SPECIAL ARTICLE

SOME APPLICATIONS OF STATISTICS TO MEDICAL RESEARCH

ROBERT T. MORRISON, B.Sc., M.D.,
Edmonton, Alta.

PART IV OF FOUR PARTS

LINEAR REGRESSION AND SIMPLE CORRELATION

ENUMERATION and measurement data have been dealt with previously; it is now necessary to introduce bivariate data. Bivariate data are those in which two different observations are made on each individual in the sample, for example, therapeutic response and dose of drug administered. It may be reasonable to assume that therapeutic response is dependent upon the dose of the drug. Thus dose is the independent variable and therapeutic response is the dependent one. It is often desirable to know the relationship between the dependent and the independent variable. If the independent variable is altered, is the dependent variable changed to the same degree? To establish the relation between the dependent and independent variables the methods of linear regression analysis might be employed. Linear regression is a measure of the relationship between two variables when the variation of one is sensibly dependent upon the variation of the other. Its units are those of the characteristics being measured such as the increase in g. % of hemoglobin per g. of ferrous sulfate administered per week.

When given a prescription for an anorexigenic agent, the obese patient frequently asks "Doctor, how much weight can I expect to lose?" This is not an unreasonable question and the answer might be determined in the following fashion. By means of a double-blind experiment* the drug is tested against a placebo and the results are compared. Suppose that there is a significant weight loss difference between the effect of the drug and the effect of the placebo. We wish to determine the relation between the dosage and the average weekly weight loss. If there is a fairly uniform increase in weight loss with uniform increase in dose of drug administered, the response is said to be linear. If the response is linear, it can be represented by a straight line, the equation for which is $Y = a + bX$. The dependent variable is customarily designated as Y and the independent variable as X. The term "a" is the value of Y when $X = 0$. "b" is the regression coefficient and the slope of the line. It is the increase in Y for unit increase in X. In the experiment we might have results as follows:

*Neither patient nor investigator knows which is placebo or which is the drug being tested until the conclusion of the experiment.

Weekly dose (mg.)	*Weekly weight loss (lb.)*
3.5	0.7
7.0	1.1
10.5	1.7
14.0	2.1
17.5	2.4
MEAN 10.5	1.6

It appears that there is a fairly uniform increase in weekly weight loss for uniform increase in weekly dose, and hence over the range of the experiment, a linear relationship exists between dose of drug and weight loss. A quick estimate shows that for each increase in dose of 3.5 mg. there is an increase in weight loss of about 0.42 lb. The regression coefficient is about 0.42/3.5 = 0.12 lb./mg./week.

A best fit straight line as shown on the graph (Fig. 5) yields an approximate value for the regression coefficient. A more accurate value is obtained using the following equation:

$$b = \frac{\Sigma(X - \bar{x})(Y - \bar{y})}{\Sigma(X - \bar{x})^2} = \frac{\Sigma(XY) - [(\Sigma X \Sigma Y)/n]}{\Sigma X^2 - [(\Sigma X)^2/n]}$$

Y = the value of the dependent variable (weekly weight loss in lb.)

$\bar{y}$ = the mean of Y

X = the value of the independent variable (weekly dose in mg.)

$\bar{x}$ = the mean of X

n = the number of observations = 5

X	Y	XY	X^2
3.5	0.7	2.45	12.25
7.0	1.1	7.70	49.00
10.5	1.7	17.85	110.25
14.0	2.1	29.40	196.00
17.5	2.4	42.00	306.25
$\Sigma X = 52.5$	$\Sigma Y = 8.0$	$\Sigma XY = 99.40$	$\Sigma X^2 = 673.75$
$\bar{x} = 10.5$	$\bar{y} = 1.6$		
$n = 5$			

$$b = \frac{99.4 - (420.0/5)}{673.75 - (2{,}756.25/5)} = \frac{99.4 - 84.0}{673.75 - 551.25}$$

$$= \frac{15.4}{122.50} = 0.13 \text{ lb./mg./week}$$

This value is close to the earlier estimate.

"a is calculated as follows: $\bar{y} = a + b\bar{x}$

$$a = \bar{y} - b\bar{x}$$
$$= 1.6 - (0.13 \times 10.5)$$
$$= 0.24$$

The equation for the regression line is

$$Y = 0.24 + 0.13X$$

Canad. M. A. J.
Mar. 18, 1961, Vol. 84

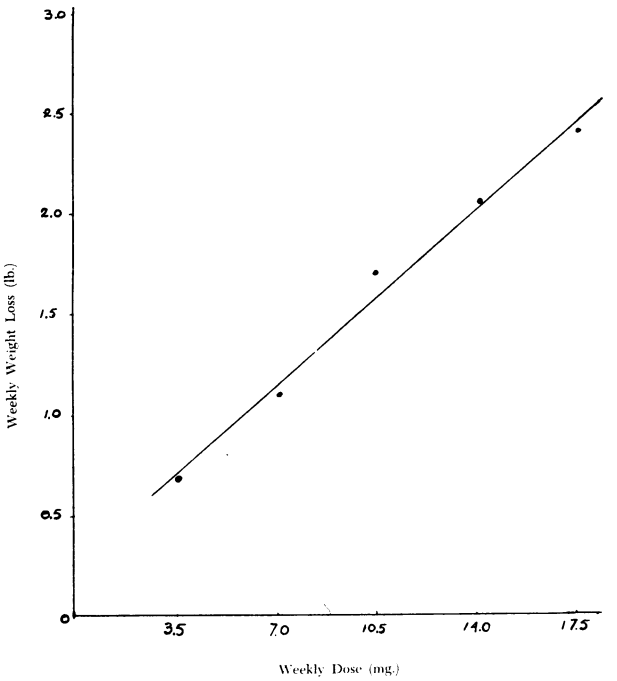

Fig. 5

Once the equation for the relationship between the two variables is known, the regression line can be drawn by plotting values of Y calculated from arbitrarily chosen values of X.

$$Y = 0.24 + 0.13X$$

at $X = 5$, $Y = 0.24 + 0.65 = 0.89$

at $X = 15$, $Y = 0.24 + 1.95 = 2.19$

The regression line is the line drawn through these points, and from it any value of Y may be found for any value of X. It is obvious that conclusions drawn from the regression line should be restricted to the area of the experimental results. The range of values for the regression coefficient is from minus infinity to plus infinity. A negative regression coefficient indicates that as the value of the independent variable increases, the value of the dependent variable decreases. A positive regression coefficient indicates that as one variable increases, so does the other.

SIMPLE CORRELATION

The methods of linear regression enable one to predict values of the dependent variable from values of the independent variable. This is appropriate for bivariate data when one variable is obviously dependent upon the other. However, some bivariate data cannot logically be classified into dependent and independent variables. Both of the variables may be logically independent of each other, yet in some way appear to be related, for example, height and weight data of healthy adult males. It is unreasonable to assume that weight is dependent upon height or vice versa, yet these two variables seem to be related. Thus the relationship between height and weight is not one of dependence and independence but rather one of covariation or correlation. The problem then is to measure the degree of correlation and this is done by calculating the correlation coefficient. Since it is a ratio, it has no units. The range of values for the correlation coefficient is from minus one to plus one. A minus value indicates negative correlation; as one variable increases, the other decreases. A value of zero indicates that there is no correlation, i.e. the variables are not related. A positive correlation coefficient indicates that a regular increase in one variable is accompanied by a regular increase in the other.

A statistically significant correlation coefficient does not prove a cause-and-effect relationship. It merely indicates the consistency with which the two variables increase or decrease together. For example, the incidence of lung cancer in North America has increased over the past few years; so has the per capita consumption of liquor. If enough data were collected comparing these two variables, it is not unlikely that a significant correlation coefficient could be calculated. On the basis of this significant correlation coefficient it is unreasonable to state that lung cancer is caused by drinking liquor or vice versa. In fact the two may not be related, yet a significant correlation coefficient can be calculated since both variables have increased over the past few years.

The correlation coefficient is calculated as follows:

$$r = \frac{\Sigma(XY) - \frac{(\Sigma X)(\Sigma Y)}{n}}{\sqrt{\left[\Sigma X^2 - \frac{(\Sigma X)^2}{n}\right]\left[\Sigma Y^2 - \frac{(\Sigma Y)^2}{n}\right]}}$$

Suppose we wish to determine the relationship between height and weight of ten healthy adult males selected at random from the population with the following results:

Height (in.)	*Weight (lb.)*			
X	Y	X^2	Y^2	XY
73	145	5,329	21,025	10,585
72	210	5,184	44,100	15,120
71	205	5,041	42,025	14,555
71	175	5,041	30,625	12,425
69	185	4,761	34,225	12,765
69	200	4,761	40,000	13,800
68	165	4,624	27,225	11,220
67	180	4,489	32,400	12,060
67	155	4,489	24,025	10,385
65	140	4,225	19,600	9,100
$\Sigma X = 692$	$\Sigma Y = 1,760$	$\Sigma X^2 = 47,944$	$\Sigma Y^2 = 315,250$	$\Sigma XY = 122,015$

The correlation coefficient is:

$$r = \frac{\Sigma(XY) - \frac{(\Sigma X)(\Sigma Y)}{n}}{\sqrt{\left[\Sigma X^2 - \frac{(\Sigma X)^2}{n}\right]\left[\Sigma Y^2 - \frac{(\Sigma Y)^2}{n}\right]}} = \frac{122{,}015 - \frac{1{,}217{,}920}{10}}{\sqrt{\left(47{,}944 - \frac{478{,}864}{10}\right)\left(315{,}250 - \frac{3{,}097{,}600}{10}\right)}}$$

$$r = \frac{223}{\sqrt{58 \times 5{,}490}} = \frac{223}{100\sqrt{31.842}}$$

$$= \frac{223}{564} = +0.40$$

We now wish to know the probability of selecting from a bivariate population in which there is no correlation, a correlation coefficient of +0.40. To determine this a table of "*r*", such as is found in Snedecor's text,[5] is used as follows. The number of degrees of freedom is equal to the number of individuals in the sample minus two. There were ten individuals in our series, so there are $10 - 2 = 8$ degrees of freedom. For eight degrees of freedom the values of "*r*" for 5% and 1% probability are 0.632 and 0.765 respectively. This means that from a bivariate population in which there is no correlation between the two variables, if samples of ten were selected at random and correlation coefficients calculated, in 5% of the cases a correlation coefficient of 0.632 would be obtained and similarly in 1% of cases a correlation coefficient of 0.765 would be obtained. It is obvious that since our calculated correlation coefficient was +0.40, the probability of obtaining this value from a bivariate population in which there is no correlation is something greater than 5%. The correlation coefficient then is not significant and the apparent relationship between height and weight is unproved. Obviously more data must be collected to obtain a significant correlation coefficient.

SUMMARY

Linear regression is a measure of the relationship between two variables when the variation of one is logically dependent upon variation of the other. When this relationship is not one of dependence and independence but rather one of covariation, the relationship is expressed by the correlation coefficient. Methods of calculating the regression coefficient and the correlation coefficient and examples of each are given.

REFERENCES

1. GOULDEN, C. H.: Methods of statistical analysis, John Wiley & Sons, Inc., New York, 2nd ed., 1952.
2. *Idem*: *Ibid.*, p. 82.
3. *Idem*: *Ibid.*, p. 443.
4. *Idem*: *Ibid.*, p. 444.
5. SNEDECOR, G. W.: Statistical methods applied to experiments in agriculture and biology, 5th ed., Iowa State College Press, Ames, Iowa, 1956.