ST. LAWRENCE HIGH SCHOOL
A JESUIT CHRISTIAN MINORITY INSTITUTION

STUDY MATERIAL-4
SUBJECT – STATISTICS

Pre-test

Chapter: BIVARIATE ANALYSIS **Class: XII**

Topic:Regression **Date: 13.05.2020**

REGRESSION & RANK CORRELATION

PART 4

- The slope of the regression line y on x is $m_1 = b_{yx}$ and the slope of the regression line x on y is $m_2 = \frac{1}{b_{xy}}$.

  So the angle between the lines is $\theta = tan^{-1}\left(\frac{m_2 - m_1}{1 + m_2\, m_1}\right) = tan^{-1}\left(\frac{1 - b_{yx}\, b_{xy}}{b_{yx} + b_{xy}}\right) = tan^{-1}\left(\frac{1 - r_{xy}^2}{b_{yx} + b_{xy}}\right)$

  Case 1 : When the lines are parallel $\theta = 0 \implies 1 - r_{xy}^2 = 0 \implies r_{xy} = \pm 1$.

  In this case the lines coincide on each other.

  Case 2 : When the lines are perpendicular $b_{yx} + b_{xy} = 0$

  $$\implies b_{yx} = -b_{xy}.$$

  Which is not possible as both the regression coefficients should be of same sign. It is mathematically true only when $r_{xy} = 0$.

  (In statistics we do not find the regression line for linearly independent variables. So we say mathematically true.)

- The mean of error is zero.

  Pf: From regression line y on x, $Y_i = \bar{y} + b_{yx}(x - \bar{x})$

  $\implies \sum Y_i = \sum y_i + b_{yx} \sum(x_i - \bar{x}) \implies n\bar{Y} = n\,\bar{y} \implies \bar{Y} = \bar{y}$

  So mean of the predicted value of y is same as the corresponding observed value of y.

  Now $\bar{e} = \frac{1}{n}\sum e_i = \frac{1}{n}\sum(y_i - Y_i) = \bar{Y} - \bar{y} = 0$

- Variance of error.

  $s_e^2 = \frac{1}{n}\sum e_i^2 - \bar{e}^2 = \frac{1}{n}\sum e_i^2 \; as \; \bar{e} = 0$

  $\implies s_e^2 = \frac{1}{n}(y_i - Y_i)^2 = \frac{1}{n}\sum((y_i - \bar{y}) - r_{xy}\frac{s_y}{s_x}(x_i - \bar{x}))^2$

$$= \frac{1}{n}\sum(y_i - \bar{y})^2 - 2\, r_{xy}\frac{s_y}{s_x}\ \frac{1}{n}\sum(x_i - \bar{x})(y_i - \bar{y}) + r_{xy}^2\frac{s_y^2}{s_x^2}\frac{1}{n}\sum(x_i - \bar{x})^2$$

$$= s_y^2 - 2\, r_{xy}\frac{s_y}{s_x}\, cov\,(x,y) + r_{xy}^2\frac{s_y^2}{s_x^2}\, s_x^2$$

$$= s_y^2 - 2\, r_{xy}\frac{s_y}{s_x}\, s_x s_y r_{xy} + r_{xy}^2\, s_y^2$$

$$= s_y^2 - r_{xy}^2\, s_y^2 = s_y^2(1 - r_{xy}^2)$$

$$\Longrightarrow s_e = s_y\sqrt{(1 - r_{xy}^2)}$$

s_e is known as standard of estimate of y in regression line y on x.

Since $s_e \geq 0 \Longrightarrow s_y\sqrt{(1 - r_{xy}^2)} \geq 0 \Longrightarrow \sqrt{(1 - r_{xy}^2)} \geq 0$

$$\Longrightarrow \left(1 - r_{xy}^2\right) \geq 0 \Longrightarrow r_{xy}^2 \leq 1 \Longrightarrow -1 \leq r_{xy} \leq 1.$$

From $s_e^2 = s_y^2\left(1 - r_{xy}^2\right)$, we see that if $r_{xy} = 0$, then $s_e^2 = s_y^2$, so that errors of estimation are as much variable as the observed values, and hence the regression line is of no help as a prediction formula. This is also seen from the fact that when $r_{xy} = 0$, we get $Y = \bar{y}$, a constant.

It is also noted that the numerical value of r_{xy} increases, s_e^2 decreases and when $r_{xy} = \pm 1$, $s_e^2 = 0$ which implies that, for each i.

$$e_i = 0\ or\ y_i = Y_i$$

So that all the points in scattered diagram lie on the regression line and, thus, the regression lines become a perfect prediction formula. From these observations, it is clear that the numerical value of r_{xy} can be taken as a measure of the efficiency of the regression equation as a prediction formula.

We get the same result if we consider the variance of the errors of estimates of x from its linear regression on y.

- Variance of predicted values of y in regression y on x.

$$s_Y^2 = \frac{1}{n}\sum(Y_i - \bar{Y})^2 = \frac{1}{n}\sum(\bar{y} + b_{yx}(x_i - \bar{x}) - \bar{y})^2$$

$$= b_{yx}^2 \ \frac{1}{n}\sum(x_i - \bar{x})^2 = r_{xy}^2 \ \frac{s_y^2}{s_x^2}\ s_x^2 = \ r_{xy}^2 \ s_y^2$$

$$\Rightarrow \ r_{xy}^2 = \frac{s_Y^2}{s_y^2}.$$

Which is known as coefficient of determination. It may be used as measure of usefulness of linear regression equation on x.

$|r_{xy}| = \frac{s_Y}{s_y}$. It shows the proportion of total variability of y which is accounted for by its linear regression on x.

It can similarly be shown that $|r_{xy}| = \frac{s_X}{s_x}$ where s_X is the standard deviation of the predicted values of x from its regression on y.

- $Cov\ (x, e) = \frac{1}{n}\sum(x_i - \bar{x})e_i = \frac{1}{n}\sum x_i e_i - \bar{x}\frac{1}{n}\sum e_i = \frac{1}{n}\sum x_i e_i - \bar{x}\bar{e}$

$$= \frac{1}{n}\sum x_i e_i = 0$$

$\Rightarrow r_{xe} = 0$

Since $\bar{e} = 0$ and $\sum x_i e_i = \sum x_i(y_i - a - bx_i)$

$$= \sum x_i y_i - na\sum x_i - b\sum x_i^2$$

= 0 (from 2nd normal equation)

- $Cov\ (Y, e) = cov(a + bx, e) = cov(a, e) + b.cov(x, e) = 0$

Since a=constant and $cov(x, e) = 0$

So $r_{Ye} = 0$

- $Cov\ (y, Y) = \ cov(Y + e, Y) = s_Y^2 + \ cov(\ Y, e) = \ s_Y^2$

  $r_{yY} = \frac{cov(y, Y)}{s_y s_Y} = \frac{s_Y^2}{s_Y s_y} = \frac{s_Y}{s_y} = \ |r_{xy}|$

  This shows that correlation between y and its predicted value Y is non-negative and numerically equal to the correlation between y and x.

Spearman's rank correlation(no tie case)

Spearman's rank correlation coefficient $r_R = 1 - \frac{6 \sum d_i^2}{n(n^2-1)}$

Where $d_i = \ u_i - v_i$

u_i = Rank given by the judge 1 to the ith candidate

v_i = Rank given by the judge 2 to the ith candidate

n = no of candidates

Example1.

The following are the results of 5 candidates in a singing competition given by two different judges.

J1:	12	18	11	8	10
J2:	15	12	9	11	13
u_i:	2	1	3	5	4
v_i:	1	3	5	4	2
d_i:	1	-2	-2	1	2
d_i^2:	0	4	4	1	4

So $\sum d_i^2 = \ 13$ and $r_R = 1 - \frac{6\sum d_i^2}{n(n^2-1)} = 1 - \frac{6X\ 13}{5X24} = \frac{7}{20}$

Derivation of Spearman's rank correlation coefficient eith no tier that

From the above example it is clear that u_i and v_i posses the natural numbers.

So $\sum_{i=1}^{n} u_i^2 = \sum_{i=1}^{n} v_i^2 = \frac{n(n+1)}{2} \ \Rightarrow \bar{u} = \ \bar{v} = \frac{n+1}{2}$

$$\sum_{i=1}^{n} u_i^2 = \sum_{i=1}^{n} v_i^2 = \frac{n(n+1)(2n+1)}{6}$$

$$s_u^2 = \frac{1}{n}\sum_{i=1}^{n} u_i^2 - (\bar{u})^2 = \frac{1}{n}\frac{n(n+1)(2n+1)}{6} - \ (\frac{n+1}{2})^2 \ = \frac{n^2-1}{12} = s_v^2$$

$$\sum_{i=1}^{n} d_i^2 = \sum_{i=1}^{n}(u_i - \ v_i)^2 = \sum_{i=1}^{n} u_i^2 + \sum_{i=1}^{n} v_i^2 - \ 2\sum_{i=1}^{n} u_i v_i$$

$$\Rightarrow \frac{1}{n}\sum_{i=1}^{n} u_i v_i = \frac{(n+1)(2n+1)}{6} - \frac{1}{2n}\sum_{i=1}^{n} d_i^2$$

$$\Rightarrow cov(u,v) = \frac{1}{n}\sum_{i=1}^{n} u_i v_i - \ \bar{u}\bar{v} = \frac{(n+1)(2n+1)}{6} - \frac{1}{2n}\sum_{i=1}^{n} d_i^2 - \ (\frac{n+1}{2})^2$$

$$\Rightarrow cov\ (u,v) = \frac{(n^2-1)}{12} - \frac{1}{2n}\sum_{i=1}^{n} d_i^2$$

So $r_R = \ r_{uv} = \frac{cov\,(u,v)}{s_u s_v} = \frac{\frac{(n^2-1)}{12} - \frac{1}{2n}\sum_{i=1}^{n} d_i^2}{\frac{(n^2-1)}{12}} \ = 1 - \ \frac{6\sum d_i^2}{n(n^2-1)}$.

Property of rank correlation coefficient:

- The rank correlation cefficient lies between -1 and +1.

  $s_d^2 = \frac{1}{n}\sum_{i=1}^{n} d_i^2 - (\bar{d})^2 = \frac{1}{n}\sum_{i=1}^{n} d_i^2$ ……(*) (since $\bar{d} = \bar{u} - \bar{v} = 0$)

  $s_d^2 = s_{u-v}^2$

  $\Rightarrow s_d^2 \leq s_{u-v}^2 + s_{u+v}^2$ since $s_{u+v}^2 \geq 0$

  $\Rightarrow s_d^2 \leq 2(s_u^2 + s_v^2) \Rightarrow s_d^2 \leq 2\,\frac{2(n^2-1)}{12} = \frac{n^2-1}{3}$……(**)

  Combining (*) and (**) we get,

  $\frac{1}{n}\sum_{i=1}^{n} d_i^2 \leq \frac{n^2-1}{3} \Rightarrow r_R \geq 1 - \frac{6n(n^2-1)}{3n(n^2-1)} \Rightarrow r_R \geq -1$ …….(***)

  $r_R = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \leq 1 \; since \; 1 - \frac{6\sum d_i^2}{n(n^2-1)} \geq o$ ……(****)

  $Combining\ (***) and\ (****), -1 \leq r_R \leq +1$

- In case of **perfect agreement** between two judges $u_i = v_i \Rightarrow d_i = 0$

  $\Rightarrow \sum_{i=1}^{n} d_i^2 = 0 \Rightarrow r_R = 1 - \frac{6\sum d_i^2}{n(n^2-1)} = 1$

  In case of **perfect disagreement** between two judges $u_i + v_i = n + 1$

  $$\sum_{i=1}^{n} d_i^2 = \sum_{i=1}^{n} (u_i - v_i)^2 = \sum_{i=1}^{n} (n + 1 - 2v_i)^2$$

  $$= \sum_{i=1}^{n} (n+1)^2 + 4\sum_{i=1}^{n} v_i^2 - 4(n+1)\sum_{i=1}^{n} v_i$$

  $$= \frac{1}{3}\, n(n^2 - 1)$$

So $r_R = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$ =-1

These two are the extreme cases . So all other cases lie in between these two case.

Hence $-1 \leq r_R \leq +1$

- ❖ Uses of rank correlation:
    - Rank correlation is used to measure the extent of association that may exist between two series of ranks for the same set of individuals.
    - Attributes cannot be measured but individuals can be ranked according to their degree of possession of an attribute. Hence rank correlation can be used to see whether two attributes are correlated or not.
    - In case of two judges, rank correlation shows the extent of agreement between the judges.
    - Sometimes rank correlation is used instead of simple correlation because rank correlation coefficient calculation is much easier.

Prepared by
Snjay Bhattacharya