Correlation and Regression

Sections 9.1 and 9.2

In sections 9.1 and 9.2 we show how the **least square method** can be used to develop a linear equation relating two variables. The variable that is being predicted is called the dependent variable and the variable that is being used to predict the value of the dependent variable is called the independent variable. We generally use y to denote the dependent variable and use x to denote the independent variable.

Example 1 The instructor in a freshman computer science course is interested in the relationship between the time using the computer system (x) and the final exam score (y). Data collected for a sample of 10 students who took the course last semester are presented below. Draw the scatter diagram.

x= Hours Using	y= Final

Computer System	Exam Score
45	40
30	35
90	75
60	65
105	90
65	50
90	90
80	80
55	45
75	65

Scatter Diagram

Looking at the scatter diagram, it is clear that the relationship between the two variables can be approximately by a straight line. Clearly, there are many straight lines that could represent the relationship between x and y. The question is, which of the straight lines that could be drawn "best" represents the relationship?

The **least square method** is a procedure that is used to find the line that provides the best approximation for the relationship between x

and y. We refer to this equation of the line developed using the least square method as the **estimated regression equation**.

Estimated Regression Equation

$$\hat{y} = ax + b$$

where
b = y-intercept of the line
a = slope of the line
$\hat{y}$ = estimated value of the dependent variable

Least Square Criterion

It minimizes the sum of the squares of the differences between the observed y-values and estimated y-values. That is,

$$\text{Minimize } \left(\sum (y - \hat{y})^2\right)$$

Using differential calculus, it can be shown that the values of a and b can be found using the following equations.

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

$$b = \bar{y} - a\bar{x}$$

where $\bar{x} = \frac{\sum x}{n}$, $\bar{y} = \frac{\sum y}{n}$, and n = total number of observations.

Example 2: Find the regression line for the data given in Example 1.Use the regression line to estimate y when x = 80.

x	Y	xy	x^2
45	40	1800	2025
30	35	1050	900
90	75	6750	8100
60	65	3900	3600
105	90	9450	11025
65	50	3250	4225
90	90	8100	8100
80	80	6400	6400
55	45	2475	3025
75	65	4875	5625
695	635	48050	53025

$\bar{x} = \frac{\sum x}{n} = 695/10 = 69.5$

$\bar{y} = \frac{\sum y}{n} = 635/10 = 63.5$

$$a = \frac{\sum xy - n\bar{x}\bar{y}}{\sum x^2 - n(\bar{x})^2} = \frac{48050 - 10(69.5)(63.5)}{53025 - 10(69.5)^2} = \frac{3917.5}{4722.5}$$

$$= 0.8295$$

$$b = \bar{y} - a\bar{x} = 63.5 - 0.8295(69.5) = 5.8498$$

Thus the regression line is

$$\hat{y} = 5.85 + 0.83x$$

When x = 80, $\hat{y} = 5.85 + .83(80) = 72.25$

The Coefficient of Determination

We have shown that the least square method can be used to generate an estimated regression line. We shall now develop a means of measuring the **goodness** of the fit of the line to the data. Recall that the least squares method finds the line that minimizes the sum of the squares of the differences between the observed y-values and estimated y-values. We define SSE, the **Sum of Squares Due to Error** as

SSE = $\left(\sum (y - \hat{y})^2\right)$

x	y	y	$y - \hat{y}$	$(y - \hat{y})^2$
45	40	43.2	-3.2	10.24
30	35	30.75	4.25	18.06
90	75	80.55	-5.55	30.80
60	65	55.65	9.35	87.42
105	90	93	-3	9
65	50	59.8	-9.8	96.04
90	90	80.55	9.45	89.3
80	80	72.25	7.75	60.06
55	45	51.5	-6.5	42.25
75	65	68.1	-3.1	9.61
			SSE →	452.78

We define, **SST, the total sum of squares**, as

SST = $\left(\sum (y - \bar{y})^2\right)$

X	Y	$\bar{y}$	$y - \bar{y}$	$(y - \bar{y})^2$
45	40	63.5	-23.5	552.25
30	35	63.5	-28.5	812.25
90	75	63.5	11.5	132.25
60	65	63.5	1.5	2.25

105	90	63.5	26.5	702.25
65	50	63.5	-13.5	182.25
90	90	63.5	26.5	702.25
80	80	63.5	16.5	272.25
55	45	63.5	-18.5	342.25
75	65	63.5	1.5	2.25
			SST→	3702.5

We define, **SSR, the sum of squares due to regression**, as

$$SSR = \left(\sum(\hat{y} - \bar{y})^2\right)$$

Relationship between SST, SSR, and SSE

$$SST = SSR + SSE$$

Where

SST = Total Sum of squares

SSR = Sum of squares due to regression, and

SSE = Sum of squares due to error

So, SSR = SST – SSE

= 3702.5 – 452.78

= 3249.72

Now, let us see how these sums of squares can be used to provide a measure of goodness of fit for the regression relationship. Notice, we would have the best possible fit if every observation happened to lie on the least square line; in that case, the line would pass through each point, and we would have SSE = 0. Hence for a perfect fit, SSR must equal to SST, and thus the ratio SSR/SST = 1. On the other hand, a poorer fit to the observed data results in a larger SSE. Since SST = SSR + SSE, the largest SSE (and hence worst fit) occurs when SSR = 0.Thus the worst possible fit yields the ratio SSR/SST = 0.
We define $\mathbf{r^2}$**, the coefficient of determination**, as

$$r^2 = SSR/SST$$

Note: The coefficient of determination always lies between 0 and 1 and is a descriptive measure of the **utility of the regression**

equation for making prediction. Values of r^2 near to zero indicate that the regression equation is not very useful for making predictions, whereas values of r^2 near 1 indicate that the regression equation is extremely useful for making predictions.

Computational Formulas for SST and SSR

$$SST = \sum y^2 - n(\bar{y})^2$$

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

$$r^2 = SSR/SST$$

The following calculations show the computation of r^2 using the above example.

x	Y	xy	x^2	y^2
45	40	1800	2025	1600
30	35	1050	900	1225
90	75	6750	8100	5625
60	65	3900	3600	4225
105	90	9450	11025	8100

65	50	3250	4225	2500
90	90	8100	8100	8100
80	80	6400	6400	6400
55	45	4875	3025	2025
75	65	63.5	5625	4225
		48050	53025	44025

$$SSR = \frac{\left(\sum xy - n\bar{x}\bar{y}\right)^2}{\sum x^2 - n(\bar{x})^2} = \frac{\left(48050 - 10(69.5)(63,5)\right)^2}{53025 - 10(69.5)^2} = 3249.72$$

$$SST = \sum y^2 - n(\bar{y})^2 = 44025 - 10(63.5)^2$$
$$= 3702.50$$

$$r^2 = SSR/SST = 3249.72/3702.5 = 0.88$$

Linear Correlation

Several statistics can be employed to measure the correlation between the two variables. The one most commonly used is the **linear correlation coefficient, r.**

The linear correlation coefficient measures the strength of the linear association between two quantitative variables. You can find r by taking the square root of r^2 and the sign of the slope of the regression line.

Rules for interpreting r:

a. The value of r always falls between –1 and 1. A positive value of r indicates positive correlation and a negative value of indicates negative correlation.

b. The closer r is to 1, the stronger the positive correlation and the closer r is to –1, the stronger the negative correlation. Values of r closer to zero indicate no linear association.

c. The larger the absolute value of r, the stronger the relationship between the two variables.

d. r measures only the strength of linear relationship between two variables.

e. An assumption for finding the regression line for a set of data points is that data points are scattered about a straight line. That same assumption applies to the linear correlation coefficient, r. r measures only the strength of linear relationship between two variables. It should be employed as a descriptive measure only when a scatter diagram indicates that the data points are scattered about a straight line.