Correlation and Regression

The main focus of both subjects was the difference between populations or subpopulations. In many other studies, however, the purpose of the research is to assess relationships among a set of variables. *For example*, the sample consists of pairs of values, say a ***mother's weight*** and ***her newborn's weight*** measured from each of 50 sets of mother and baby, and the research objective is concerned with the association between these weights.

Regression analysis is a technique for investigating relationships between variables; it can be used both for assessment of association and for prediction. Consider, for example, an analysis of whether or not a woman's age is predictive of her systolic blood pressure. As another example, the research question could be whether or not a leukemia patient's white blood count is predictive of his survival time. Research designs may be classified as experimental or observational.

Regression analyses are applicable to both types; yet the confidence one has in the results of a study can vary with the research type. In most cases, one variable is usually taken to be the response or dependent variable, that is, a variable to be predicted from or explained by other variables. The other variables are called predictors, or explanatory variables or independent variables.

The examples above, and others, show a wide range of applications in which the dependent variable is a continuous measurement. Such a variable is often assumed to be normally distributed and a model is formulated to express the mean of this normal distribution as a function of potential independent variables under investigation. The dependent variable is denoted by Y, and the study often involves a number of risk factors or predictor variables: X1;X2; . . . ;Xk.

The symbol **b** refer to regression, if b value is positive that is meaning all increasing in X values coming to increase in Y values, and opposite is true.(When **X** represents **independence variable,** and **Y** represents **satellite variable**).

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

$\hat{Y} = a + bX$

$\hat{Y}$:. Is the expect value to satellite variable.

a :. Is point cut to regression line with vertical axis.

b :. Regression factor.

X :. Independence variable.

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

Correlation

Correlation is the relationship between two variables, and its measured the type and strong the joint between two factors. When we need to learn if there are relationship between two variables or no, we must calculate ***the correlation factor***, the symbol **"r"** is refer to this factor.

$$\mathbf{r} = \frac{\sum xiyi - \frac{(\sum xi)(\sum yi)}{n}}{\sqrt{(\sum xi^2 - \frac{(\sum xi)^2}{n})(\sum yi^2 - \frac{(\sum yi)^2}{n}}}$$

the "r" value about -1 and +1 , in other word:.

$-1 \leq r \leq +1$

If the correlation is ***weak***, the correlation factor "**r**" is coming to ***zero***.

If the correlation is ***not found***, **r = 0.**

If the correlation is ***strong*** and ***positive***, the correlation factor **"r"** is coming to **+1**.(*i.e. the increasing (or decreasing)in one variable causing of increasing (or decreasing)in other variable).*

If the correlation is ***strong*** and ***negative***, the correlation factor **"r"** is coming to **-1**. (*i.e. the increasing(or decreasing) in* *one variable causing of decreasing(or increasing) in other variable).*

Coefficient of determinate

r^2 refer to ***coefficient of determinate*** or naming ***the capacity predictable*** , and this ratio explain the linear relationship between two variables.

Coefficient of indeterminate

K refer to ***coefficient of indeterminate*** , this coefficient consider the ratio that not explain the linear relationship between two variables.

$$K = 1 - r^2$$

if we want to tested the correlation relationship between two variables by **t test, should be compare between **t table** at α and df.=n-2and **t calculate**.*

$$t_{cal.} = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} \quad or = r\sqrt{\frac{n-2}{1-r^2}}$$

***Example*:.**

In Table below, the two columns give the values for the birth weight (x, in ounces) and the increase in weight between days 70 and 100 of life, expressed as a percentage of the birth weight (y) for 12 infants.

X	Y
112	63
111	66
107	72
119	52
92	75
80	118
81	120
84	114
118	42
106	72
103	90
94	91

If there are relationship between these two variable?

Answer:.

Variable	X	Y	X^2	Y^2	XY
	112	63	12544	3969	7056
	111	66	12321	4356	7326
	107	72	11449	5184	7704
	119	52	14161	2704	6188
	92	75	8464	5625	6900
	80	118	6400	13924	9440
	81	120	6561	14400	9720
	84	114	7056	12996	9576
	118	42	13924	1764	4956
	106	72	11236	5184	7632
	103	90	10609	8100	9270
	94	91	8836	8281	8554
Total	***1207***	***975***	***123,561***	***86,487***	***94,322***

$$\mathbf{r} = \frac{\sum xiyi - \frac{(\sum xi)(\sum yi)}{n}}{\sqrt{(\sum xi^2 - \frac{(\sum xi)^2}{n})(\sum yi^2 - \frac{(\sum yi)^2}{n})}}$$

$$\mathbf{r} = \frac{94322 - \frac{1207\times975}{12}}{\sqrt{(123561 - \frac{(1207)^2}{12})(86487 - \frac{(975)^2}{12})}}$$

r = - 0.946

we conclude there are negative strong correlation between two variables*(i.e. if the birth weight is increased that cause to decrease the percentage of weight after 70 to 100 days).*

r table at 0.01 and df. 12-2 =10 is ***0.7079***

absolute of r cal. > r tab.

Therefore there is significance correlation.

Applying the formulas, we obtain estimates for the slope and intercept as follows:

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

$$b = \frac{94322 - \frac{(1207)(975)}{12}}{123561 - \frac{(1207)^2}{12}}$$

$$= -\ 1.74$$

$$\overline{X} = \frac{1207}{12}$$

$$= 100.6$$

$$\bar{y} = \frac{975}{12}$$

$$= 81.3$$

$$a = \bar{y} - b\overline{X}$$

$$= 81.3 - (-\ 1.74)(100.6)$$

$$= 256.3$$

For example, if the birth weight is 95 oz, it is predicted that the increase between days 70 and 100 of life would be

$$\hat{Y} = a + bX$$

$$= 256.3 + (-1.74)(95)$$

$$= 90.1 \text{ \% of birth weight.}$$

Question:.

In the following table the first two columns give the values for age (x, in years) and systolic blood pressure (y, in mmHg) for 15 women. Calculate **r, r^2, k, b, $\hat{Y}$(50), $\hat{Y}$(45).**

Variable	X	Y	X^2	Y^2	XY
	42	130	1764	16900	5460
	46	115	2116	13225	5290
	42	148	1764	21904	6216
	71	100	5041	10000	7100
	80	156	6400	24336	12480
	74	162	5476	26224	11988
	70	151	4900	22801	10570
	80	156	6400	24336	12480
	85	162	7225	26224	13770
	72	158	5184	24964	11376
	64	155	4096	24025	9920
	81	160	6561	25600	12960
	41	125	1681	15625	5125
	61	150	3721	22500	9150
	75	165	5625	27225	12375
Total	***984***	***2193***	***67954***	***325889***	***146260***