom of the curve, where the gradient
- is zero.



Slope of the regression line

- Formula 12.2 is an equation for computing the value of the sample slope.

SLOPE OF THE REGRESSION LINE (12.2)

$$b_1 = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sum(x-\bar{x})^2} = \frac{\sum xy - n\bar{x}\bar{y}}{\sum x^2 - n\bar{x}^2} = \frac{\sum xy - \frac{(\sum x)(\sum y)}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}}$$

$$\underline{SS_{xy}} = \sum(x-\bar{x})(y-\bar{y}) = \sum xy - \frac{(\sum x)(\sum y)}{n}$$

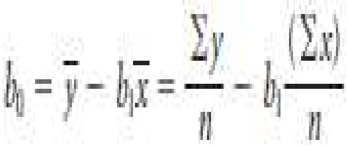
$$\underline{SS_{xx}} = \sum(x-\bar{x})^2 = \sum x^2 - \frac{(\sum x)^2}{n}$$

ALTERNATIVE FORMULA FOR SLOPE (12.3)

$$\boxed{b_1 = \frac{SS_{xy}}{SS_{xx}}}$$

Y intercept of the regression line

y INTERCEPT OF THE REGRESSION LINE (12.4)



$$b_0 = \bar{y} - b_1\bar{x} = \frac{\sum y}{n} - b_1\frac{(\sum x)}{n}$$

Exercise

What is the Best-fit Line for this data?

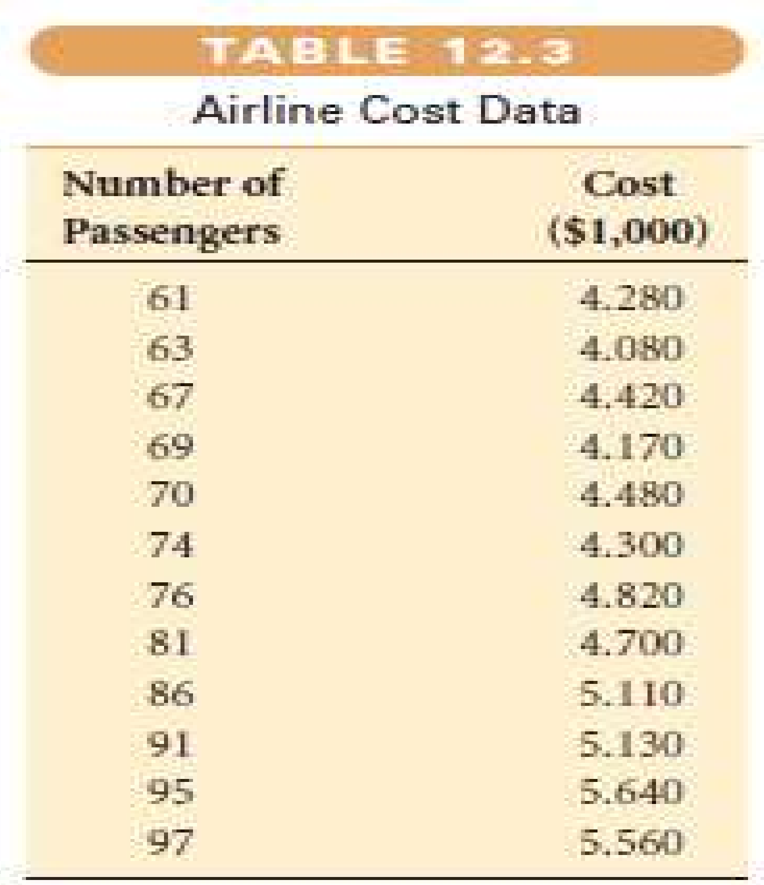


TABLE 12.3

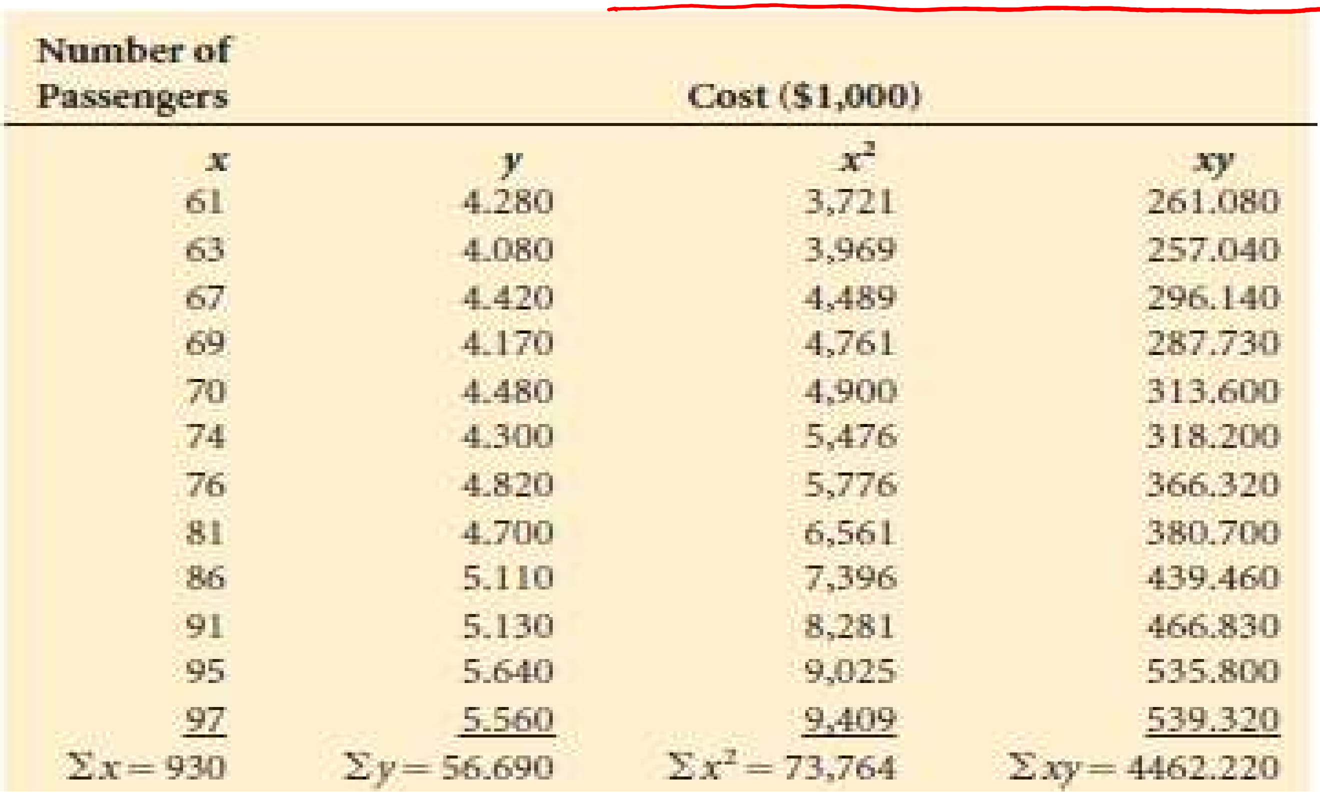
Airline Cost Data

Number of Passengers	Cost ($1,000)
61	4.280
63	4.080
67	4.420
69	4.170
70	4.480
74	4.300
76	4.820
81	4.700
86	5.110
91	5.130
95	5.640
97	5.560

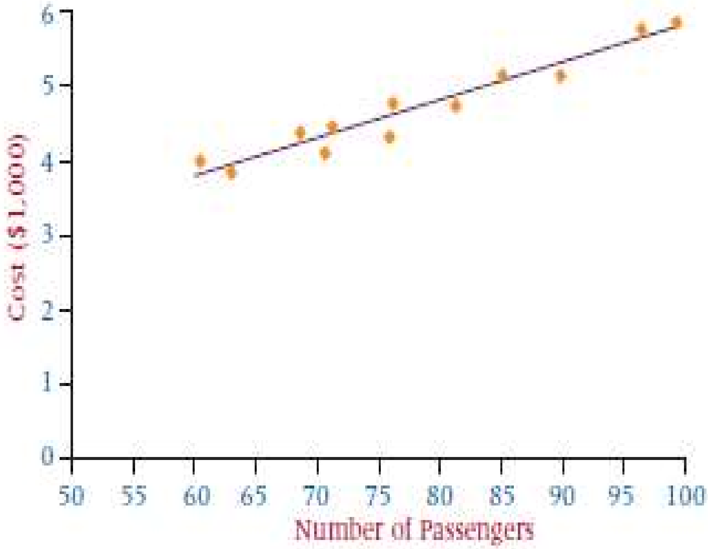
Solution

$$b_0 = \frac{\Sigma y}{n} - b_1 \frac{\Sigma x}{n} = 1.569$$

$$\boxed{\hat{y} = 1.569 + 0.0407x}$$



Number of Passengers	Cost ($1,000)		
x	y	x^2	xy
61	4.280	3,721	261.080
63	4.080	3,969	257.040
67	4.420	4,489	296.140
69	4.170	4,761	287.730
70	4.480	4,900	313.600
74	4.300	5,476	318.200
76	4.820	5,776	366.320
81	4.700	6,561	380.700
86	5.110	7,396	439.460
91	5.130	8,281	466.830
95	5.640	9,025	535.800
97	5.560	9,409	539.320
$\Sigma x = 930$	$\Sigma y = 56.690$	$\Sigma x^2 = 73{,}764$	$\Sigma xy = 4462.220$



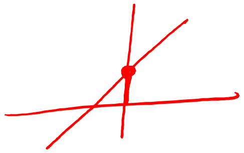
$$S_{xy} = \Sigma xy - \frac{(\Sigma x)(\Sigma y)}{n} = 4462.220 - \frac{(930)(56.690)}{12} = 68.745$$

$$S_{xx} = \Sigma x^2 - \frac{(\Sigma x)^2}{n} = 73{,}764 - \frac{(930)^2}{12} = 1689$$

$$b_1 = \frac{S_{xy}}{S_{xx}} = 0.0407$$

Interpretation of slope and intercept

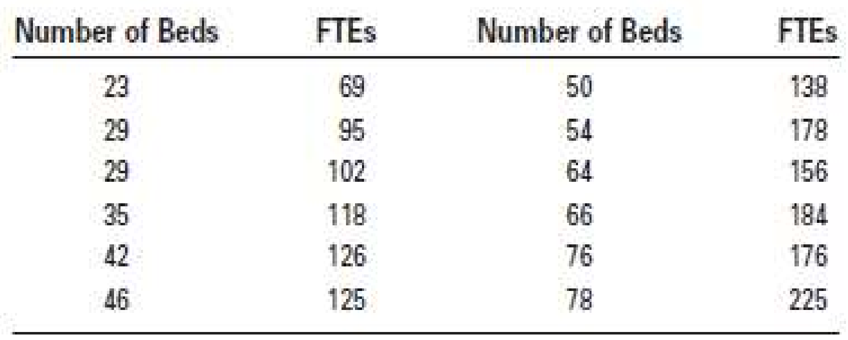
- b_1 estimated change in the average value of y as a result of unit change in x



- b_0 estimated average value of y when the value of x is 0

Exercise (HW)

- A specialist in hospital administration stated that the number of FTEs (full-time employees) in a hospital can be estimated by counting the number of beds in the hospital (a common measure of hospital size). A healthcare business researcher decided to develop a regression model in an attempt to predict the number of FTEs of a hospital by the number of beds. She surveyed 12 hospitals and obtained the following data. The data are presented in sequence, according to the number of beds.



Number of Beds	FTEs	Number of Beds	FTEs
23	69	50	138
29	95	54	178
29	102	64	156
35	118	66	184
42	126	76	176
46	125	78	225

$\Sigma x = 592$

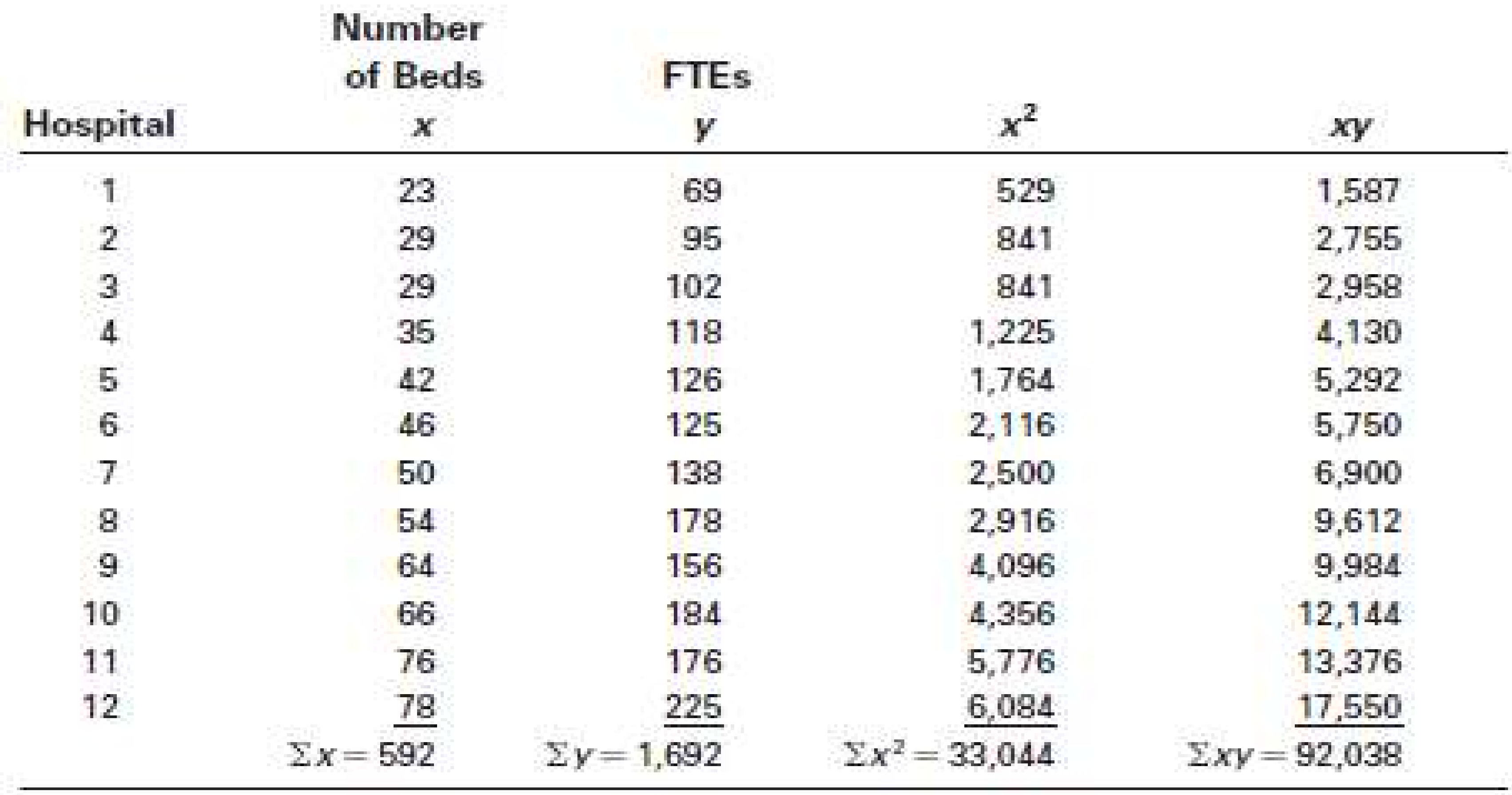
$\Sigma y = 1692$

$\Sigma x^2 = 33044$

$\Sigma xy = 92,038$

SS_{xy}? SS_{xx} ?

Solution



Hospital	Number of Beds x	FTEs y	x^2	xy
1	23	69	529	1,587
2	29	95	841	2,755
3	29	102	841	2,958
4	35	118	1,225	4,130
5	42	126	1,764	5,292
6	46	125	2,116	5,750
7	50	138	2,500	6,900
8	54	178	2,916	9,612
9	64	156	4,096	9,984
10	66	184	4,356	12,144
11	76	176	5,776	13,376
12	78	225	6,084	17,550
	$\Sigma x = 592$	$\Sigma y = 1,692$	$\Sigma x^2 = 33,044$	$\Sigma xy = 92,038$

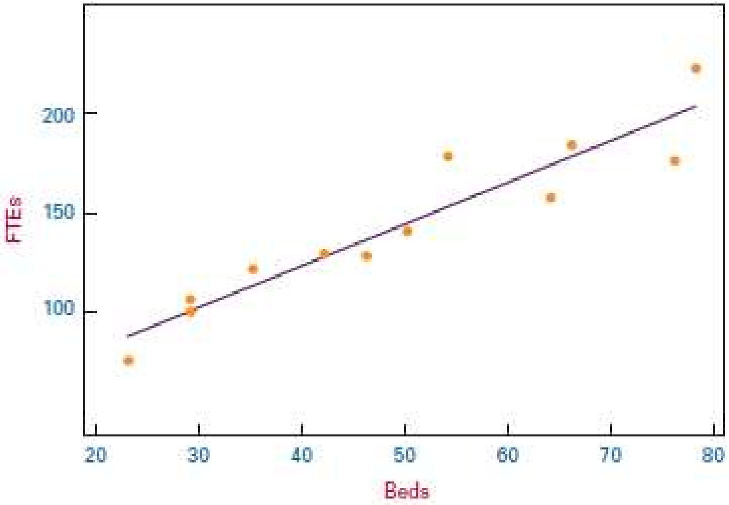
Using these values, the researcher solved for the sample slope (b_1) and the sample y-intercept (b_0).

$$SS_{xy} = \Sigma xy - \frac{(\Sigma x)(\Sigma y)}{n} = 92{,}038 - \frac{(592)(1692)}{12} = 8566$$

$$SS_{xx} = \Sigma x^2 - \frac{(\Sigma x)^2}{n} = 33{,}044 - \frac{(592)^2}{12} = 3838.667$$

$$b_1 = \frac{SS_{xy}}{SS_{xx}} = \frac{8566}{3838.667} = 2.232$$

$$b_0 = \frac{\Sigma y}{n} - b_1\frac{\Sigma x}{12} = \frac{1692}{12} - (2.232)\frac{592}{12} = 30.888$$



The least squares equation of the regression line is

$$\hat{y} = 30.888 + 2.232x$$

Coefficient of determination

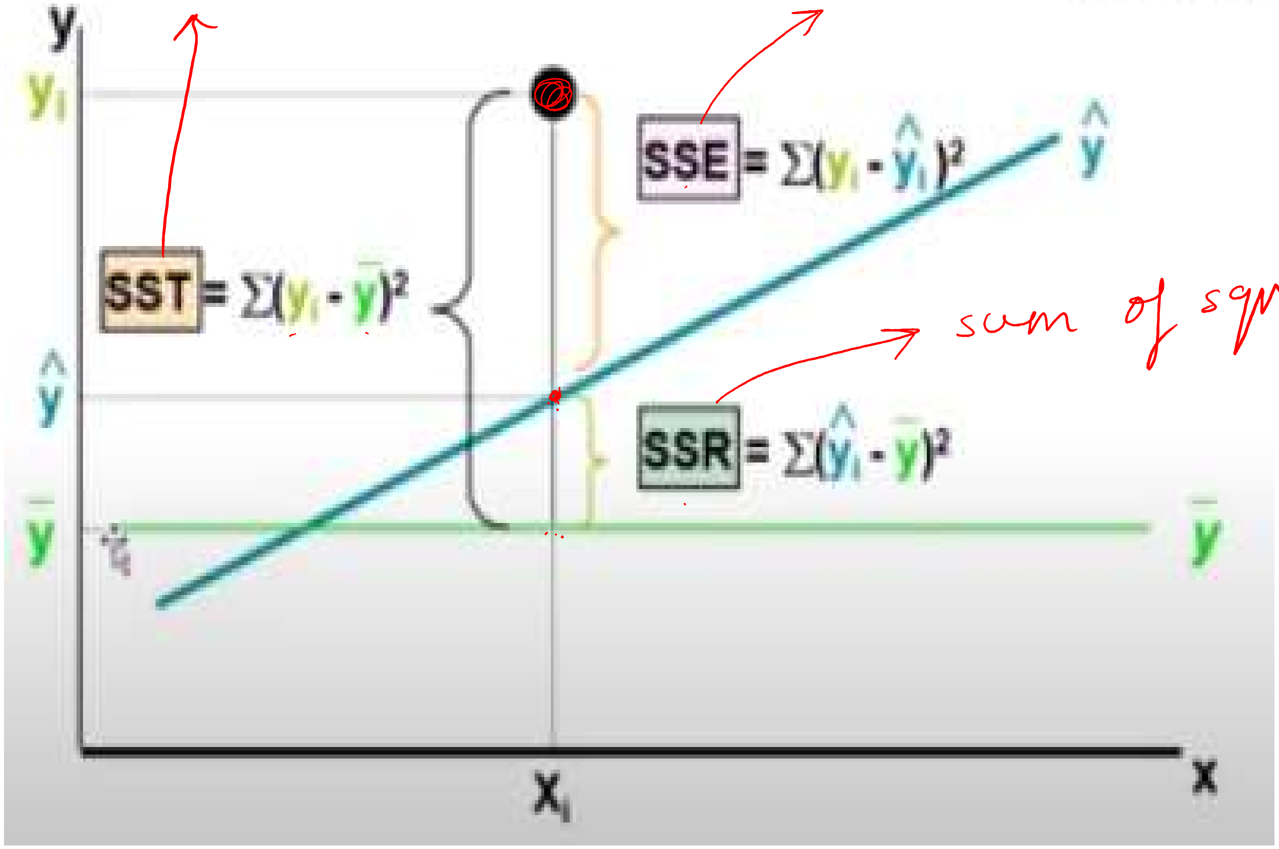
- The coefficient of determination is *the proportion of variability of the dependent variable (y) accounted for or explained by the independent variable (x).*
- The coefficient of determination ranges from 0 to 1.
- An r^2 of zero means that the predictor accounts for none of the variability of the dependent variable and that there is no regression prediction of y by x.
- An r^2 of 1 means perfect prediction of y by x and that 100% of the variability of y is accounted for by x.

Sum of Square of Totals

SST = SSE + SSR

sum of square of totals

sum of square of error



y
y_i
$\hat{y}$
$\bar{y}$
$SST = \Sigma(y_i - \bar{y})^2$
$SSE = \Sigma(y_i - \hat{y}_i)^2$
$SSR = \Sigma(\hat{y}_i - \bar{y})^2$
$\hat{y}$
$\bar{y}$
X_i
X

sum of square of regression

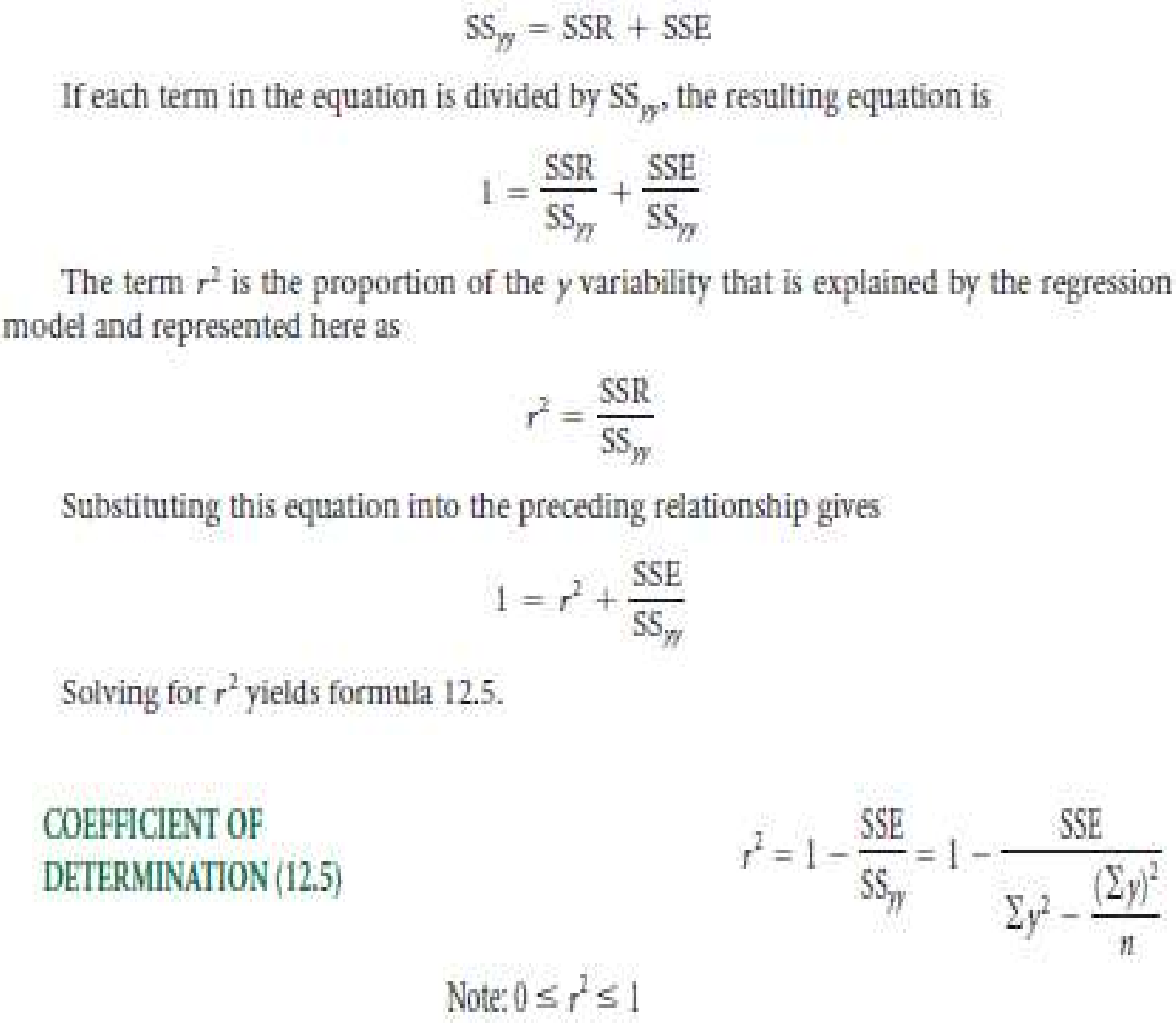
Image: Google

- The dependent variable, *y*, being predicted in a regression model has a variation that is measured by the sum of squares of *y* (SS_{yy}):

$$\text{SST} \leftarrow \underline{SS_{yy}} = \Sigma(y-\bar{y})^2 = \Sigma y^2 - \frac{(\Sigma y)^2}{n}$$

- This variation can be broken into two additive variations: the explained variation, measured by the sum of squares of regression (**SSR**), and the unexplained variation, measured by the sum of squares of error (**SSE**). This relationship can be expressed in equation form as

- Coefficient of determination, r^2= SSR/SST
- SST also called as SS_{yy}



$$SS_{yy} = SSR + SSE$$

If each term in the equation is divided by SS_{yy}, the resulting equation is

$$1 = \frac{SSR}{SS_{yy}} + \frac{SSE}{SS_{yy}}$$

The term r^2 is the proportion of the y variability that is explained by the regression model and represented here as

$$r^2 = \frac{SSR}{SS_{yy}}$$

Substituting this equation into the preceding relationship gives

$$1 = r^2 + \frac{SSE}{SS_{yy}}$$

Solving for r^2 yields formula 12.5.

COEFFICIENT OF DETERMINATION (12.5)

$$r^2 = 1 - \frac{SSE}{SS_{yy}} = 1 - \frac{SSE}{\sum y^2 - \frac{(\sum y)^2}{n}}$$

Note: $0 \le r^2 \le 1$

Significance of R-squared

In Graph 1:

All the points lie on the line

and the R2 value is a perfect 1

In Graph 2:

Some points deviate from

the line and the error is represented

by the lower R2 value of 0.70

In Graph 3:

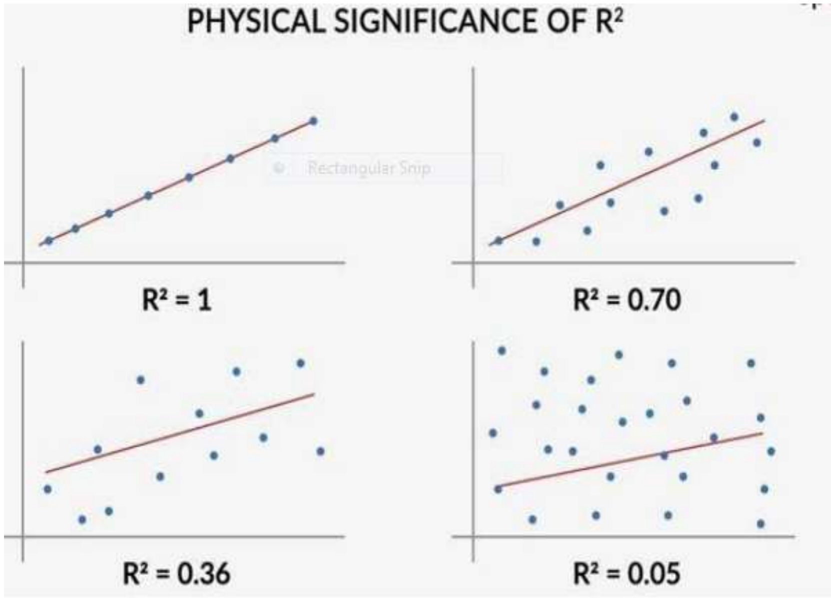
The deviation further

increases and the R2 value further

goes down to 0.36

In Graph 4:

The deviation is further higher with a very low R2 value of 0.05


PHYSICAL SIGNIFICANCE OF R²
Rectangular Snip
R² = 1
R² = 0.70
R² = 0.36
R² = 0.05