CORRELATION AND REGRESSION

Dr. Sabiha Khan

- **Correlation and linear regression are the most commonly used techniques for investigating the relationship between two quantitative variables.**
- The goal of correlation analysis is to see whether two measurement variables co vary, and to quantify the **strength** of the relationship between the variables, whereas **regression** expresses the relationship in the form of an equation.
- For example, in students taking a Maths and English test, we could use correlation to determine whether students who are good at Maths tend to be good at English as well, and regression to determine whether the marks in English can be predicted for given marks in Maths.

Why Use Correlation?

We can use the correlation coefficient, such as the Pearson Product Moment Correlation , to test if there is a linear relationship between the variables. To quantify the strength of the relationship, we can calculate the correlation coefficient (r). Its numerical value ranges from +1.0 to -1.0. $r > 0$ indicates positive linear relationship, $r < 0$ indicates negative linear relationship while $r = 0$ indicates no linear relationship.

- **A Caveat**

It must, however, be considered that there may be a third variable related to both of the variables being investigated, which is responsible for the apparent correlation. Correlation does not imply causation. Also, a nonlinear relationship may exist between two variables that would be inadequately described, or possibly even undetected, by the correlation coefficient.

Why Use Regression

- In regression analysis, the problem of interest is the nature of the relationship itself between the dependent variable (response) and the (explanatory) independent variable.
- The analysis consists of choosing and fitting an appropriate model, done by the method of least squares, with a view to exploiting the relationship between the variables to help estimate the expected response for a given value of the independent variable. For example, if we are interested in the effect of age on height, then by fitting a regression line, we can predict the height for a given age.

ASSUMPTIONS:

Some underlying assumptions governing the uses of correlation and regression are as follows.

- The observations are assumed to be independent. For correlation, both variables should be random variables,
- But for regression only the dependent variable Y must be random. In carrying out hypothesis tests, the response variable should follow Normal distribution and the variability of Y should be the same for each value of the predictor variable.
- A scatter diagram of the data provides an initial check of the assumptions for regression.

Uses of Correlation and Regression

There are three main uses for correlation and regression.

- One is to test hypotheses about cause – and – effect relationships. In this case, the experimenter determines the values of the X-variable and sees whether variation in X causes variation in Y. For example, giving people different amounts of a drug and measuring their blood pressure.
- The second main use for correlation and regression is to see whether two variables are associated, without necessarily inferring a cause – and – effect relationship. In this case, neither variable is determined by the experimenter; both are naturally variable. If an association is found, the inference is that variation in X may cause variation in Y, or variation in Y may cause variation in X, or variation in some other factor may affect both X and Y.
- The third common use of linear regression is estimating the value of one variable corresponding to a particular value of the other variable.

CORRELATION

Correlation is a statistical technique that can show whether and how strongly pairs of variables are related.

- **Correlation** is a statistical measure that indicates the extent to which two or more variables fluctuate together.
- A **positive correlation** indicates the extent to which those variables increase or decrease in parallel.
- A **negative correlation** indicates the extent to which one variable increases as the other decreases.

PURPOSE

- **Correlation analysis** is a method of statistical evaluation used to study the strength of a relationship between two, numerically measured, continuous variables (e.g. height and weight).
- This particular type of **analysis** is useful when a researcher wants to establish if there are possible connections between variables.

Describing Relationships

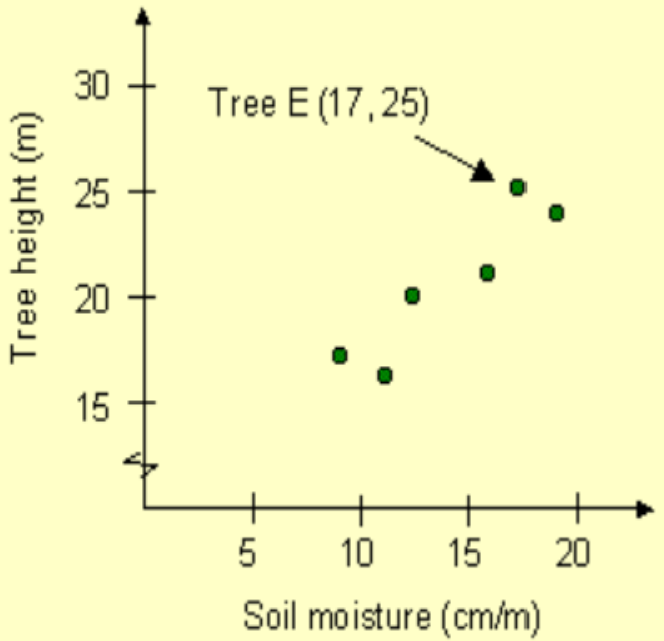
- There are two ways to summarize the relationship between two variables:
- Graphically using scatterplots
- Numerically using a correlation coefficient.

SCATTER PLOTS

- These graphs are an excellent way to get a sense of the relationship that exists between two variables. Scatterplots are usually based on the Cartesian coordinate system, where one variable is represented on the vertical or Y-axis and the other is represented on the horizontal or X-axis. The resulting pattern of points in the scatterplot shows the relationship between the variables, which can be described in terms of three criteria:
- *strength:* weak, moderate, strong
- *direction:* negative, positive, neutral
- *shape:* linear, curved, sinusoidal (s-shaped)

Example: The table below contains 6 observations collected for a botany project. The diagram on the right presents a scatterplot of the data. What kind of relationship exists between tree height and soil moisture?

Tree ID	Soil moisture (cm/m)	Tree height (m)
A	9	17
B	11	16
C	13	20
D	16	21
E	17	25
F	19	24



There is a positive relationship between soil moisture and tree height.

What kind of relationship would you expect between the following variables:

- number of cars per capita and carbon monoxide emissions?
- stream velocity and bed load particle size?
- elevation and average air temperature?

Correlation coefficient

- Methods of correlation summarize the relationship between two variables in a single number called the **correlation coefficient**.
- The correlation coefficient is usually given the symbol **r** and it ranges from -1 to +1.
- A correlation coefficient can be produced for ordinal, interval or ratio level variables, but has little meaning for variables which are measured on a scale which is no more than nominal.

Cont….

- A correlation coefficient quite close to 0, but either positive or negative, implies little or no relationship between the two variables.
- A correlation coefficient close to plus 1 means a positive relationship between the two variables, with increases in one of the variables being associated with increases in the other variable.
- A correlation coefficient close to -1 indicates a negative relationship between two variables, with an increase in one of the variables being associated with a decrease in the other variable.

- For ordinal scales, the correlation coefficient which is usually calculated is **Spearman's rho**.

- For interval or ratio level scales, the most commonly used correlation coefficient is **Pearson's** r, ordinarily referred to as simply the correlation coefficient.

Pearson correlation

- Pearson correlation evaluates the linear relationship between two continuous variables. A relationship is linear when a change in one variable is associated with a proportional change in the other variable.
- For example, you might use a Pearson correlation to evaluate whether increases in temperature at your production facility are associated with decreasing thickness of your chocolate coating.
- Pearson's coefficient measures the *linear relationship* between the two, i.e. how well a straight line describes the relationship between them.

Pearson correlation coefficient

$$r = \frac{\sum(X-\overline{X})(Y-\overline{Y})}{\sqrt{\sum(X-\overline{X})^2}\sqrt{(Y-\overline{Y})^2}}$$

Where, $\overline{X}$ = mean of X variable

$\overline{Y}$ = mean of Y variable

Spearman correlation

- The Spearman correlation evaluates the monotonic relationship between two continuous or ordinal variables. In a monotonic relationship, the variables tend to change together, but not necessarily at a constant rate. The Spearman correlation coefficient is based on the ranked values for each variable rather than the raw data.
- Spearman correlation is often used to evaluate relationships involving ordinal variables. For example, to evaluate whether the order in which employees complete a test exercise is related to the number of months they have been employed.
- Spearman's coefficient measures the *rank order* of the points. It does not care exactly where they are.

Spearman correlation coefficient

$$\rho = 1 - \frac{6\sum d^2}{n(n^2 - 1)}$$

THANK YOU

Disclaimer: The content displayed in the PPT has been taken from variety of different websites and book sources. This study material has been created for the academic benefits of the students alone and I do not seek any personal advantage out of it.