Statistical Analysis

- Descriptive Statistics
- Graphs and Charts
- Correlation and Regression Analysis

Descriptive Statistics, Graphs and Charts

- *Before starting with any advanced analysis, it is a good habbit to start with some **descriptive statistics** and **simple graphics**, to see what is going on in your data!*
- Some simple charts can be obtained, such as bar charts, pie charts and histograms. (A histogram is a graphical display of counts for ranges of data values.)

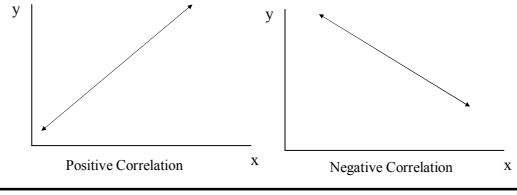
There are many statistical tools to measure whether two or more variables are associated with each other.

- **Correlation Analysis**
  - is used for describing relationship (positive or negative) between 2 variables.
- **Regression Analysis**
  - is used to model relationship between **dependent** and **independent variables**.
  - is an ***Estimation Technique***. (OLS)
  - in linear regression, the function is a linear straight line equation. ($Y = \alpha + Bx$)

Correlation Analysis

- The correlations table displays **Pearson correlation coefficients** and significance values.
- The values of the correlation coefficient range from -1 to 1.
- The sign of the correlation coefficient indicates the direction of the relationship (positive or negative).
- **The Correlation Matrix** - is a table that shows the *Pearson's r* scores for each combination pair for a table of two or more variables.

- A positive correlation exists - if the two variables move together: (As X increases Y also increases.)
- A negative correlation exists - if the two variables move in opposite directions: (As X increases Y decreases.)



Correlation Analysis

- The significance of each correlation coefficient is also displayed in the correlation table.
- The significance level (or p-value) is the probability of obtaining results as extreme as the one observed.
- If the significance level is very small (less than 0.05) then the correlation is significant and the two variables are linearly related.
- If the significance level is relatively large (for example, 0.50) then the correlation is not significant and the two variables are not linearly related.

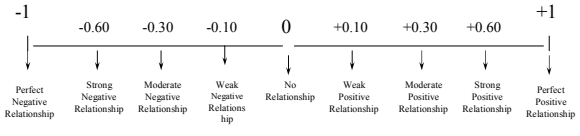
Correlation Analysis

- Hypothesis
  - H_0 There is no relationship between two variables
  - H_1 There is relationship between two variables

If p-value is below 0.05, we reject the null hypothesis (H_0) and accept the alternative (H_1).

- We need a way to measure the degree of correlation that exists between two or variables:
  - ***Pearson's Correlation Coefficient*** – is a measure often used to measure correlation is
- The value of the correlation coefficient varies between –1 and +1:
  - -1 indicates a perfect negative correlation
  - 0 indicates no relationship
  - +1 indicates a perfect positive correlation

The Correlation Coefficient Scale



The absolute value of the correlation coefficient indicates the strength, with larger absolute values indicating stronger relationships. The correlation coefficients on the main diagonal are always 1, because each variable has a perfect positive linear relationship with itself.

- e.g. if

  x: years of education possessed by a person

  y: the beginning salary of a person

  Results in

  r = 0.40

  This means that:

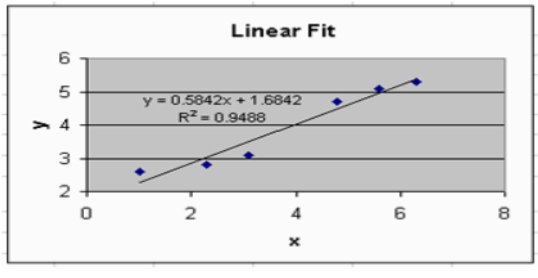
  there is a ***moderately positive correlation*** between the years of education and the beginning salary of a person.

Introduction to Regression Analysis

- Regression is an important tool researchers use to understand the relationship among two or more variables.
- Regression analysis is used to produce an equation that will predict a dependent variable using one more independent variable.
- OLS come in two forms:
  - Bi-variate Regression: $y = \alpha + \beta_1 x_1 + e$
  - Multiple Regression: $y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots \beta_i x_i + e$

Regression as a best fitting line

Let us begin with just two variables (Y and X). We refer to this case as simple regression.



Example: Assume that the price of a house (Y) depends on the size (X)

- Y=34,000+7X
  - The prediction equation is
    - The Price of the House is = 34000+7(Size)
    - Telling you that house price is predicted to increase 7 when size goes up by 1 unit.
    - If size were 5,000 square feet, this model says that the price of house should be $69,000.
- However, this model might be misleading becase there are other factors that may also affect the price of a house. Ex. Number of badrooms,...etc.
- There might be a source of error which is due to missing variables. Y=34,000+7X+e e is errors.

The Output Summary Indicates Several Interesting Points

The prediction Equation:

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots. \beta_i x_i + e$$

Y	: Dependent Variable
$X_1, X_2, X_3 \ldots$	: Independent Variable
β	: Coeffient or multipliers that describe the size of the effect the independent variables are having on your dependent variable

Coefficients: The size of the coeff. For each independent variable gives you the size of the effect that variable is having on your dependent variable, and the sign on the coefficient (+'ve or –'ve) gives you the direction of the effect.

•In regression with multiple independent variables, the coefficient of tells you how much the dependent variable is expected to increase (or decrease) when the independent variable increases (or decreases) by one , holding all the other independent variables constant.

- **R-Square** (Goodness of Fit)
  - How well the model explains (fits) the data?
  - Higher the R-square would suggest that the variation in the dependent variable that is predicted by independent variables. (i.e. Total variation of Y explained by X_1, X_2.....)
  - If R-square=91%
    - quite well, suggest that the models explains 91% of the variation. i.e. X1,X2,X3...variables explain 91% of the variability of the data
  - If R-square=13.7%
    - we expect 13.7% of the original variability, and left 86.3% residual variability.

To determine if a relationship exists between the independent variable(s) ($x_1, x_2, x_3 \ldots x_i$) and the dependent variable (y)

- A *Hypothesis Test* is conducted doing an
  - **F-test** on the independent variables as a whole,
  - **t-test** on each independent variable.

F-Test:

- Hypotheses test is conducted on the independent variables as a whole.
  - H_0 There is no relationship on the independent variables as a whole
  - H_1 There is relationship independent variables as a whole
  - Eg. If significant 0.0000 (or 0.05 at 5% significance level)
    - If p-value is below 0.05, we reject the null hypothesis (H_0) and accept the alternative (H_1).
    - The regression itself is also quite significant as the large value of F-ratio shows.

T- Statistics (t-test coefficients)

- Hypotheis test shows whether each variables are significant.
  - H_0 There is no relationship between each independent variable and dependent variable
  - H_1 There is relationship between each independent variable and dependent variable

  Eg. 0.000

  – At 5% significance level (or 95% confidence level)
    - if this value is greater than 0.05, then there is no evidence that the variable is significant. Hence no relation between dependent variable and independent variable.