Mustansiriyah Uni.
College of science
Atmospheric Science Dept.

الجامعة المستنصرية
كلية العلوم
قسم علوم الجو

المرحلة ألرابعة

Lecture Title عنوان المحاضرة

Regression Analysis تحليل الانحدار

Lecturer Name اسم التدريسي

Dr. Ali Raheem Alnassar د. علي رحيم النصار

لجنة التعليم الالكتروني

The Coefficient of Determination (R^2):

The coefficient of determination gives an idea of how many data points fall within the results of the line formed by the regression equation.

The higher the coefficient, the higher percentage of points the line passes through when the data points and line are plotted.

If the coefficient is 0.80, then 80% of the points should fall within the regression line.

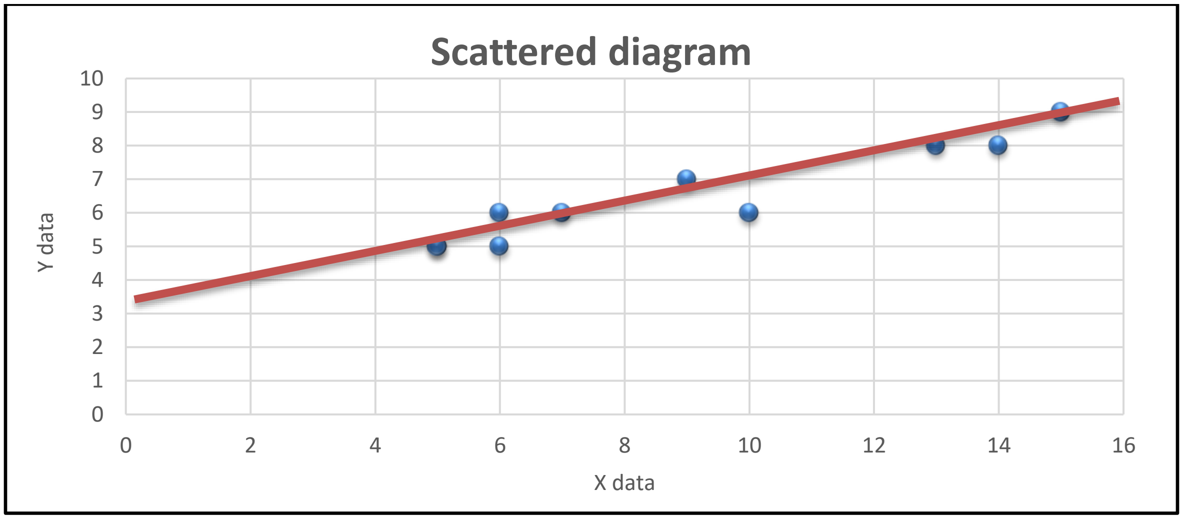
Values of 1 or 0 would indicate the regression line represents all or none of the data, respectively. A higher coefficient is an indicator of a better goodness of fit for the observations.

$$R^2 = \frac{b^2 \sum (X_i - \overline{X})^2}{\sum (Y_i - \overline{Y})^2} = \frac{b^2 \left(\sum X_i^2 - \frac{(\sum X_i)^2}{n} \right)}{\left(\sum Y_i^2 - \frac{(\sum Y_i)^2}{n} \right)}$$

Example: Find the regression equation for the following data

y	6	8	9	8	7	6	5	6	5	5
x	10	13	15	14	9	7	6	6	5	5

Solution



N	X	Y	X.Y	X^2
1	10	6	60	100
2	13	8	104	169
3	15	9	135	225
4	14	8	112	196
5	9	7	63	81
6	7	6	42	49
7	6	5	30	36
8	6	6	36	36
9	5	5	25	25
10	5	5	25	25
Total	**90**	**65**	**632**	**942**

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

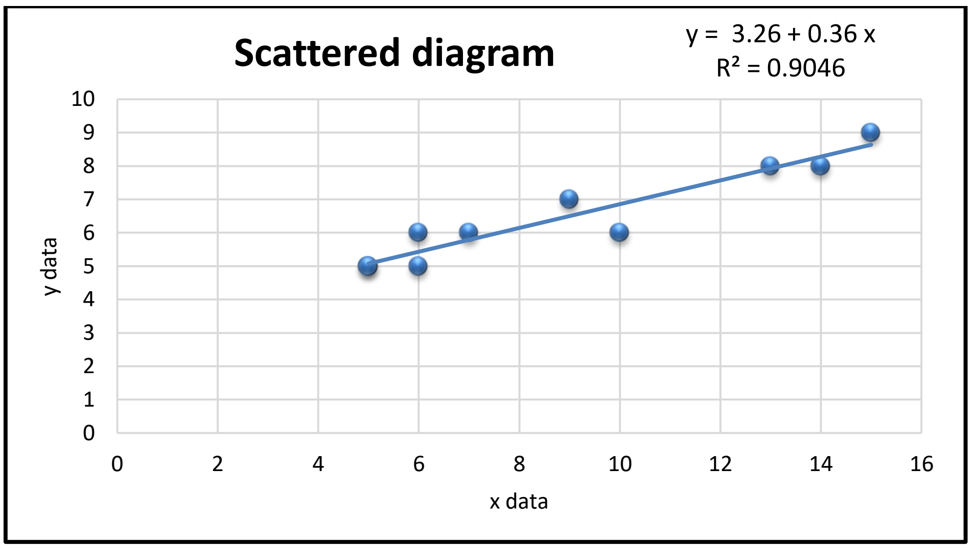
$$b = \frac{(10 * 632) - (90) * (65)}{(10 * 942) - (90)^2} = 0.36$$

$$a = \frac{\sum y - b\sum x}{n}$$

$$a = \frac{65 - (0.36 * 90)}{10} = 3.26$$

$$y = a + b.x$$

$$y = 3.26 + 0.36.x$$



Standard Error of Estimate

Standard Error of Estimate is the measure of the variation of an observation made around the computed regression line. Simply, it is used to check the accuracy of predictions made with the regression line.

Likewise, a standard deviation which measures the variation in the set of data from its mean, the standard error of estimate also measures the variation in the actual values of Y from the computed values of Y (predicted) on the regression line. It is computed as a standard deviation, and here the deviations are the vertical distance of every dot from the line of the average relationship. The deviation of each dot from the regression line is expressed as Y-Ye, thus the square root of the mean of standard deviation is:

$$\text{Se} = \sqrt{\frac{\sum\left(Ya - Ye\right)^2}{n-2}}$$

Ya = actual values
Ye= estimated values

This formula is not convenient as it requires to calculate the estimated value of Y i.e. Ye. Thus, more convenient and easy formula is given below:

$$\text{Se} = \sqrt{\frac{Syy - b.Sxy}{n-2}}$$

Syx is a measure of the variation of observed Y values from the regression line. Relatively low Syx indicates good fit

$$Syy = \sum Y_i^2 - \frac{\left(\sum Y_i\right)^2}{n}$$

$$Sxy = \sum X_i Y_i - \frac{\left(\sum X_i \sum Y_i\right)}{n}$$

The smaller the value of a standard error of estimate the closer are the dots to the regression line and the better is the estimate based on the equation of the line. If the standard error is zero, then there is no variation corresponding to the computed line and the correlation will be perfect.

Thus, the standard error of estimate measures the accuracy of the estimated figures, i.e. it is possible to ascertain the goodness and representativeness of the regression line as a description of the average relationship between the two series.

Example:

The following data represent the quantity of rice production (in 1000 kg) and the area planted in km from 2001 to 2010.

Quantity of rice(1000Kg)	112	128	130	138	158	162	140	175	125	142
Area planted in km	35	40	38	44	67	64	59	69	25	50

1- Determine the dependent variable and the independent variable.
2- Estimation of the regression equation.
3- What is the expected amount of rice when increasing the area to 80 km?

Solution

The dependent variable (Y) is the quantity produced from the rice crop and the area cultivated is the independent variable (X).

Years	Y_i	X_i	$\sum X_i Y_i$	$\sum X_i^2$	$\sum Y_i^2$
2001	112	35	3920	1225	12544
2002	128	40	5120	1600	16384
2003	130	38	4940	1444	16900
2004	138	44	6072	1936	19044
2005	158	67	10586	4489	24964
2006	162	64	10368	4096	26244
2007	140	59	8260	3481	19600
2008	175	69	12075	4761	30625
2009	125	25	3125	625	15625
2010	142	50	7100	2500	20164
$\sum$	1410	491	71566	26157	202094

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

$$a = \frac{\sum y - b\sum x}{n}$$

$$y = a + b.x$$

$$Y = 85.026 + 1.140\ X$$

From the above equation, it can be said that if the cultivated area is not increased, the amount of production will be 85.026 (1000kg), and the more the cultivated area increases by one km, it will lead to an increase in the quantity of production by 1.140 (1000kg).

- The expected quantity of rice crop production when increasing the cultivated area to 80 km, i.e. X = 80 is:

$$\hat{Y} = 85.026 + 1.140(80)$$

$$\hat{Y} = 85.026 + 91.2$$

$$\hat{Y} = 176.2(1000\text{kg}).$$

$$R^2 = \frac{(1.140)^2\left(26157 - \frac{(491)^2}{10}\right)}{202094 - \frac{(1410)^2}{10}}$$

$$= \frac{1.2996(2049)}{202094 - 198810} = \frac{2662.88}{3284} = 0.81$$

This means that 81% of the total changes in the amount of rice production (Y) are due to changes in the area cultivated with wheat (X), and that 19% are due to other changes and random changes.

To calculate Standard Error of Estimate:

a- Calculate **Standard Error of Estimate** by the first way:

Y_i	$\hat{Y}_i = 85.026 + 1.140(x)$	$Y_i - \hat{Y}_i$	$(y_i - \hat{y}_i)^2$
112	124.93	-12.93	167.18
128	130.63	-2.63	6.91
130	128.35	1.65	2.72
138	135.18	2.81	7.89
158	161.34	-3.377	11.55
162	157.98	4.02	16.16
140	152.28	-12.22	149.32
175	163.67	11.33	128.36
125	113.53	11.47	131.56
142	142.02	-0.02	0.0004
1410		0	621

$$S_e = \sqrt{\frac{\sum\left(Y_i - \hat{Y}_i\right)^2}{n-2}}$$

$$S_e = \sqrt{\frac{621}{10-2}}$$

$$S_e = \sqrt{77.625}$$

$$S_e = 8.81$$

b-calculate Standard Error of Estimate by the second way:

$\sum Y_i = 1410$ ، $\sum Y_i^2 = 202094$ b=1.140 ،n=10 ، $\sum X_i = 491$ ، $\sum X_i^2 = 26157$ ، $\sum X_i Y_i = 71566$

$$S_e = \sqrt{\frac{Syy - bSxy}{n-2}}$$:

$$Syy = \sum Y_i^2 - \frac{\left(\sum Y_i\right)^2}{n}$$

$$= 202094 - \frac{(1410)^2}{10}$$

$$= 202094 - 198810$$

$$= 3284$$

$$Sxy = \sum X_i Y_i - \frac{\left(\sum X_i \sum Y_i\right)}{n}$$

$$= 71566 - \frac{(491)(1410)}{10}$$

$$= 71566 - 69231$$

$$= 2335$$

$$\therefore \quad S_e = \sqrt{\frac{3284 - (1.140)2335}{10 - 2}}$$

$$S_e = \sqrt{\frac{622.1}{8}}$$

$$S_e = \sqrt{77.7625}$$

$$\therefore S_e = 8.81$$