BUS 172
Descriptive Statistics

Lecture 20

Measures of Relationships Between Variables

✓Covariance: Covariance is a measure of the linear relationship between two variables. A positive value indicates a direct or increasing linear relationship and a negative value indicates a decreasing linear relationship.

A sample covariance is

$$Cov(x, y) = S_{xy} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{n-1}$$

✓**Correlation**: If the change in one variable effects a change in the other variable, the variables are said to be correlated.

✓If the increase (decrease) in one variable results in the corresponding increase in the other i.e. if the changes are in the same direction, the variables are positively correlated. e.g. Height and weight of a group of people.

✓If the increase (decrease) in one variable results in the corresponding decrease (increase) in the others, i.e. if the changes are in the opposite direction, the variables are negatively correlated. e.g. Volume and pressure of perfect gas.

$$\text{Correlation Coefficient } (x, y) = r_{xy} = \frac{S_{xy}}{S_x S_y}$$

$$= \frac{\sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n}}{\sqrt{\{\sum x_i^2 - \frac{(\sum x_i)^2}{n}\}\{\sum y_i^2 - \frac{(\sum y_i)^2}{n}\}}}$$

✓**Correlation**:

✓The correlation coefficient ranges from -1 to +1.

✓When r =0 there is no linear relationship between x and y but not necessarily a lack of relationship.

✓The closer “r” is to +1, represents strong positive relationship.

✓The closer “r” is to -1, represents strong negative relationship.

✓Correlation indicates whether there is any relation between the variables and correlation coefficient measures the extent of relationship between them.

✓**Regression**: Regression measures the probable movement of one variable in term of the other. Therefore regression is used for prediction or forecasting purpose.

✓Suppose the movement of the variable Y is dependent on the movement of X variable. Hence Y is dependent variable and X is independent variable. Let the regression line of Y on X be

Y=a+ bX, where, a = intercept or constant, b= slope coefficient

$$b = \frac{\sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n}}{\sum x_i^2 - \frac{(\sum x_i)^2}{n}} = \frac{S_{xy}}{S_x^2} \qquad a = \bar{y} - b\bar{x}$$

Example

The table shows the math achievement test scores for a random sample of n = 10 college freshmen, along with their final calculus grades.

Student	1	2	3	4	5	6	7	8	9	10
Math test, x	39	43	21	64	57	47	28	75	34	52
Calculus grade, y	65	78	52	82	92	89	73	98	56	75

Use your calculator to find the sums and sums of squares.

$$\sum x = 460 \qquad \sum y = 760$$

$$\sum x^2 = 23634 \quad \sum y^2 = 59816$$

$$\sum xy = 36854$$

$$\bar{x} = 46 \qquad \bar{y} = 76$$

Example

$$S_{xx} = 23634 - \frac{(460)^2}{10} = 2474$$

$$S_{yy} = 59816 - \frac{(760)^2}{10} = 2056$$

$$S_{xy} = 36854 - \frac{(460)(760)}{10} = 1894$$

$$b = \frac{1894}{2474} = .76556 \quad \text{and} \quad a = 76 - .76556(46) = 40.78$$

$$\text{Bestfitting line:} \quad \hat{y} = 40.78 + .77x$$

Goodness of fit:

- The overall goodness of fit of the regression is measured by the coefficient of determination, r^2.
- It explains what proportion of variation in the dependent variable is explained by the explanatory variable.
- $0 \le r^2 \le 1$: the closer it is to 1, the better is the fit.
- e.g. if $r^2 = 0.92$, it means that 92% of variation in Y is explained by X.
- In the case of multivariate regression, the coefficient of determination is denoted by R^2.

$$r^2 = \frac{\sum(\hat{y}_i - \overline{y})^2}{\sum(y_i - \overline{y})^2} = \frac{SSR}{TSS} = 1 - \frac{SSE}{TSS} = \frac{s_{xy}^2}{s_{xx}s_{yy}},$$

y_i = actual value, $\hat{y}_i$ = predicted value

The Analysis of Variance

TSS= Total Sum of Squares = $\sum (y_i - \bar{y})^2 = S_{yy}$ =SSR+SSE

SSR= Sum of Squares of Regression= $\sum (\hat{y}_i - \bar{y})^2 = \dfrac{s_{xy}^2}{s_{xx}}$

=Variation in y explained by regression

SSE= Sum of Squares of Error =$\sum \hat{u}_i^2 = \sum (y_i - \hat{y}_i)^2$,

$= S_{yy} - \dfrac{s_{xy}^2}{s_{xx}}$ = unexplained variation in y

y_i = actual value, $\hat{y}_i$ = predicted value

The Analysis of Variance

We calculate

$$\text{SSR} = \frac{(S_{xy})^2}{S_{xx}} = \frac{1894^2}{2474}$$

$$= 1449.9741$$

SSE= Total SS- SSR

$$= S_{yy} - \frac{(S_{xy})^2}{S_{xx}}$$

$$= 2056 - 1449.9741$$

$$= 606.0259$$

The ANOVA Table

Total df = **n -1**

Regression df = **K=1**

Error df = **n –k – 1 = n - 2**

Mean Squares

MSR = SSR/(1)

MSE = SSE/(n-2)

Source	df	SS	MS	F
Regression	K=1	SSR	SSR/(1)	MSR/MSE 1/(n-2) df
Error	(n -k-1)=n-2	SSE	SSE/(n-2)	
Total	n -1	Total SS		

The Calculus Problem

$$\text{SSR} = \frac{(S_{xy})^2}{S_{xx}} = \frac{1894^2}{2474} = 1449.9741$$

$$\text{SSE} = \text{Total SS} - \text{SSR} = S_{yy} - \frac{(S_{xy})^2}{S_{xx}}$$

$$= 2056 - 1449.9741 = 606.0259$$

Source	df	SS	MS	F
Regression	1	1449.9741	1449.9741	19.14
Error	8	606.0259	75.7532	
Total	9	2056.0000		