Finding the relation between heights and weights of the 200 Students

Wae Phyoe Aung*

Abstract

The data was collected from 200 second year students majoring in Mathematics at Yadanabon University. Then the data was classified and the regression line and the correlation coefficient were calculated.

1. Introduction

Statistics, or Statistical methods as it is sometimes called, is playing an increasingly important role in nearly all phases of human endeavor. Formerly dealing only with affairs of the state, thus accounting for its name, the influence of statistics has now spread to agriculture, biology, business, chemistry, communications, economics, education, electronics, medicine, physics, political science, psychology, sociology and numerous other fields of science and engineering.

The purpose of this paper is to present the relation between the weights and the heights of 200 second year mathematics students in Yadanabon University. And also we went to know the correlation coefficient for their relation.

2. Correlation and Regression

We considered the problem of **regression** or estimation of one variable (the dependent variable) from one or more related variables(the independent variables).

We consider the closely related problem of **correlation**, or the degree of relationship between variables, which seeks to determine how well a linear of other equation describes or explains the relationship between variables.

3. The Least Square Regression Lines

We consider the problem of how well a straight line explains the relationship between two variables. To do this the least square regression line of Y on X is

$$Y = a_0 + a_1 X, \tag{1}$$

where

$$a_0 = \frac{(\sum f_y Y)(\sum f_x X^2) - (\sum f_x X)(\sum fXY)}{N\sum f_x X^2 - (\sum f_x X)^2} \quad \text{and} \quad a_1 = \frac{N\sum fXY - (\sum f_y Y)(\sum f_x X)}{N\sum f_x X^2 - (\sum f_x X)^2}.$$

* Lecturer, Department of Mathematics, Yadanabon University

f_x and f_y are the corresponding class frequencies and f represent the various cell frequencies corresponding to the pairs of class marks (X, Y).

Similarly,

$$X = b_0 + b_1 Y, \tag{2}$$

where

$$b_0 = \frac{(\sum f_x X)(\sum f_y Y^2) - (\sum f_y Y)(\sum fXY)}{N\sum f_y Y^2 - (\sum f_x Y)^2} \quad \text{and}$$

$$b_1 = \frac{N\sum fXY - (\sum f_y Y)(\sum f_x X)}{N\sum f_y Y^2 - (\sum f_y Y)^2}.$$

The equations (1) and (2) can also can be written respectively as

$$y = \left(\frac{\sum xy}{\sum x^2}\right)x \quad \text{and} \quad x = \left(\frac{\sum xy}{\sum y^2}\right)y,$$

where $x = X - \bar{X}$ and $y = Y - \bar{Y}$.

3.1 Least square regression equations for the grouped data

The following table is heights and weights of 200 Second Year mathematics students in Yadanabon University.

Table(1) Heights X(inches)

Weights Y (pounds)

	55–57	58–60	61–63	64–66	67–69	70–72	Total
69–87	4	11	12	1	1		29
88–106		19	42	19	5		85
107–125		4	19	21	20		64
126–144		2	3	9	2	3	19
145–163		1	1	1			3
Total	4	37	77	51	28	3	200

Then, we get

$$\sum fXY = 1326498,$$

$$N\sum fXY = 265299600.$$

The following frequency table is heights of 200 Second Year mathematics students in Yadanabon University.

Table(2)

X	Frequency f_x	f_xX	X^2	f_xX^2
56	4	224	3136	12544
59	37	2183	3481	128797
62	77	4774	3844	295988
65	51	3315	4225	215475
68	28	1904	4624	129472
71	3	213	5041	15123

From Table (2),

$$\sum f_xX = 12613,$$

$$\sum f_xX^2 = 797399,$$

$$\left(\sum f_xX\right)^2 = 159087769,$$

$$N\sum f_xX^2 = 159479800.$$

Y	Frequency f_y	f_yY	Y^2	f_yY^2
78	29	2262	6084	176436
97	85	8245	9409	799765
116	64	7424	13456	861184
135	19	2565	18225	346275
154	3	462	23716	71148

The following frequency table is weights of 200 Second Year mathematics students in Yadanabon University.

Table (3)

From Table (3),

$$\sum f_yY = 20958,$$

$$\sum f_yY^2 = 2254808,$$

$$\left(\sum f_yY\right)^2 = 439237764,$$

$$N\sum f_yY^2 = 450961600.$$

And then, we get

$$(\sum f_y Y)(\sum f_x X^2) = 16711888242,$$

$$(\sum f_x X)(\sum fXY) = 16731119274,$$

$$(\sum f_y Y)(\sum f_x X^2) - (\sum f_x X)(\sum fXY) = -19231032,$$

$$N\sum f_x X^2 - (\sum f_x X)^2 = 392031,$$

$$a_0 = \frac{(\sum f_y Y)(\sum f_x X^2) - (\sum f_x X)(\sum fXY)}{N\sum f_x X^2 - (\sum f_x X)^2}$$

$$= -49.05487576,$$

$$(\sum f_y Y)(\sum f_x X) = 264343254,$$

$$N\sum fXY - (\sum f_y Y)(\sum f_x X) = 956346,$$

$$a_1 = \frac{N\sum fXY - (\sum f_y Y)(\sum f_x X)}{N\sum f_x X^2 - (\sum f_x X)^2}.$$

$$= 2.439465246.$$

The regression line of Y on X is $Y = 2.439465246X - 49.05487576$.

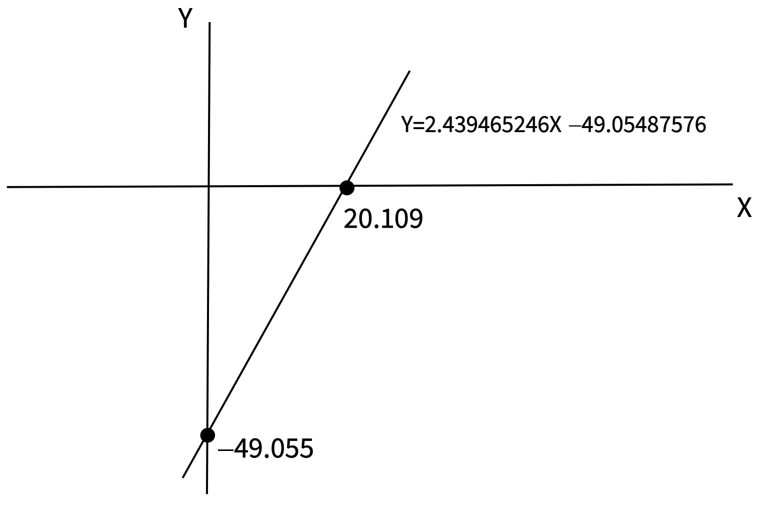

Figure (1)

$$(\sum f_x X)(\sum f_y Y^2) = 28439893304,$$

$$(\sum f_y Y)(\sum fXY) = 27800745084,$$

$$(\sum f_x X)(\sum f_y Y^2) - (\sum f_y Y)(\sum fXY) = 639148220,$$

$$N\sum f_y Y^2 - (\sum f_y Y)^2 = 11723836,$$

$$b_0 = \frac{(\sum f_x X)(\sum f_y Y^2) - (\sum f_y Y)(\sum fXY)}{N\sum f_y Y^2 - (\sum f_x Y)^2}$$

$$= 54.51698744,$$

$$b_1 = \frac{N\sum fXY - (\sum f_y Y)(\sum f_x X)}{N\sum f_y Y^2 - (\sum f_y Y)^2}$$

$$= 0.081572789.$$

The regression line of X on Y is $X = 0.081572789Y + 54.51698744$.

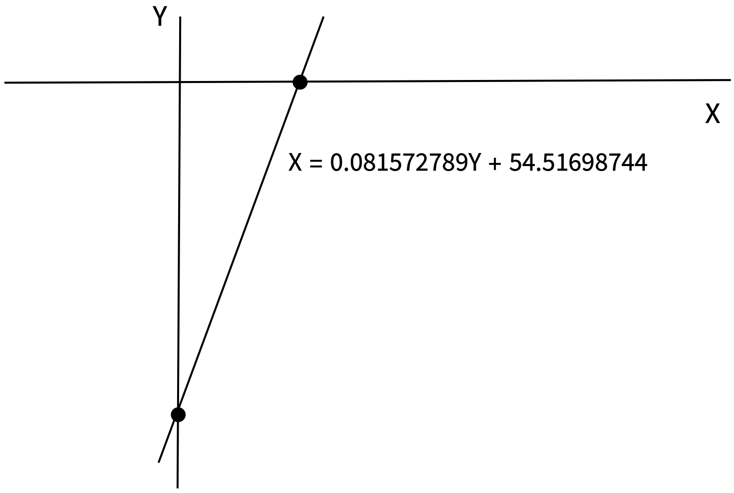

Figure (2)

4. Coefficient of Correlation

4.1 Standard error of estimate

We consider $Y_{esti.}$ represent the value of Y for given values of X as estimated form (1), a measure of the scatter about the regression line of Y on X is supplied by the quantity

$$S_{Y.X} = \sqrt{\frac{\sum (Y - Y_{esti.})^2}{N}}$$

which is called the **standard error of estimate** of Y on X.

4.2 Explained and unexplained variation

The total variation of Y is defined as $\sum\left(Y - \overline{Y}\right)^2$.

This can be expressed as

$$\sum\left(Y - \overline{Y}\right)^2 = \sum\left(Y - Y_{esti.}\right)^2 + \sum\left(Y_{esti.} - \overline{Y}\right)^2 \qquad (3)$$

In the above equation, $\sum\left(Y - Y_{esti.}\right)^2$ is unexplained variable and $\sum\left(Y_{esti.} - \overline{Y}\right)^2$ is explained variable.

4.3. Formula for Coefficient of Correlation

The coefficient of correlation $r = \pm\sqrt{\frac{\text{explained variation}}{\text{total variation}}} = \pm\sqrt{\frac{\sum\left(Y_{esti.} - \overline{Y}\right)^2}{\sum\left(Y - \overline{Y}\right)^2}} \qquad (4)$

and varies between – 1 and + 1.

If a linear relationship between two variables is assumed, equation(4) become

$$r = \frac{\sum xy}{\sqrt{\left(\sum x^2\right)\left(\sum y^2\right)}}. \tag{5}$$

In (5),

$$r = \frac{\sum\left(X-\bar{X}\right)\left(Y-\bar{Y}\right)}{\sqrt{\left[\sum\left(X-\bar{X}\right)^2\right]\left[\sum\left(Y-\bar{Y}\right)^2\right]}}.$$

However $\sum\left(X-\bar{X}\right)\left(Y-\bar{Y}\right) = \sum\left(XY - \bar{X}Y - X\bar{Y} - \bar{X}\,\bar{Y}\right)$

$$= \sum XY - \bar{X}\sum Y - \bar{Y}\sum X + N\bar{X}\,\bar{Y}$$

$$= \sum XY - N\bar{X}\,\bar{Y} - N\bar{Y}\bar{X} + N\bar{X}\,\bar{Y}$$

$$= \sum XY - \frac{\left(\sum X\right)\left(\sum Y\right)}{N},$$

where, $\bar{X} = \frac{\sum X}{N}$ and $\bar{Y} = \frac{\sum Y}{N}$.

And then, $\sum\left(X-\bar{X}\right)^2 = \sum\left(X^2 - 2X\bar{X} + \bar{X}^2\right) = \sum X^2 - \frac{\left(\sum X\right)^2}{N}$.

In a similar way $\sum\left(Y-\bar{Y}\right)^2 = \sum Y^2 - \frac{\left(\sum Y\right)^2}{N}$.

Therefore, the linear correlation coefficient is

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

4.4 Correlation Coefficient for the grouped data

We consider the f represent the various cell frequencies corresponding to the pairs of class marks (X,Y) , then we obtain

$$r = \frac{N\sum fXY - (\sum f_y Y)(\sum f_x X)}{\sqrt{\left[N\sum f_x X^2 - \left(\sum f_x X\right)\right]\left[N\sum f_y Y^2 - \left(\sum f_y Y\right)^2\right]}},$$

$$N\sum f_x X^2 - \left(\sum f_x X\right)^2 = 392031,$$

$$N\sum f_y Y^2 - \left(\sum f_y Y\right)^2 = 11723836,$$

$$\sqrt{\left[N\sum f_x X^2 - \left(\sum f_x X\right)\right]\left[N\sum f_y Y^2 - \left(\sum f_y Y\right)^2\right]} = 214385.342,$$

Coefficient of correlation is $r = 0.446087417$.

Conclusion

From the observed data, our coefficient of correlation is 0.446087417 or 44.61%.This mean that the graph of weights and heights are not bell shaped.

Acknowledgements

I would like to pay my deep thanks to Dr Si Si Khin, Pro–Rector and Dr Tint Moe Thuzar, Pro–Rector, Yadanabon University for their exhorting and encouragement to submit this paper. And then, I deeply indebted to Dr Win Kyaw, Professor and Head of Department of Mathematics, Dr Hla Win, Professor, Department of Mathematics and Dr Nan Mya Ngwe, Professor, Department of Mathematics, Yadanabon University, for their incitement and fruitful comments on this research.

References

[1] Murray R. S., " Theory and Problems of Statistics", McGraw–Hill, Inc., USA, 1961.