The process of Econometric Analysis

- Econometric analysis consists mainly of:
- Estimating economic relationships from sample data in respect of:

→Existence of relationships
→Direction of the relationships
→Magnitude of the relationships.

- Prediction
- Four major elements of Econometric Model.

- Data: Collecting and Coding of sample data.
- Model specification: Two components of an Econometric Model:

Economic Model:

> Specifies the dependent and independent variables and the functional form of the relationships.
> Based on economic theory/informal intuition/observation

Statistical Model:

Specifies the statistical elements (particularly statistical properties of the random disturbance term).

- Estimation: Computing estimates of the unknown parameters using the sample data.
- Inference: Using the estimates to test hypotheses about the unknown population parameters

Statistical errors:

Two types of statistical errors:

State of Nature	Do not Reject H_0	Reject H_0
H_0	Correct Decision	Type I error
H_A	Type II error	Correct Decision

Types of data:

Cross-sectional data: Generated at one point of time across cross-sectional units.

Time-series data: Consist of repeated observations on a cross sectional unit over an interval of time.

Panel Data: Have both cross-sectional and time-series dimensions.

Types of Models:

- Bivariate Regression Model
- Multiple Regression Model
- Dynamic Model
- Non-linear Model.
- Simultaneous Equation Model
- Other Models (e.g. Logit, Probit, Tobit)

Model Specification

- Variables in the Equation(s) and their functional forms.
- A priori restriction on the parameters
- Stochastic Assumptions

Why Stochastic Component?

- Omission of variables from the functions
- Random behaviour of human beings
- Imperfect mathematical specification of the model
- Errors of aggregation
- Errors of measurement.

Stochastic Assumptions

- Normal distribution
- Zero Mean
- Constant variance
- Independence across observations
- Independence from the explanatory variable(s).

Major Econometric Problems

- Autocorrelation
- Heteroscedasticity
- Multicollinearity
- Endogenity
- Specification Bias.

Regression analysis

Regression analysis is concerned with the study of the dependence of one variable, ~~with~~ the dependent variable, on one or more other variables, the explanatory variables, with a view to estimating and/or predicting the (population) mean or average value of the former in terms of the known or fixed (in repeated sampling) values of the latter.

* Statistical v/s Deterministic Relationships

In statistical relationships among variables we deal with random or stochastic variables, that is, variables that have probability distributions.

In functional or deterministic dependency, on the other hand, we also deal with variables, but these variables are not random or stochastic. i.e. if the value of the variables is known, the value of the function can be found out with absolute certainty.

* Regression v/s Causation

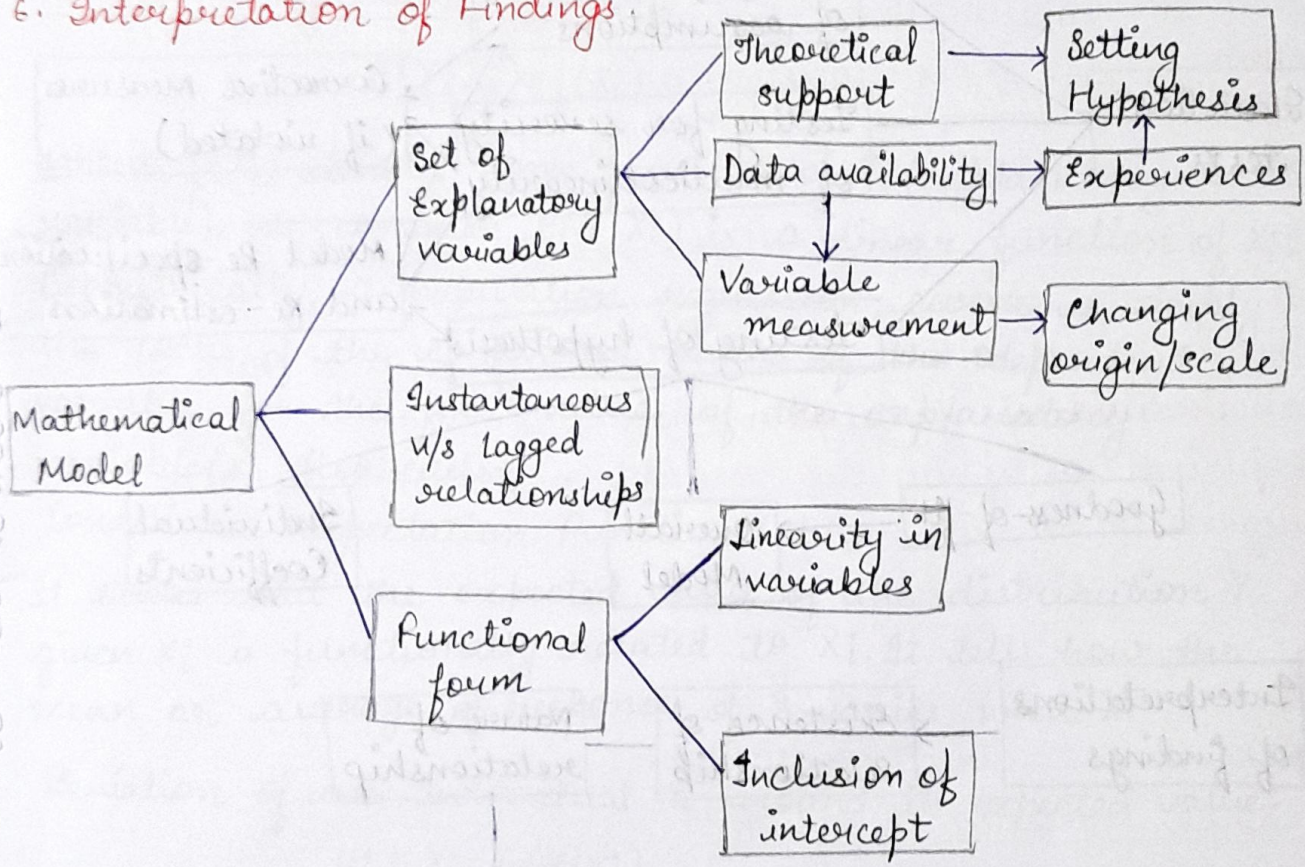
- A statistical relationship in itself cannot logically imply causation.
- To ascribe causality, one must appeal to a priori or theoretical considerations.

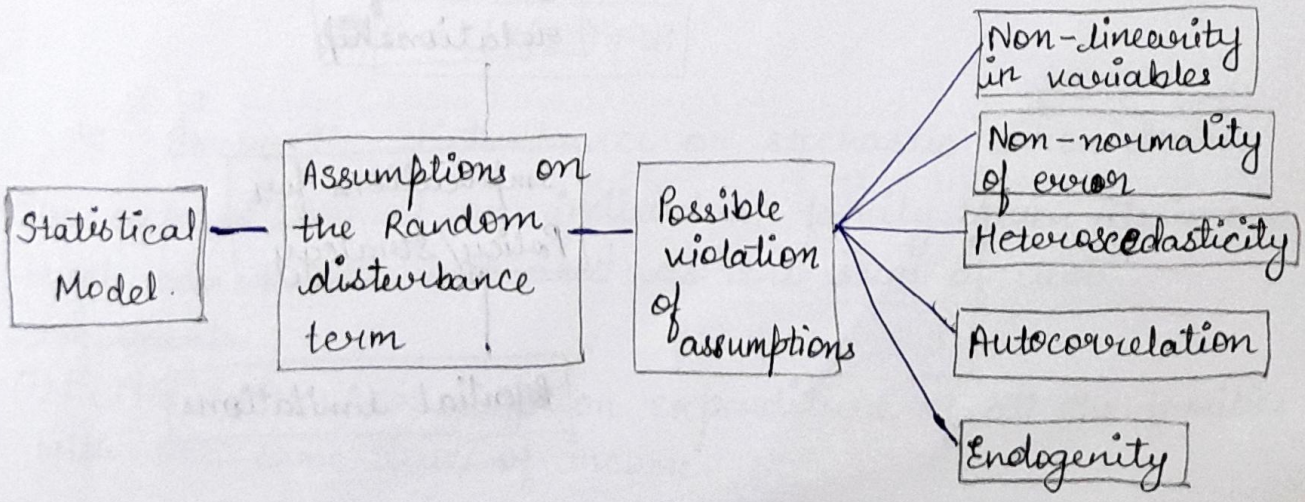
* Regression v/s Correlation

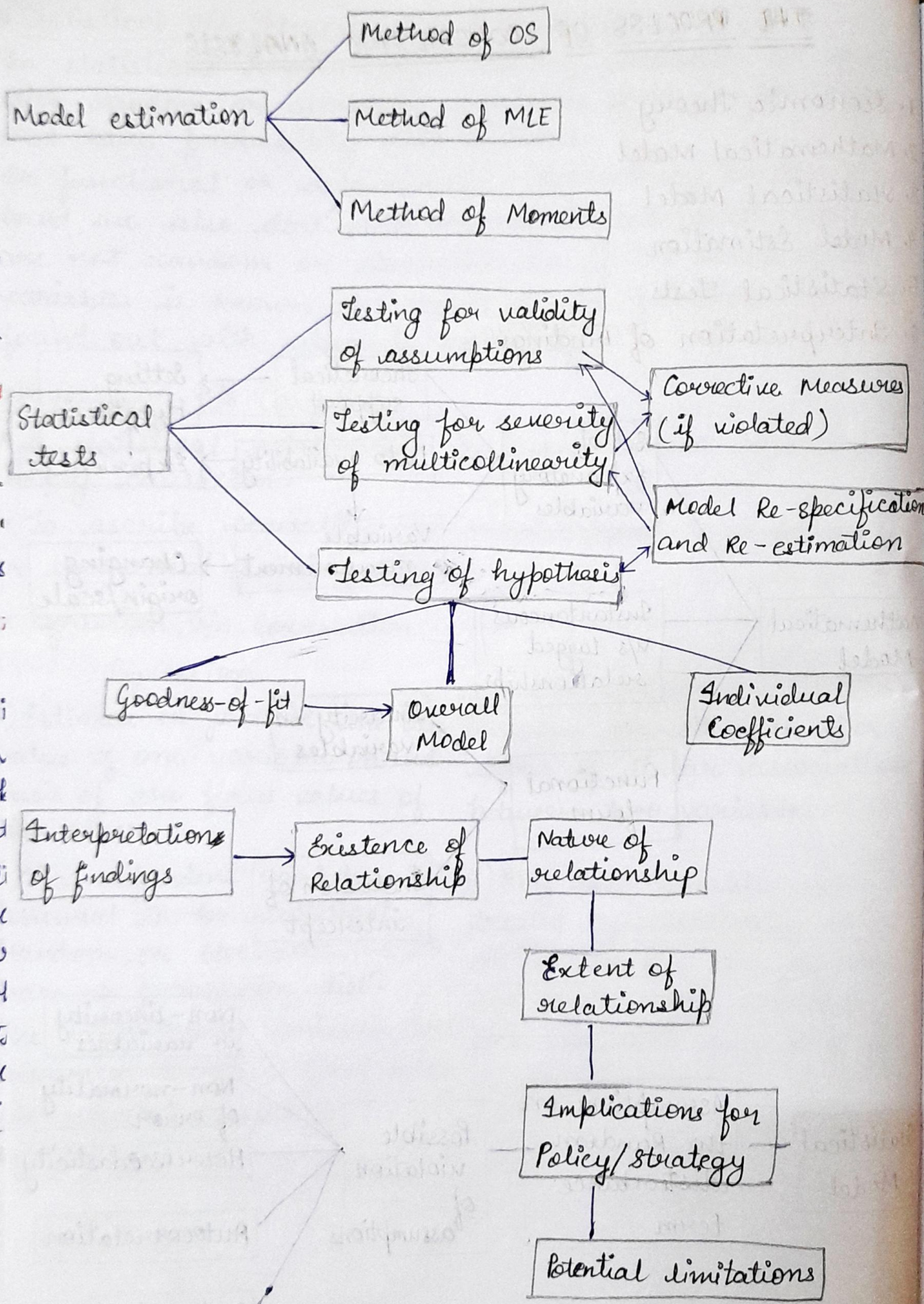
Regression	Correlation
i) Estimate or predict the avg. value of one variable on the basis of the fixed values of the other.	i) Measure the strength or degree of linear association between two variables.
ii) The dependent variable is assumed to be statistical, random, or stochastic, i.e. to have a probability dist^n. The explanatory variables, are assumed to have fixed values (in repeated sampling).	ii) Any two variables are treated symmetrically, no distinction b/w the dependent and explanatory variables. Both variables are assumed to be random.

THE PROCESS OF ECONOMETRIC ANALYSIS

1. Economic Theory
2. Mathematical Model
3. Statistical Model
4. Model Estimation
5. Statistical Tests
6. Interpretation of Findings






Method of OS
Model estimation
Method of MLE
Method of Moments
Testing for validity of assumptions
Statistical tests
Testing for severity of multicollinearity
Corrective Measures (if violated)
Model Re-specification and Re-estimation
Testing of hypothesis
Goodness-of-fit
Overall Model
Individual Coefficients
Interpretation of findings
Existence of Relationship
Nature of relationship
Extent of relationship
Implications for Policy/strategy
Potential limitations

BIVARIATE REGRESSION MODEL

$$Y_i = \alpha + \beta X_i + u_i$$

Y = Dependent Variable (e.g. Quantity Demanded)

X = Independent variable (e.g. Price of the commodity)

u = Random disturbance term

★Population Regression function

$$E(Y|X_i) = f(X_i)$$

where $f(X_i)$ denotes some function of X_i (explanatory variable). We consider $E(Y|X_i)$ is a linear function of X_i. Geometrically, a population regression curve is simply the locus of the conditional means of the dependent variable for the fixed values of the explanatory variable(s). Also called

Conditional Expectation Function (CEF)

It states that the expected value of the distribution Y given X_i is functionally related to X_i. It tells how the mean or average response of Y varies with X.

Deviation of an Individual Y_i around its expected value:

$$u_i = Y_i - E(Y|X_i)$$

$$\boxed{Y_i = E(Y|X_i) + u_i}$$

u_i → Stochastic disturbance or stochastic error term.

The expenditure of an individual family, given its income level, can be the expressed as the sum of two components.

(i) $E(Y|X_i)$: Mean consumption expenditure of all the families with the same level of income.

This component is known as the systematic, or deterministic component.

(ii) u_i, which is the random, or nonsystematic component.

$$E(Y_i|X_i) = E(E(Y|X_i)) + E(u_i|X_i)$$
$$= E(Y|X_i) + E(u_i|X_i)$$
$$\Rightarrow \boxed{E(u_i|X_i) = 0}$$

Thus, the assumption that the regression line passes through the conditional means of Y implies that the conditional mean values of u_i (conditional upon the given X's) are zero.

The stochastic specification has the advantage that it clearly shows that there are other variables besides income that affect consumption expenditure and that an individual family's consumption cannot be fully explained only by the variable(s) included in the regression model.

★ The Method of Ordinary least squares

$$Y_i = \alpha + \beta X_i + u_i$$

- PRF is not directly observable.

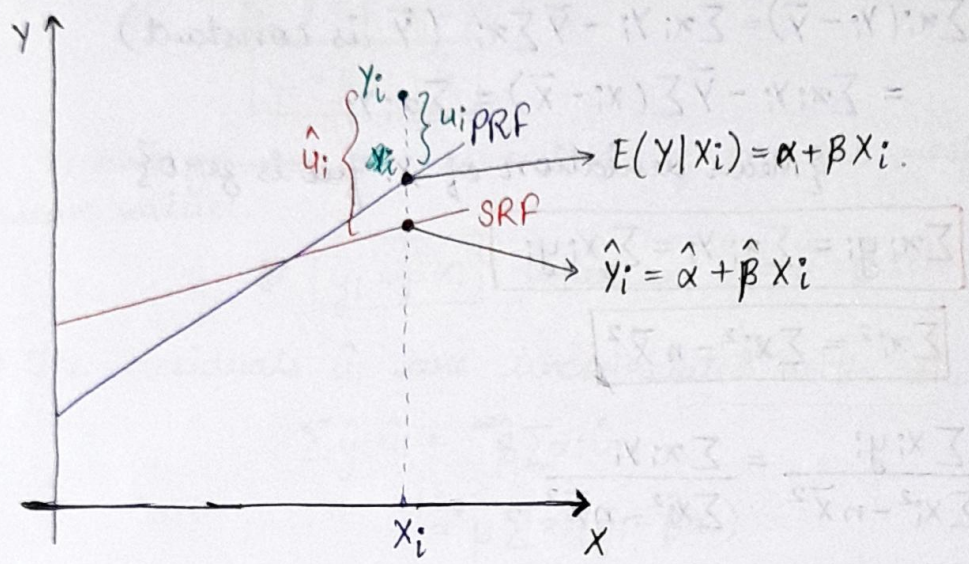
Sample Regression Function

$$Y_i = \hat{\alpha} + \hat{\beta} X_i + \hat{u}_i$$
$$Y_i = \hat{Y}_i + \hat{u}_i$$

where $\hat{Y}_i$ is the estimated (conditional mean) value of Y_i.

$$\hat{u}_i = Y_i - \hat{Y}_i$$
$$= Y_i - \hat{\alpha} - \hat{\beta} X_i$$

$$\boxed{\hat{u}_i = Y_i - \hat{\alpha} - \hat{\beta} X_i}$$



Model: $Y_i = \alpha + \beta X_i + u_i$

Minimize: $\sum \hat{u}_i^2 = \sum (Y_i - \hat{\alpha} - \hat{\beta} X_i)^2$ with respect to $\hat{\alpha}$ & $\hat{\beta}$

Differentiating both sides w.r.t. $\hat{\alpha}, \hat{\beta}$

$$\frac{\partial \sum \hat{u}_i^2}{\partial \alpha} = -2 \sum (Y_i - \hat{\alpha} - \hat{\beta} X_i) = 0 \qquad \boxed{\sum \hat{u}_i = 0} \quad \text{—— (1)}$$

$$\frac{\partial \sum u_i^2}{\partial \beta} = -2 \sum (Y_i - \hat{\alpha} - \hat{\beta} X_i)(X_i) = 0 \qquad \boxed{\sum \hat{u}_i X_i = 0} \quad \text{—— (2)}$$

(1) and (2) are called Normal Equations

Solving (1) and (2), for $\hat{\alpha}$ and $\hat{\beta}$ we get

$$\boxed{\hat{\beta} = \frac{\sum x_i y_i}{\sum x_i^2} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2} = \frac{cov(X, Y)}{var(X)}}$$

$$\boxed{\hat{\alpha} = \bar{Y} - \hat{\beta} \bar{X}}$$

OLS Estimators of $\hat{\alpha}$ and $\hat{\beta}$.

$$\sum x_i^2 = \sum (X_i - \bar{X})^2 = \sum X_i^2 - 2\bar{X} \sum X_i + \sum \bar{X}^2 = \sum X_i^2 - 2\bar{X} \sum X_i + n\bar{X}^2$$

Since $\bar{X}$ is constant.

$$\sum X_i = n\bar{X} \text{ and } \sum \bar{X}^2 = n\bar{X}^2$$

$$\boxed{\sum x_i^2 = \sum X_i^2 - n\bar{X}^2}$$

$$\sum x_i y_i = \sum x_i(Y_i - \bar{Y}) = \sum x_i Y_i - \bar{Y}\sum x_i \quad (\bar{Y} \text{ is constant})$$

$$= \sum x_i Y_i - \bar{Y}\sum(X_i - \bar{X}) = \sum x_i Y_i$$

{Mean deviation of X. ~~for~~ is zero}

Similarly, $\boxed{\sum x_i y_i = \sum x_i Y_i = \sum X_i y_i}$

$$\boxed{\sum x_i^2 = \sum X_i^2 - n\bar{X}^2}$$

$$\hat{\beta} = \frac{\sum x_i y_i}{\sum X_i^2 - n\bar{X}^2} = \frac{\sum x_i Y_i}{\sum X_i^2 - n\bar{X}^2}$$

*OLS Estimators are Point Estimators.

Properties of Sample Regression line

- It passes through the sample means of X and Y.

$$\bar{Y} = \hat{\alpha} + \hat{\beta}\bar{X}$$

- The mean value of the estimated $Y = \hat{Y}_i$ is equal to the mean value of the actual Y for

$$\hat{Y}_i = \hat{\alpha} + \hat{\beta}X_i$$

$$= (\bar{Y} - \hat{\beta}\bar{X}) + \hat{\beta}X_i$$

$$= \bar{Y} + \hat{\beta}(X_i - \bar{X})$$

$$\sum \hat{Y}_i = \sum \bar{Y} + \hat{\beta}\sum(X_i - \bar{X})$$

$$\Rightarrow \sum \hat{Y}_i = \sum \bar{Y} \qquad \boxed{\bar{\hat{Y}} = \bar{Y}}$$

- The mean value of the residuals $\hat{u}_i$ is zero.

$$-2\sum(Y_i - \hat{\alpha} - \hat{\beta}X_i) = 0$$

$$\hat{u}_i = Y_i - \hat{\alpha} - \hat{\beta}X_i \qquad \sum \hat{u}_i = 0$$

$$Y_i = \hat{\alpha} + \hat{\beta}X_i + \hat{u}_i$$

$$\sum Y_i = n\hat{\alpha} + \hat{\beta}\sum X_i + \sum \hat{u}_i \quad (\hookrightarrow 0)$$

$$\boxed{\bar{Y} = \hat{\alpha} + \hat{\beta}\bar{X}}$$

$$Y_i - \bar{Y} = \hat{\beta}(X_i - \bar{X}) + \hat{u}_i$$

$$\boxed{y_i = \hat{\beta} x_i + \hat{u}_i}$$

x_i and y_i are deviations from the respective (sample) mean values.

$$\boxed{\hat{y}_i = \hat{\beta} x_i} \quad \text{whereas} \quad \hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i$$

- The residuals $\hat{u}_i$ are uncorrelated with the predicted Y_i.

$$\sum \hat{y}_i \hat{u}_i = \hat{\beta} \sum x_i \hat{u}_i$$

$$= \hat{\beta} \sum x_i (y_i - \hat{\beta} x_i)$$

$$= \hat{\beta} \sum x_i y_i - \hat{\beta}^2 \sum x_i^2 \quad \left\{ \hat{\beta} = \sum x_i y_i / \sum x_i^2 \right\}$$

$$= \hat{\beta}^2 \sum x_i^2 - \hat{\beta}^2 \sum x_i^2 = 0.$$

- The residuals $\hat{u}_i$ are uncorrelated with X_i; i.e.

$$\sum \hat{u}_i X_i = 0$$

$$\Rightarrow \sum \hat{u}_i (x_i + \bar{X}) = \sum \hat{u}_i x_i + \bar{X} \sum \hat{u}_i = \sum \hat{u}_i x_i = 0$$

$$\sum \hat{u}_i X_i = \sum \hat{u}_i x_i = 0$$

$$\Rightarrow \sum \hat{u}_i \hat{Y}_i = \sum \hat{u}_i (\hat{\alpha} + \hat{\beta} X_i) = \hat{\alpha} \sum \hat{u}_i + \hat{\beta} \sum \hat{u}_i X_i = 0$$

$$\sum \hat{u}_i \hat{Y}_i = 0$$

$$\Rightarrow \sum \hat{u}_i \hat{y}_i = \sum \hat{u}_i (\hat{Y}_i - \bar{Y}) = \sum \hat{u}_i \hat{Y}_i - \bar{Y} \sum \hat{u}_i = \sum \hat{u}_i \hat{Y}_i = 0$$

$$\sum \hat{u}_i \hat{y}_i = 0$$

$$\Rightarrow \sum \hat{u}_i y_i = \sum \hat{u}_i (\hat{y}_i + \hat{u}_i) = \sum \hat{u}_i \hat{y}_i + \sum \hat{u}_i^2 = \sum \hat{u}_i^2 = RSS.$$

$$\sum \hat{u}_i y_i = \sum \hat{u}_i^2 = RSS.$$

★ Assumptions of OLS

1) Linear Regression Model.

The Regression Model is linear in the parameters.

$$Y_i = \alpha + \beta X_i + u_i$$

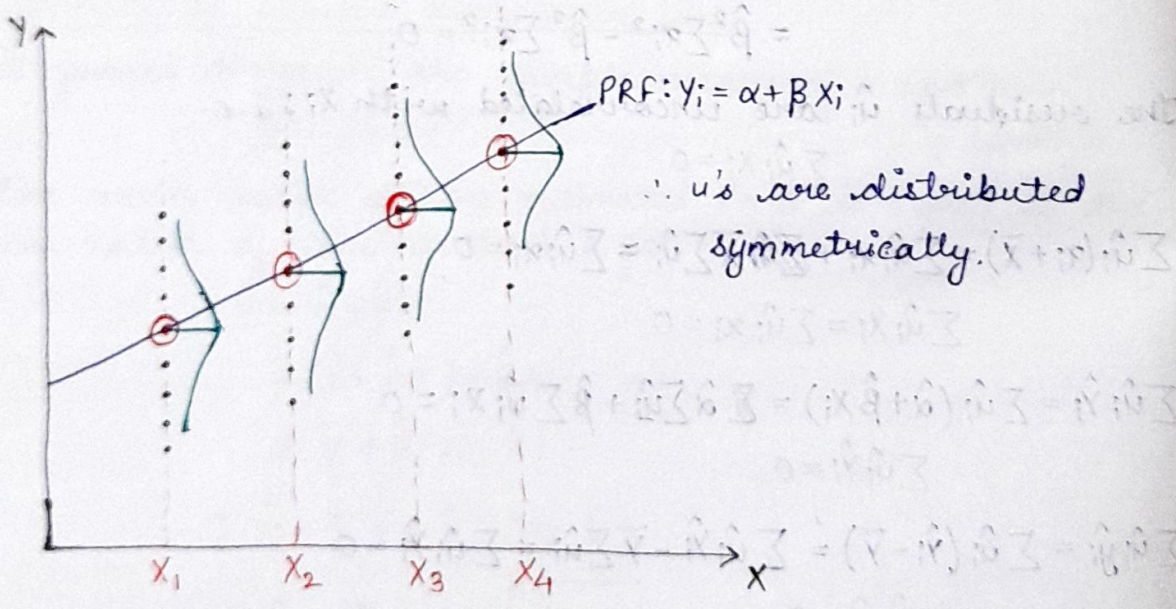
2) X values are fixed in repeated sampling.

Values taken by the regressor X are considered fixed in repeated samples. X is assumed non-stochastic.

3) Zero mean value of disturbance u_i

Given the value of X, the mean, or expected value of the random disturbance term u_i is zero. Technically, the conditional mean value of u_i is zero.

$$E(u_i|X_i) = 0$$



The positive u_i values cancel out the negative u_i values so that their average or mean effect on Y is zero.

4) Heteroscedasticity or equal variance of u_i

Given the value of X, the variance of u_i is the same for all observations, i.e., the conditional variances of u_i are identical.

$$\text{var}(u_i|X_i) = E[u_i - E(u_i|X_i)]^2$$

$$= E(u_i^2|X_i)$$

$$E[u_i - E(u_i|X_i)]^2 = E(u_i^2) - 2E(u_i)E(u_i|X_i) + E(u_i^2|X_i)$$

↳0

The Y populations corresponding to ~~diff~~ various X values have the same variance. The variation around the regression line is the same across the X values; it neither increases or decreases as X varies.

Example: Let Y represent weekly consumption expenditure, and X weekly ~~consumption~~ income. As income increases the average consumption expenditure also increases. If the variance also increases with increase in income. In other words, richer families on the average consume more than poorer families, but there is ~~always~~ also more variability in the consumption expenditure of the former.

If, let

$$var(u|X_1) < var(u|X_2) < \cdots < Var(u|X_i)$$

Therefore, the likelihood is that the Y observations coming from the population with $X = X_1$, would be closer to the PRF than those coming from populations corresponding to $X = X_2$, $X = X_3$, and so-on.

- Not all Y values corresponding to the various X's will be equally reliable, reliability being judged by how closely or distantly the Y values are distributed around their means, that is, the points on the PRF.

Assumption 4 also implies,

$$var(Y_i|X_i) = \sigma^2.$$

In case of heteroscedascity, if we prefer to sample from those Y populations that are closer to their mean than those that are widely spread might restrict the variation we obtain X values.

5) No autocorrelation between the disturbances.

Given any two values X_i and X_j $(i \neq j)$, the correlation b/w any two u_i and u_j $(i \neq j)$ is zero.

$$cov(u_i, u_j|X_i, X_j) = E\{[u_i - E(u_i)]|X_i\}\{[u_j - E(u_j)]|X_j\}$$
$$= E(u_i|X_i)(u_j|X_j)$$

$$\boxed{cov(u_i, u_j|X_i, X_j) = 0}$$ where i & j are two observations

Suppose in our PRF

$Y_t = \alpha + \beta X_t + u_t$, u_t and u_{t-1} are positively correlated.

Then Y_t depends not only of X_t but also on u_{t-1} for u_{t-1} to some extent determines u_t.

6) Zero covariance between u_i and X_i, or $E(u_i X_i) = 0$

$$\begin{aligned} \text{cov}(u_i, X_i) &= E[u_i - E(u_i)][X_i - E(X_i)] \\ &= E[u_i(X_i - E(X_i))] \\ &= E(u_i X_i) - E(X_i)E(u_i) \end{aligned}$$

$X_i \rightarrow$ non-stochastic

$= E(u_i X_i)$ since $E(u_i) = 0$

$= 0$ by assumption.

The rationale for this assumption is as follows: When we express the PRF, we assumed that X and u (which may represent the influence of all the omitted variables) have seperate (and additive) influence on Y. But if X and u are correlated, it is not possible to assess their individual effects on Y.

Regression theory holds true even if the X's are stochastic or random, provided they are independent or atleast uncorrelated with the disturbances u_i.

7) The number of observations n must be greater than the number of parameters to be estimated.

$n >$ no. of explanatory variables.

8) Variability in X values. $\text{var}(X) \rightarrow$ finite positive number.

9) The regression model is correctly specified. There is no specification bias or error in the model used in empirical analysis.

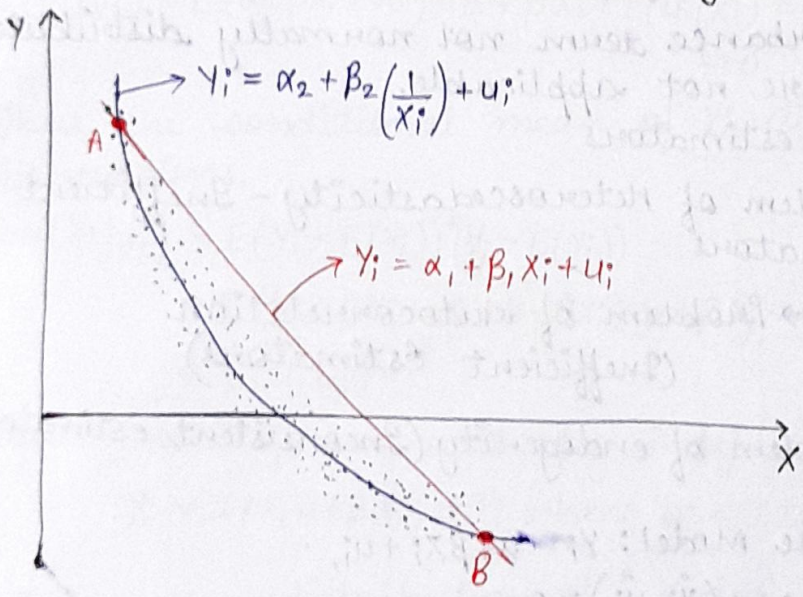
Example: Two models to depict the underlying relationship between the rate of change of money wages and the unemployment rate:

$$Y_i = \alpha_1 + \beta_1 X_i + u_i \quad , \quad \beta_1 < 0 \quad \text{—— (1)}$$

$$Y_i = \alpha_2 + \beta_2 \left(\frac{1}{X_i}\right) + u_i \quad , \quad \beta_2 > 0 \quad \text{—— (2)}$$

Y_i = rate of change of money wages

X_i = rate ~~of change~~ of unemployment.



Between points A & B, for any given X_i the model (1) is going to overestimate the true mean value of Y, whereas to the left of A (or right of B) it is going to underestimate (or overestimate, in absolute terms) the true mean value of Y.

10) There is no perfect multicollinearity. That is, there is no perfect linear relationships among the explanatory variables.

Assumptions on Random Disturbance Term (Summary)

- Normal distribution of the random disturbance term
- $E(u_i|X_i)=0$
- $E(u_i^2|X_i)=\sigma^2$ (constant for all i,j)
- $E(u_iu_j|_{X_iX_j})=0 \ \forall\ i\neq j$
- $E(X_iu_i)=0 \ \forall\ i$

Consequence of Violation of Assumptions

- The random disturbance term not normally distributed
  ⇒ Statistical tests are not applicable.
- $E(u_i|X_i)=0$ ⇒ Biased estimators
- $E(u_i^2|X_i)=\sigma^2$ ⇒ Problem of Heteroscedasticity - Inefficient estimators
- $E(u_iu_j|_{X_iX_j})\neq 0 \ \forall\ i\neq j$ ⇒ Problem of Autocorrelation (Inefficient Estimators)
- $E(X_iu_i)\neq 0 \ \forall\ i$ ⇒ Problem of endogenity (Inconsistent estimato

1. Prove that for the Model: $Y_i=\alpha+\beta X_i+u_i$,
   $cov(X_i,\hat{u}_i)=0$ and $cov(\hat{Y}_i,\hat{u}_i)=0$.

Proof: $cov(X_i,\hat{u}_i)=\frac{1}{n}\sum(X_i-\bar{X})(\hat{u}_i-\bar{\hat{u}}) \quad \{\bar{\hat{u}}=0\}$

$$=\frac{1}{n}\sum(X_i-\bar{X})\hat{u}_i=\frac{1}{n}\sum X_i\hat{u}_i-\bar{X}\sum\hat{u}_i=0$$

($\sum X_i\hat{u}_i \to 0$, $\bar{X}\sum\hat{u}_i \to 0$)

$$cov(\hat{Y}_i,\hat{u}_i)=\frac{1}{n}\sum(\hat{Y}_i-\bar{\hat{Y}})(\hat{u}_i-\bar{\hat{u}})=\frac{1}{n}\sum(\hat{Y}_i-\bar{\hat{Y}})\hat{u}_i$$

$$=\frac{1}{n}\sum\hat{Y}_i\hat{u}_i-\bar{\hat{Y}}\sum\hat{u}_i=0$$

($\sum\hat{Y}_i\hat{u}_i \to 0$, $\bar{\hat{Y}}\sum\hat{u}_i \to 0$)

$$\boxed{cov(X_i,\hat{u}_i)=0}$$

$$\boxed{cov(\hat{Y}_i,\hat{u}_i)=0}$$

2) For the model $Y_i = \alpha + \beta X_i + u_i$

if $u_i \sim IIN(0, \sigma^2)$

$Y_i \sim IN(\alpha + \beta X_i, \sigma^2)$

Proof:

Since Y_i is a linear transformation of u_i with X_i being non-stochastic, it will follow normal distribution when u_i does so.

$$E(Y_i|_{X_i}) = E(\alpha + \beta X_i + u_i) = \alpha + \beta X_i + \underbrace{E(u_i|X_i)}_{\hookrightarrow 0} = \alpha + \beta X_i$$

Thus, the conditional mean of Y_i differs across the observations.

$$\begin{aligned} var(Y_i|_{X_i}) &= E(Y_i - E(Y_i))(Y_j - E(Y_j)) \\ &= E((\alpha + \beta X_i + u_i - \alpha - \beta X_i)(\alpha + \beta X_j + u_j - \alpha - \beta X_j)) \\ &= E(u_i u_j) = 0 \end{aligned}$$

Thus, Y_i and Y_j are mutually independent.

$Y_i \sim IN(\alpha + \beta X_i, \sigma^2)$ when $u_i \sim IIN(0, \sigma^2)$.

3) Derive the distribution of $\hat{\alpha}$ and $\hat{\beta}$.

$$\hat{\beta} = \frac{\sum x_i y_i}{\sum x_i^2} = \frac{\sum x_i (Y_i - \bar{Y})}{\sum x_i^2} = \frac{\sum x_i Y_i}{\sum x_i^2} - \bar{Y}\underbrace{\frac{\sum x_i}{\sum x_i^2}}_{\hookrightarrow 0} = \sum k_i Y_i$$

$$k_i = \frac{x_i}{\sum x_i^2}$$

(i) $\sum k_i = \sum \left(\frac{x_i}{\sum x_i^2}\right) = \frac{\sum x_i}{\sum x_i^2} = 0$

(ii) $\sum k_i X_i = \sum \left(\frac{x_i(x_i + \bar{X})}{\sum x_i^2}\right) = \frac{\sum x_i^2}{\sum x_i^2} + \bar{X}\underbrace{\frac{\sum x_i}{\sum x_i^2}}_{\hookrightarrow 0} = 1$

$\therefore \hat{\beta}$ is a linear function of Y_i

$$\hat{\beta} = \sum k_i Y_i \quad k_i = \frac{x_i}{\sum x_i^2}$$

Since Y_i is normally distributed, so is $\hat{\beta}$.

$$\hat{\beta} = \sum k_i Y_i = \sum k_i(\alpha + \beta X_i + u_i) = \alpha \sum k_i + \beta \sum k_i X_i + \sum k_i u_i$$

$$= \beta + \sum k_i u_i$$

$$\Rightarrow E(\hat{\beta}) = E(\beta + \sum k_i u_i) = \beta + \sum k_i E(u_i) = \beta. \quad (E(u_i) \downarrow 0)$$

$$\Rightarrow \text{var}(\hat{\beta}) = E(\hat{\beta} - E(\hat{\beta}))^2 = E(\hat{\beta} - \beta)^2 = E(\sum k_i u_i)^2$$

$$= E(\sum k_i^2 u_i^2 + 2 \sum k_i k_j u_i u_j) \quad \forall i \neq j$$

$$\Rightarrow \text{var}(\hat{\beta}) = \sum k_i^2 E(u_i^2) + 2 \sum k_i k_j E(u_i u_j) \quad (E(u_i u_j) \to 0)$$

$$= \frac{\sum x_i^2 \sigma^2}{(\sum x_i^2)^2} = \frac{\sigma^2}{\sum x_i^2}$$

$$\boxed{E(\hat{\beta}) = \beta}$$

$$\boxed{\text{Var}(\hat{\beta}) = \frac{\sigma^2}{\sum x_i^2}}$$

$$\hat{\beta} \sim N\left(\beta, \frac{\sigma^2}{\sum x_i^2}\right)$$

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X} = \frac{\sum Y_i}{n} - \bar{X} \sum k_i Y_i = \sum \left(\frac{1}{n} - \bar{X} k_i\right) Y_i$$

$\therefore \hat{\alpha}$ is a linear function of Y_i. Since Y_i is normally distributed, so is $\hat{\alpha}$.

$$\hat{\alpha} = \sum \left[\frac{1}{n} - \bar{X} k_i\right](\alpha + \beta X_i + u_i)$$

$$\Rightarrow \hat{\alpha} = \sum \left[\frac{\alpha}{n} - \alpha \bar{X} k_i + \frac{\beta X_i}{n} - \beta \bar{X} k_i X_i\right] + \sum \left[\frac{u_i}{n} - \bar{X} u_i k_i\right]$$

$$\Rightarrow \hat{\alpha} = \alpha + \beta \bar{X} - \beta \bar{X} + \sum \left[\frac{1}{n} - \bar{X} k_i\right] u_i$$

$$\Rightarrow \hat{\alpha} = \alpha + \sum \left[\frac{1}{n} - \bar{X} k_i\right] u_i$$

$$E(\hat{\alpha}) = E\left\{\alpha + \sum \left[\frac{1}{n} - \bar{X} k_i\right] u_i\right\}$$

$$= \alpha + \sum \left[\frac{1}{n} - \bar{X}k_i\right] E(u_i) = \alpha$$

($E(u_i) \to 0$)

$$var(\hat{\alpha}) = E(\hat{\alpha} - E(\hat{\alpha}))^2 = E(\hat{\alpha} - \alpha)^2 = E\left\{\sum \left[\frac{1}{n} - \bar{X}k_i\right] u_i\right\}^2$$

$$\Rightarrow var(\hat{\alpha}) = E\left\{\sum \left[\frac{1}{n} - \bar{X}k_i\right]^2 u_i^2 + 2\sum \left[\frac{1}{n} - \bar{X}k_i\right]\left[\frac{1}{n} - \bar{X}k_j\right] u_i u_j\right\} \quad \forall i \neq j$$

$$\Rightarrow var(\hat{\alpha}) = \sum \left[\frac{1}{n} - \bar{X}k_i\right]^2 E(u_i^2) + 2\sum \left[\frac{1}{n} - \bar{X}k_i\right]\left[\frac{1}{n} - \bar{X}k_j\right] E(u_i u_j)$$

($E(u_i u_j) \to 0$)

$$\Rightarrow var(\hat{\alpha}) = \sigma^2 \sum \left(\frac{1}{n^2} + \bar{X}^2 k_i^2 - 2\bar{X}\frac{k_i}{n}\right) = \sigma^2 \left(\frac{1}{n} + \bar{X}^2 \frac{\sum x_i^2}{(\sum x_i^2)^2} - 2\bar{X}\frac{\sum k_i}{n}\right)$$

($\frac{\sum k_i}{n} \to 0$)

$$\Rightarrow var(\hat{\alpha}) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{X}^2}{\sum x_i^2}\right)$$

$$\boxed{E(\hat{\alpha}) = \alpha}$$

$$\boxed{var(\hat{\alpha}) = \sigma^2 \left[\frac{1}{n} + \frac{\bar{X}^2}{\sum x_i^2}\right]}$$

$$\hat{\alpha} \sim N\left[\alpha, \sigma^2 \left(\frac{1}{n} + \frac{\bar{X}^2}{\sum x_i^2}\right)\right]$$

* Features of the variances of $\hat{\alpha}$ & $\hat{\beta}$

1) The variance of $\hat{\beta}$ is directly proportional to σ^2 & inversely proportional to $\sum x_i^2$ i.e. variance of X. i.e. given σ^2, the larger the variation in the X values, the smaller the variance of β and hence the greater the precision with which β can be ~~calculated~~ estimation. Also given $\sum x_i^2$, the larger the variance σ^2, the larger the variance of β. As sample size n increases, the number of terms in the sum, $\sum x_i^2$ will increase, ∴ the precision with which β can be estimated also increases.

2) The variance of $\hat{\alpha}$ is directly proportional to σ^2 & $\bar{X}^2$ but inversely proportional to $\sum x_i^2$ and the sample size n.

8) Since $\hat{\alpha}$ & $\hat{\beta}$ are estimators, they will not only vary from sample to sample but in a given sample they are likely to be dependent on each other., this dependence being measured by the covariance b/w them.

$$\boxed{cov(\hat{\alpha}, \hat{\beta}) = -\bar{X}\, var(\hat{\beta})}$$

$$cov(\hat{\alpha}, \hat{\beta}) = -\bar{X}\left(\frac{\sigma^2}{\sum x_i^2}\right)$$

∴ If the slope coefficient β is overestimated (i.e. the slope is too steep), the intercept coefficient α will be underestimated (i.e. the intercept will be too small).

* Sum of Squares

- Explained Sum of Squares (ESS)

$$ESS = \sum(\hat{Y}_i - \bar{\hat{Y}})^2 = \sum \hat{y}_i^2$$

$Y_i = \alpha + \beta X_i + u_i$, $\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i$ and $\bar{\hat{Y}} = \hat{\alpha} + \hat{\beta}\bar{X}$

$$ESS = \sum(\hat{\alpha} + \hat{\beta} X_i - \hat{\alpha} - \hat{\beta}\bar{X})^2 = \hat{\beta}^2 \sum (X_i - \bar{X})^2 = \hat{\beta}^2 \sum x_i^2$$

$$\boxed{ESS = \hat{\beta}^2 \sum x_i^2 = \sum \hat{y}_i^2}$$

Degrees of freedom : $k-1$

k = Number of coefficients including the intercept.

- Residual Sum of Squares (RSS)

$$\boxed{RSS = \sum(Y_i - \hat{Y}_i)^2 = \sum \hat{u}_i^2}$$

Degrees of freedom : $n-k$.

- Total Sum of Squares (TSS)

$$\boxed{TSS = \sum(Y_i - \bar{\hat{Y}})^2 = \sum y_i^2}$$

$Y_i = \hat{Y}_i + \hat{u}_i$, $\sum \hat{u}_i = 0$, $\bar{Y} = \bar{\hat{Y}}$

$$\Rightarrow TSS = \sum(Y_i - \bar{Y})^2 = \sum\{(Y_i - \hat{Y}_i) + (\hat{Y}_i - \bar{\hat{Y}})\}^2$$

$$\Rightarrow TSS = \sum(Y_i - \hat{Y}_i)^2 + \sum(\hat{Y}_i - \bar{\hat{Y}})^2 + 2\sum(Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{\hat{Y}})$$

$$\Rightarrow TSS = \sum(Y_i - \hat{Y}_i)^2 + \sum(\hat{Y}_i - \bar{Y})^2 + 2\sum(\hat{Y}_i - \bar{Y})\hat{u}_i$$

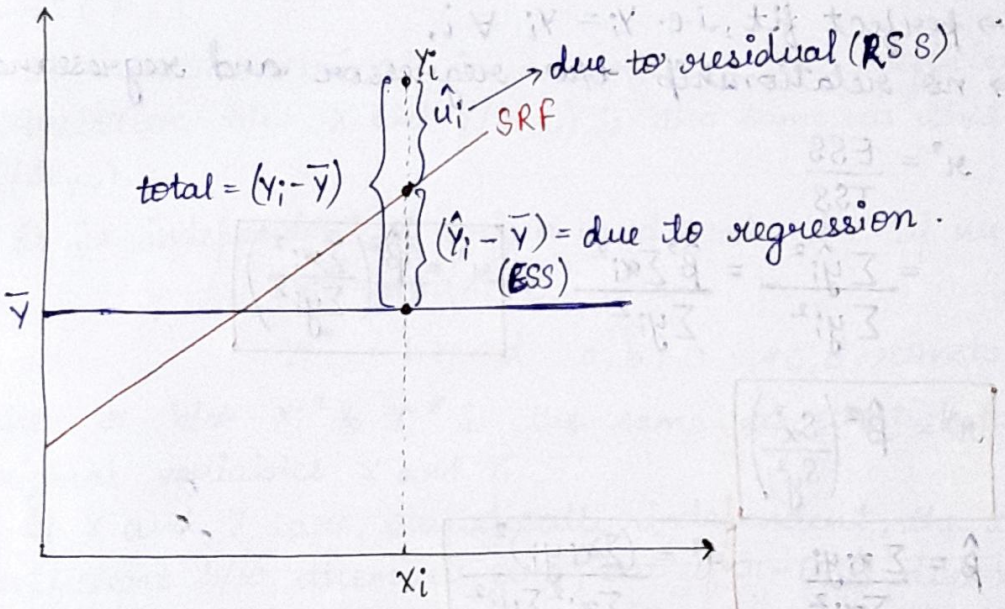
$$\Rightarrow TSS = \sum(Y_i - \hat{Y}_i)^2 + \sum(\hat{Y}_i - \bar{Y})^2 + 2\sum \hat{Y}_i\hat{u}_i - 2\bar{Y}\sum u_i$$

$\sum \hat{Y}_i\hat{u}_i \rightarrow 0$ and $\sum u_i \rightarrow 0$.

$$\Rightarrow TSS = \sum(Y_i - \hat{Y}_i)^2 + \sum(\hat{Y}_i - \bar{Y})^2$$

$$\boxed{TSS = RSS + ESS}$$

Degrees of freedom $= n - 1$.



$$1 = \frac{ESS}{TSS} + \frac{RSS}{TSS}$$

$$= \frac{\sum(\hat{Y}_i - \bar{Y})^2}{\sum(Y_i - \bar{Y})^2} + \frac{\sum(Y_i - \hat{Y}_i)^2}{\sum(Y_i - \bar{Y})^2}$$

$$= \frac{\sum(\hat{Y}_i - \bar{Y})^2}{\sum(Y_i - \bar{Y})^2} + \frac{\sum\hat{u}_i^2}{\sum(Y_i - \bar{Y})^2}$$

$$\boxed{r^2 = \frac{\sum(\hat{Y}_i - \bar{Y})^2}{\sum(Y_i - \bar{Y})^2}} \qquad \boxed{r^2 = 1 - \frac{\sum\hat{u}_i^2}{\sum(Y_i - \bar{Y})^2}}$$

$$r^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}$$

The quantity r^2 thus defined is known as the (sample) coefficient of determination and is the most commonly used measure of the goodness of fit of a regression line.
r^2 measures the proportion or percentage of the total variation in Y explained by the ~~mod~~ regression model.

Properties:

i) It is a nonegative quantity

ii) $0 \leq r^2 \leq 1$.

If $r^2 = 1 \Rightarrow$ perfect fit, i.e. $\hat{Y}_i = Y_i \ \forall \ i$,

If $r^2 = 0 \Rightarrow$ no relationship b/w regressor and regressand.

$$r^2 = \frac{ESS}{TSS}$$

$$= \frac{\sum \hat{y}_i^2}{\sum y_i^2} = \frac{\hat{\beta}^2 \sum x_i^2}{\sum y_i^2}$$

$$\boxed{r^2 = \hat{\beta}^2 \left(\frac{\sum x_i^2}{\sum y_i^2} \right)}$$

$$\boxed{r^2 = \hat{\beta}^2 \left(\frac{S_x^2}{S_y^2} \right)}$$

we know, $\hat{\beta} = \frac{\sum x_i y_i}{\sum x_i^2}$

$$\boxed{r^2 = \frac{(\sum x_i y_i)^2}{\sum x_i^2 \sum y_i^2}}$$

$$ESS = r^2 TSS$$

$$= r^2 \sum y_i^2$$

$$\boxed{\begin{aligned} ESS &= r^2 \sum y_i^2 \\ RSS &= (1 - r^2) \sum y_i^2 \end{aligned}}$$

$$RSS = TSS - ESS$$

$$= TSS \left[1 - \frac{ESS}{TSS} \right]$$

$$= TSS[1 - r^2] = \sum y_i^2 (1 - r^2)$$

Coefficient of correlation

$$r = \pm \sqrt{r^2}$$

$$r = \frac{\sum x_i y_i}{\sqrt{\sum x_i^2 \sum y_i^2}}$$

$$r = \frac{n\sum x_i y_i - (\sum x_i)(\sum y_i)}{\sqrt{\left[n\sum x_i^2 - (\sum x_i)^2\right]\left[n\sum y_i^2 - (\sum y_i)^2\right]}}$$

Sample Correlation Coefficient.

Properties of r

1. Can be +ve or -ve, depending on the numerator which measures the sample covariation of two variables.
2. $-1 \leq r \leq 1$.
3. It is symmetrical in nature; i.e. the coefficient of correlation b/w X and Y (r_{XY}) is the same as that b/w Y and X (r_{YX}).
4. It is independent of origin and scale; i.e. if we define

$$x_i^* = ax_i + c$$

$$y_i^* = by_i + d \quad a, b > 0, \text{ & } c, d \rightarrow \text{constants}$$

then r b/w x_i^* & y_i^* is the same as that between the original variables X and Y.

5. If X and Y are statistically independent, the correlation coefficient b/w them is zero; but if $r = 0$, it does not mean that two variables are independent. In other words, zero correlation does not necessarily imply independence.
6. It is a measure of linear association or linear dependence only; it has no meaning for describing nonlinear relations.
7. Although it is a measure of linear association between two variables, it does not necessarily imply any cause-and-effect relationship.

$$r^2 = \frac{\left[\sum (Y_i - \bar{Y})(\hat{Y}_i - \bar{Y})\right]^2}{\sum (Y_i - \bar{Y})^2 \sum (\hat{Y}_i - \bar{Y})^2}$$

$$r^2 = \frac{(\sum y_i \hat{y}_i)^2}{\sum y_i^2 \sum \hat{y}_i^2}$$

- Mean of ESS

$$MESS = \frac{ESS}{df_{ESS}} = \frac{\sum(\hat{y}_i - \bar{y})^2}{k-1}$$

- Mean of RSS

$$MRSS = \frac{RSS}{df_{RSS}} = \frac{\sum(y_i - \hat{y}_i)^2}{n-k}$$

- Mean of TSS

$$MTSS = \frac{TSS}{df_{TSS}} = \frac{\sum(y_i - \bar{y})^2}{n-1}$$

1. Prove that for the model $y_i = \alpha + \beta x_i + u_i$, $R^2 = r^2$ where r is the pair-wise correlation coefficient between X and Y.

For the model $y_i = \alpha + \beta x_i + u_i$

$$R^2 = \frac{ESS}{TSS} = \frac{\hat{\beta}^2 \sum x_i^2}{\sum y_i^2} \quad \hat{\beta} = \frac{\sum x_i y_i}{\sum x_i^2}$$

$$R^2 = \frac{\hat{\beta}^2 \sum x_i^2}{\sum y_i^2} = \frac{(\sum x_i y_i)^2}{(\sum x_i^2)^2} \frac{\sum x_i^2}{\sum y_i^2} = \frac{(\sum x_i y_i)^2}{\sum x_i^2 \sum y_i^2}$$

$$= \left(\frac{\sum x_i y_i}{\sqrt{\sum x_i^2 \sum y_i^2}}\right)^2 = \left[\frac{cov(X,Y)}{\sigma_X \sigma_Y}\right]^2 = r^2$$

2. Prove that, for the model $y_i = \alpha + \beta x_i + u_i$,

$$|r_{Y\hat{Y}}| = |r_{XY}|$$

where r is the pair-wise correlation coefficient between X and Y.

$$|r_{Y\hat{Y}}| = \frac{cov(Y,\hat{Y})}{\sigma_Y \sigma_{\hat{Y}}} = \frac{\sum(y_i - \bar{y})(\hat{y}_i - \bar{y})}{\sqrt{\sum(\hat{y}_i - \bar{y})^2 \sum(y_i - \bar{y})^2}}$$

~~$Y_i = \alpha + \beta X_i +$~~ ~~$\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i$~~

$y_i = \alpha + \beta x_i + u_i$ $\quad \hat{y}_i = \hat{\alpha} + \hat{\beta} x_i$ $\quad \bar{\hat{y}} = \hat{\alpha} + \hat{\beta} \bar{x}$

~~$r_{XY} = \sum(x_i - \bar{x})(y_i - \bar{y})$~~.

$$r_{y\hat{y}} = \frac{\sum(y_i-\bar{y})(\hat{\alpha}+\hat{\beta}x_i-\hat{\alpha}-\hat{\beta}\bar{x})}{\sqrt{\sum(y_i-\bar{y})^2}\sqrt{\sum(\hat{\alpha}+\hat{\beta}x_i-\hat{\alpha}-\hat{\beta}\bar{x})^2}}$$

$$= \frac{\hat{\beta}}{|\hat{\beta}|}\frac{\sum(y_i-\bar{y})(x_i-\bar{x})}{\sqrt{\sum(y_i-\bar{y})^2}\sqrt{\sum(x_i-\bar{x})^2}} = \frac{\hat{\beta}}{|\hat{\beta}|} r_{xy}$$

If $\hat{\beta}>0$, $r_{y\hat{y}} = r_{xy}$

If $\hat{\beta}<0$, $r_{y\hat{y}} = -r_{xy}$

$\Big\}\ |r_{y\hat{y}}| = |r_{xy}|$

3. If $x_i^* = a+bx_i$, $y_i^* = c+dy_i$, what will be the impact on the
a) OLS estimators of $\hat{\alpha}$ & $\hat{\beta}$
b) value of R^2.

$$x_i^* = a+bx_i \quad , \quad y_i^* = c+dy_i$$

$$\bar{x_i}^* = a+b\bar{x} \quad , \quad \bar{y_i}^* = c+d\bar{y}$$

$$x_i^* = (a+bx_i - a-b\bar{x}) \quad , \quad y_i^* = (c+dy_i - \bar{c}-d\bar{y}) = dy_i$$
$$= bx_i$$

$$\hat{\beta}^* = \frac{\sum x_i^* y_i^*}{\sum x_i^{*2}} = \frac{bd\sum x_i y_i}{b^2\sum x_i^2} = \frac{d}{b}\hat{\beta}$$

$$\hat{\alpha}^* = \bar{y}^* - \hat{\beta}^*\bar{x}^* = c+d\bar{y} - \frac{d}{b}\hat{\beta}(a+b\bar{x}) = \left(c - \frac{ad}{b}\hat{\beta}\right) + d(\bar{y}-\hat{\beta}\bar{x})$$

$$\hat{\alpha}^* = \theta + d\hat{\alpha}$$

$$R^{*2} = \frac{\hat{\beta}^{*2}\sum x_i^{*2}}{\sum y_i^{*2}} = \frac{d^2}{b^2}\hat{\beta}^2\frac{b^2\sum x_i^2}{d^2\sum y_i^2} = \hat{\beta}^2\frac{\sum x_i^2}{\sum y_i^2} = R^2.$$

★ If $x_i^* = a+bx_i$, $y_i^* = c+dy_i$

$$\hat{\beta}^* = \frac{d}{b}\hat{\beta} \qquad \hat{\alpha}^* = \theta + d\hat{\alpha} \qquad \theta = \left(c - \frac{ad}{b}\hat{\beta}\right)$$

$$R^{*2} = R^2.$$

4. For the model $Y_i = \alpha + \beta X_i + u_i$, if the scales of Y_i and X_i are changed by w_1 and w_2 respectively, i.e.

$$Y_i^* = w_1 Y_i \;,\quad X_i^* = w_2 X_i \;,$$

prove that $\text{var}(\hat{\alpha}^*) = w_1^2 \,\text{var}(\hat{\alpha})$ and $\text{var}(\hat{\beta}^*) = \left(\frac{w_1}{w_2}\right)^2 \text{var}(\hat{\beta})$.

Proof: $Y_i^* = w_1 Y_i \quad X_i^* = w_2 X_i$

$$\bar{Y}^* = w_1 \bar{Y} \quad \bar{X}^* = w_2 \bar{X}$$

$$y_i^* = Y_i^* - \bar{Y}^* \quad x_i^* = X_i^* - \bar{X}^*$$

$$y_i^* = w_1 Y_i - w_1 \bar{Y} = w_1 y_i$$

$$x_i^* = w_2 X_i - w_2 \bar{X} = w_2 x_i$$

Hence,

$$\sigma^{*2} = \frac{\sum u_i^{*2}}{n-2}$$

$$= \frac{\sum \left\{ w_1 Y_i - w_1 \hat{\alpha} - \left(\frac{w_1}{w_2}\right) \hat{\beta} w_2 X_i \right\}^2}{n-2}$$

$$\sigma^{*2} = \frac{w_1^2 \sum (Y_i - \hat{\alpha} - \hat{\beta} X_i)^2}{n-2} = \frac{w_1^2 \sum \hat{u}_i^2}{n-2} = w_1^2 \hat{\sigma}^2$$

$$\text{var}(\beta^*) = \frac{\sigma^{*2}}{\sum x_i^{*2}} = \frac{w_1^2 \hat{\sigma}^2}{w_2^2 \sum x_i^2} = \left(\frac{w_1}{w_2}\right)^2 \text{var}(\hat{\beta})$$

$$\text{var}(\alpha^*) = \sigma^{*2} \left(\frac{1}{n} + \frac{\bar{X}^{*2}}{\sum x_i^{*2}} \right) = w_1^2 \hat{\sigma}^2 \left(\frac{1}{n} + \frac{w_2^2 \bar{X}^2}{w_2^2 \sum x_i^2} \right)$$

$$= w_1^2 \hat{\sigma}^2 \left(\frac{1}{n} + \frac{\bar{X}^2}{\sum x_i^2} \right) = w_1^2 \,\text{var}(\hat{\alpha})$$

★ **The least-squares estimator of σ^2**

$$Y_i = \alpha + \beta X_i + u_i$$

$$\Rightarrow \bar{Y} = \alpha + \beta \bar{X} + \bar{u}$$

$$\Rightarrow y_i = Y_i - \bar{Y} = \beta x_i + (u_i - \bar{u})$$

$$\hat{u}_i = y_i - \hat{\beta} x_i$$

$$\Rightarrow \hat{u}_i = \beta x_i + (u_i - \bar{u}) - \hat{\beta} x_i$$

$$\Rightarrow \sum \hat{u}_i^2 = (\hat{\beta} - \beta)^2 \sum x_i^2 + \sum (u_i - \bar{u})^2 - 2(\hat{\beta} - \beta) \sum x_i (u_i - \bar{u})$$

$$\Rightarrow E\left(\sum \hat{u}_i^2\right) = \sum x_i^2 E(\hat{\beta} - \beta)^2 + E\left[\sum (u_i - \bar{u})^2\right] - 2E\left[(\hat{\beta} - \beta) \sum x_i (u_i - \bar{u})\right]$$

$$= \sum x_i^2 \operatorname{var}(\hat{\beta}) + (n-1) \operatorname{var}(u_i) - 2E\left[\sum k_i u_i \sum x_i (u_i - \bar{u})\right]$$

$$= \sum x_i^2 \operatorname{var}(\hat{\beta}) + (n-1) \operatorname{var}(u_i) - 2\left[\sum k_i u_i (x_i u_i)\right]$$

$$= \sigma^2 + (n-1)\sigma^2 - 2E\left[\sum k_i x_i u_i^2\right]$$

$$E\left[\frac{\sum x_i^2 u_i^2}{\sum x_i^2}\right] = \frac{1}{\sum x_i^2} \sum x_i^2 E(u_i^2)$$

$$= \sigma^2$$

$$= \sigma^2 + (n-1)\sigma^2 - 2\sigma^2 = (n-2)\sigma^2$$

$$\Rightarrow E \sum (u_i - \bar{u})^2 = E\left[\sum u_i^2 - n\bar{u}^2\right]$$

$$= E\left[\sum u_i^2 - n\left(\frac{\sum u_i}{n}\right)^2\right]$$

$$= E\left[\sum u_i^2 - \frac{1}{n}\left(\sum u_i\right)^2\right] = n\sigma^2 - \frac{n}{n}\sigma^2 = (n-1)\sigma^2$$

{u_i are uncorrelated and the variance of each u_i is σ^2}.

$$E\left(\sum \hat{u}_i^2\right) = (n-2)\sigma^2$$

$$\hat{\sigma}^2 = \frac{\sum \hat{u}_i^2}{n-2}$$

$$E(\hat{\sigma}^2) = E\left(\frac{\sum \hat{u}_i^2}{n-2}\right) = \sigma^2.$$

$\hat{\sigma}^2$ is an unbiased estimator of σ^2

5. Prove that, for the model $Y_i = \alpha + \beta X_i + u_i$,

$\hat{\beta} = \frac{\sum x_i y_i}{\sum x_i^2}$ is unbiased, efficient and consistent estimator of β.

i) Unbiasedness of $\hat{\beta}$

$$\hat{\beta} = \frac{\sum x_i y_i}{\sum x_i^2} = \frac{\sum x_i (Y_i - \bar{Y})}{\sum x_i^2} = \frac{\sum x_i Y_i}{\sum x_i^2} - \bar{Y}\frac{\sum x_i}{\sum x_i^2} = \frac{\sum x_i Y_i}{\sum x_i^2} = \sum k_i Y_i$$

$$= \sum k_i (\alpha + \beta X_i + u_i)$$

$$= \alpha \sum k_i + \beta \sum k_i X_i + \sum k_i u_i$$

$$= \beta + \sum k_i u_i$$

$$E(\hat{\beta}) = E(\beta + \sum k_i u_i) = \beta + E(\sum k_i u_i)$$

$$= \beta + \sum k_i E(u_i) = \beta. \quad \{E(u_i) = 0\}$$

$\hat{\beta}$ is an unbiased estimator of β.

ii) Efficiency of $\hat{\beta}$

$$\text{var}(\hat{\beta}) = \frac{\sigma^2}{\sum x_i^2}$$

Let us consider another unbiased estimator of β:

$$\beta^* = \sum w_i Y_i = \sum w_i (\alpha + \beta X_i + u_i)$$

$$\beta^* = \sum w_i Y_i = \alpha \sum w_i + \beta \sum w_i X_i + \sum w_i u_i$$

Since, β^* is assumed to be an unbiased estimator of β,

$E(\beta^*) = \beta$. $\sum w_i = 0$ and $\sum w_i X_i = 1$

$$\beta^* = \beta + \sum w_i u_i$$

$$\text{var}(\beta^*) = E(\beta^* - E(\beta^*))^2 = E(\beta^* - \beta)^2 = E(\sum w_i u_i)^2$$

$$= E(\sum w_i^2 u_i^2 + 2\sum w_i w_j u_i u_j) = \sum w_i^2 E(u_i^2) + 2\sum w_i w_j E(u_i u_j)$$

$$= \sigma^2 \sum w_i^2$$

$$\text{var}(\beta^*) = \sigma^2 \sum \{(w_i - k_i) + k_i\}^2 = \sigma^2 \{\sum (w_i - k_i)^2 + \sum k_i^2 + 2\sum (w_i - k_i) k_i\}$$

~~$\text{var}(\beta^*) > \sigma^2$~~

$$\text{var}(\beta^*) = \sigma^2 \left\{ \sum (w_i - k_i)^2 + \sum \left(\frac{x_i}{\sum x_i^2}\right)^2 + 2\sum \left(w_i - \frac{x_i}{\sum x_i^2}\right)\left(\frac{x_i}{\sum x_i^2}\right) \right\}$$

$$\Rightarrow \text{var}(\beta^*) = \sigma^2 \left\{ \sum (w_i - k_i)^2 + \frac{\sum x_i^2}{(\sum x_i^2)^2} + 2\sum \left[\frac{w_i x_i}{\sum x_i^2} - \frac{x_i^2}{\sum x_i^2} \right] \right\}$$

$$\Rightarrow \text{var}(\beta^*) = \sigma^2 \left\{ \sum (w_i - k_i)^2 + \frac{1}{\sum x_i^2} + 2\left(\frac{\sum w_i x_i}{\sum x_i^2} - \frac{\sum x_i^2}{(\sum x_i^2)^2} \right) \right\}$$

$$\Rightarrow \text{var}(\beta^*) = \sigma^2 \left\{ \sum (w_i - k_i)^2 + \frac{1}{\sum x_i^2} + 2\left(\frac{1}{\sum x_i^2} - \frac{1}{\sum x_i^2} \right) \right\}$$

$$\Rightarrow \text{var}(\beta^*) = \sigma^2 \left\{ \sum (w_i - k_i)^2 + \frac{1}{\sum x_i^2} \right\}$$

as $\sum w_i x_i = \sum w_i (X_i - \bar{X}) = \sum w_i X_i - \bar{X} \sum w_i = 1$

Hence, $\text{var}(\beta^*)$ will be minimum when $\sum (w_i - k_i)^2 = 0$ or

$$w_i = k_i = \frac{x_i}{\sum x_i^2} \quad \text{as} \quad \frac{1}{\sum x_i^2} > 0$$

→ when $\text{var}(\beta^*)$ is minimum, $\text{var}(\beta^*) = \text{var}(\hat{\beta})$.

Alternatively, β^* will have the minimum variance when $\beta^* = \hat{\beta}$. Thus $\hat{\beta}$ is an efficient estimator of β.

iii) Consistency of $\hat{\beta}$

If $\hat{\beta}$ is a consistent estimator of β, $\hat{\beta} = \beta$ as $n \to \infty$ or

$\text{var}(\hat{\beta}) = 0$ as $n \to \infty$.

$$\text{var}(\hat{\beta}) = \frac{\sigma^2}{\sum x_i^2} = \frac{\left(\frac{\sigma^2}{n}\right)}{\frac{\sum x_i^2}{n}}$$

Here, $\frac{\sum x_i^2}{n} = \text{var}(X) \neq 0$ {by Assumption}

Hence, as $n \to \infty$, $\text{var}(\hat{\beta}) \to 0$. This means $\hat{\beta}$ is a consistent estimator of β.

6. Prove that, for the model $Y_i = \alpha + \beta X_i + u_i$, $\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}$ is an unbiased, efficient and consistent estimator of α.

$$\hat{\alpha} = \bar{Y} - \frac{\sum x_i y_i}{\sum x_i^2}\bar{X}$$

$$\Rightarrow \hat{\alpha} = \frac{\sum Y_i}{n} - \frac{\sum x_i y_i}{\sum x_i^2}\bar{X} \Rightarrow \hat{\alpha} = \sum\left[\frac{1}{n} - k_i\bar{X}\right]Y_i$$

$$\Rightarrow \hat{\alpha} = \sum\left[\frac{1}{n} - k_i\bar{X}\right](\alpha + \beta X_i + u_i)$$

$$\Rightarrow \hat{\alpha} = \alpha - \bar{X}\alpha\underbrace{\sum k_i}_{0} + \beta\bar{X} - \beta\bar{X}\sum k_i X_i + \frac{\sum u_i}{n} - \bar{X}\sum k_i u_i$$

$$\Rightarrow \hat{\alpha} = \alpha + \frac{\sum u_i}{n} - \bar{X}\frac{\sum k_i u_i}{n}$$

$$E(\hat{\alpha}) = \alpha + \frac{\sum E(u_i)}{n} - \bar{X}\frac{\sum k_i E(u_i)}{n}$$

($\sum E(u_i)/n \to 0$, $\sum k_i E(u_i)/n \to 0$)

$$E(\hat{\alpha}) = \alpha \text{ (Unbiased)}.$$

Let $\alpha^* = \sum p_i Y_i$ be an unbiased estimator of α.

$$\text{var}(\hat{\alpha}) = \sigma^2\left(\frac{1}{n} + \frac{\bar{X}^2}{\sum x_i^2}\right)$$

$$\alpha^* = \sum p_i(\alpha + \beta X_i + u_i) = \alpha\sum p_i + \beta\sum p_i X_i + \sum p_i u_i$$

as α^* is an unbiased estimator of α

$\therefore \sum p_i = 1 \quad \sum p_i X_i = 0$

Let $p_i = \frac{1}{n} - w_i\bar{X}$

$\sum p_i \Rightarrow \sum w_i = 0 \quad \sum p_i X_i \Rightarrow \bar{X} - \bar{X}\sum w_i X_i = 0$

$\Rightarrow \sum w_i X_i = 1$

$$\text{var}(\beta^*) = E(\beta^* - E(\beta^*))^2 = E(\sum p_i u_i)^2$$

$$= E\left(\sum\left(\frac{1}{n} - w_i\bar{X}\right)u_i\right)^2$$

$$= E\left(\frac{\sum u_i}{n} - \bar{X}\sum w_i u_i\right)^2$$

$$= E(-\bar{X}\sum w_i u_i)^2$$

$$= \bar{X}^2 E(\sum w_i^2 u_i^2 + 2\sum w_i w_j u_i u_j)$$

$$= \bar{X}^2 \left[E(\sum w_i^2 u_i^2) + 2\sum w_i w_j E(u_i u_j)\right]$$

($E(u_i u_j) \to 0$)

$$= \bar{X}^2 E(\sum w_i^2 u_i^2)$$

$$= \bar{X}^2 \sum w_i^2 E(u_i^2) = \bar{X}^2 \sigma^2 \sum w_i^2$$

$$\text{var}(\alpha^*) = \bar{X}^2\sigma^2 \sum \{(w_i - k_i) + k_i\}^2$$

$$= \bar{X}^2\sigma^2 \sum \{(w_i - k_i)^2 + k_i^2 + 2(w_i - k_i)k_i\}$$

$$= \bar{X}^2\sigma^2 \sum \left\{(w_i - k_i)^2 + \frac{x_i^2}{(\sum x_i^2)^2} + 2\left(w_i - \frac{x_i}{\sum x_i^2}\right)\frac{x_i}{\sum x_i^2}\right\}$$

$$= \bar{X}^2\sigma^2 \left\{\sum (w_i - k_i)^2 + \frac{\sum x_i^2}{(\sum x_i^2)^2} + 2\frac{\sum w_i x_i}{\sum x_i^2} - 2\frac{\sum x_i^2}{(\sum x_i^2)^2}\right\}$$

($\sum w_i x_i \to 0$)

$$= \bar{X}^2\sigma^2 \left\{\sum (w_i - k_i)^2 + \frac{1}{\sum x_i^2} + 2\left(\frac{1}{\sum x_i^2} - \frac{1}{\sum x_i^2}\right)\right\}$$

$$= \bar{X}^2\sigma^2 \left\{\sum (w_i - k_i)^2 + \frac{1}{\sum x_i^2}\right\}$$

for $\sum (w_i - k_i)^2$ to be minimum, $w_i = k_i = \frac{x_i}{\sum x_i^2}$

$\text{var}(\alpha^*)$ is minimum when $\alpha^* = \hat{\alpha}$.

iii) Consistency

$$\text{var}(\hat{\alpha}) = \sigma^2\left(\frac{1}{n} + \frac{\bar{X}^2}{\sum x_i^2}\right)$$

$$= \frac{\sigma^2}{n} + \bar{X}^2\frac{\sigma^2}{\sum x_i^2}$$

$$\lim_{n\to\infty} \text{var}(\hat{\alpha}) = \frac{\sigma^2}{n} + \bar{X}^2 \frac{\sigma^2/n}{\sum x_i^2/n} = 0.$$

$\frac{\sum x_i^2}{n} \to$ variance of x_i (finite)

$\frac{\sigma^2}{n} \to 0$

$\therefore \hat{\alpha}$ is a consistent estimator of α.

Correlation Coefficient v/s Regression coefficient -

Correlation coefficient represents the degree of association between two variables.

Regression coefficient indicates the nature and ~~impact of~~ extent of impact (how the independent variable causes the dependent variable).

For two variables X and Y, the pairwise correlation coefficient,

$$r = \frac{\sum x_i y_i}{\sqrt{\sum x_i^2 \sum y_i^2}}$$

However, there are two regression coefficients depending on whether

a) Y is regressed on X

$$y_i = \alpha + \beta x_i + u$$

$$\hat{\beta} = \frac{\sum x_i y_i}{\sum x_i^2} = \left(\frac{\sum x_i y_i}{\sqrt{\sum x_i^2 \sum y_i^2}} \frac{\sqrt{\sum y_i^2}}{\sqrt{\sum x_i^2}} \right) = r \frac{\sigma_y}{\sigma_x} \Rightarrow \boxed{\hat{\beta} = r \frac{\sigma_y}{\sigma_x}}$$

b) X is regressed on Y.

$$x_i = \gamma + \delta y_i + v_i$$

$$\hat{\delta} = \frac{\sum x_i y_i}{\sum y_i^2} = \left(\frac{\sum x_i y_i}{\sqrt{\sum x_i^2 \sum y_i^2}} \frac{\sqrt{x_i^2}}{\sqrt{y_i^2}} \right) = r \frac{\sigma_x}{\sigma_y} \Rightarrow \boxed{\hat{\beta} = r \frac{\sigma_x}{\sigma_y}}$$

* The pairwise correlation coefficient r lies between -1 and +1, i.e., $-1 \leq r \leq +1$.

Proof: Let us consider n pairs of observations of two variables X and Y.

$$x_i = \frac{X_i - \bar{X}}{\sigma_X} \quad \text{and} \quad y_i = \frac{Y_i - \bar{Y}}{\sigma_Y}$$

$$\sum x_i^2 = \frac{\sum (X_i - \bar{X})^2}{\sigma_X^2} = \frac{n\sigma_X^2}{\sigma_X^2} = n$$

$$\sum y_i^2 = \frac{\sum (Y_i - \bar{Y})^2}{\sigma_Y^2} = \frac{n\sigma_Y^2}{\sigma_Y^2} = n$$

$$\sum x_i y_i = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sigma_X \sigma_Y} = \frac{n\,cov(X,Y)}{\sigma_X \sigma_Y} = nr$$

$$\sum (x_i + y_i)^2 \geq 0 \Rightarrow \sum x_i^2 + \sum y_i^2 + 2\sum x_i y_i \geq 0$$

$$\Rightarrow n + n + 2nr \geq 0$$

$$r \geq -1$$

$$\sum (x_i - y_i)^2 \geq 0 \Rightarrow \sum x_i^2 + \sum y_i^2 - 2\sum x_i y_i \leq 0$$

$$\Rightarrow n + n - 2nr \leq 0 \Rightarrow r \leq 1$$

$$\Rightarrow -1 \leq r \leq 1$$

• Decomposition of TSS

$$R^2 = \frac{ESS}{TSS} \qquad ESS = R^2 TSS$$

$$TSS = ESS + RSS \Rightarrow RSS = TSS - ESS$$

$$= TSS - R^2 TSS = (1 - R^2)\,TSS$$

$$\boxed{TSS = [R^2 \times TSS] + [(1 - R^2) \times TSS]}$$

For a bivariate model, $R^2 = r^2$ $\quad TSS = [r^2 \times TSS] + [(1 - r^2)\,TSS]$

★ Spearman's Rank Correlation Coefficient

$$r_s = 1 - 6\left[\frac{\sum_{i=1}^{n} d_i^2}{n(n^2-1)}\right]$$

d_i = Difference in ranks of $|u_i|$ and x_i.

Let us consider n pairs of observations of the two variables X and Y. If the values of these observations are ranked in ascending or descending order, both X and Y take values of the first n natural numbers. Hence, $X_i = 1, 2, \ldots n$ and $Y_i = 1, 2, \cdots, n$.

$$\sum X_i = \sum Y_i = \frac{n(n+1)}{2} \qquad \bar{X} = \bar{Y} = \frac{n+1}{2}$$

$$\sum X_i^2 = \sum Y_i^2 = \frac{n(n+1)(2n+1)}{6}$$

$$\sigma_X^2 = \frac{\sum X_i^2}{n} - \left(\frac{\sum X_i}{n}\right)^2 = \frac{(n+1)(2n+1)}{6} - \frac{(n+1)^2}{4}$$

$$= \frac{n^2-1}{12} = \sigma_Y^2 \quad (\text{Since } \sum X_i = \sum Y_i \; ; \; \bar{X} = \bar{Y})$$

Let d_i is the difference between the rank of X and Y for the observation i, i.e., $d_i = X_i - Y_i$

$$\sum d_i^2 = \sum \{(X_i - \bar{X}) - (Y_i - \bar{Y})\}^2 \quad (\because \bar{X} = \bar{Y})$$

$$\sum d_i^2 = \sum (X_i - \bar{X})^2 + \sum (Y_i - \bar{Y})^2 - 2\sum (X_i - \bar{X})(Y_i - \bar{Y})$$

$$\frac{\sum d_i^2}{n} = \frac{\sum (X_i - \bar{X})^2}{n} + \frac{\sum (Y_i - \bar{Y})^2}{n} - \frac{2\sum \{(X_i - \bar{X})(Y_i - \bar{Y})\}}{n}$$

$$= \sigma_X^2 + \sigma_Y^2 - 2\,\text{cov}(X, Y)$$

$$= 2\sigma_X^2 - 2\,\text{cov}(X, Y) \quad (\because \sigma_X^2 = \sigma_Y^2)$$

$$\text{cov}(X, Y) = \sigma_X^2 - \frac{\sum d_i^2}{2n}$$

$$r_s = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{cov(X,Y)}{\sigma_X^2}$$

$$= 1 - \frac{\sum d_i^2}{2n\sigma_X^2} = 1 - \frac{12\sum d_i^2}{2n(n^2-1)}$$

$$r_s = 1 - 6\left[\frac{\sum d_i^2}{n(n^2-1)}\right]$$

★ Relation between Ordinary R^2 and adjusted R^2

Adjusted R^2: $\bar{R}^2$

$$\bar{R}^2 = R^2 - \left(\frac{k-1}{n-k}\right)(1-R^2) = 1 - \left(\frac{n-1}{n-k}\right)(1-R^2)$$

$$\boxed{\bar{R}^2 = 1 - \left[\frac{n-1}{n-k}\right](1-R^2)}$$

★ Hypothesis Testing

• Properties of Variance

- $E(X-\mu)^2 = E(X^2) - \mu^2$
- $var(aX+b) = a^2 var(X)$
- If X and Y are independent

  $var(X+Y) = var(X) + var(Y)$

  $var(X-Y) = var(X) + var(Y)$
- $var(aX+bY) = a^2 var(X) + b^2 var(Y)$

• Properties of Covariance

$$cov(X,Y) = E\{(X-\mu_X)(Y-\mu_Y)\} = E(XY) - \mu_X\mu_Y$$

- If X and Y are independent,

  $cov(X,Y) = 0$
- ~~$cov(aX, bY)$~~ $cov(a+bX, c+dY) = bd\,cov(X,Y)$

Correlation Coefficient

$$\rho = \frac{cov(X,Y)}{\sqrt{\{var(X)\,var(Y)\}}} = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$$

$$cov(X,Y) = \rho \sigma_X \sigma_Y$$

- $var(X+Y) = var(X) + var(Y) + 2\rho\sigma_X\sigma_Y$
- $var(X-Y) = var(X) + var(Y) - 2\rho\sigma_X\sigma_Y$
- $var\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} var(X_i) + 2\sum_{i<j}\sum \rho_{ij}\sigma_i\sigma_j$

Skewness

$$S = \frac{E(X-\mu)^3}{\sigma^3} = \frac{\text{third moment about the mean}}{\text{cube of the standard deviation}}$$

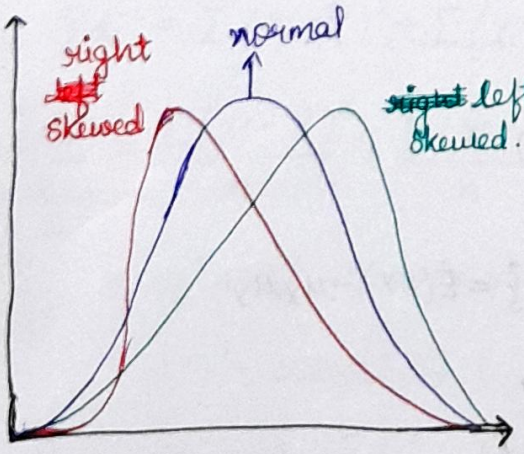
Kurtosis

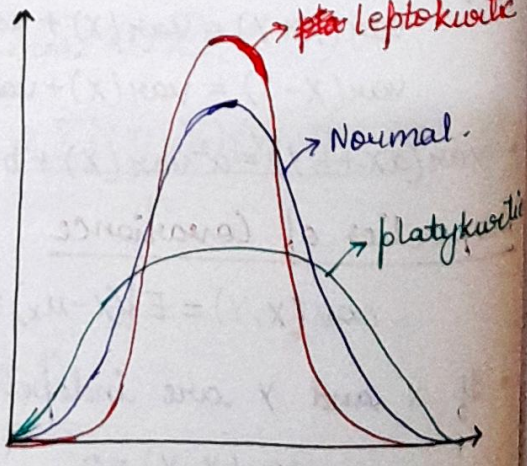
$$K = \frac{E(X-\mu)^4}{[E(X-\mu)^2]^2} = \frac{\text{Fourth moment about mean}}{\text{Square of variance.}}$$

$K < 3 \Rightarrow$ platykurtic (fat or short-tailed)

$K > 3 \Rightarrow$ leptokurtic (slim or long-tailed)

$K = 3 \Rightarrow$ normal distribution.





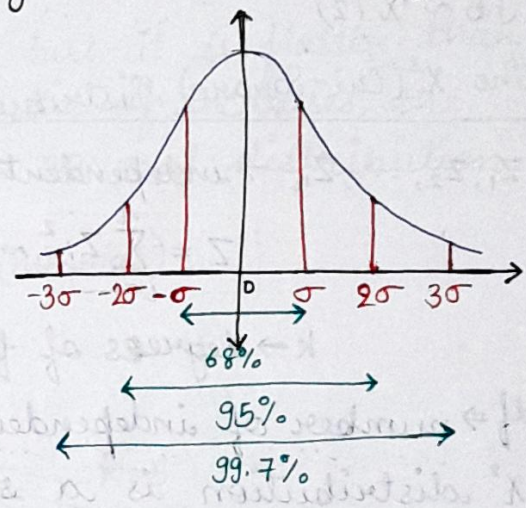
★Normal Distribution

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{\frac{-1}{2}\frac{(x-\mu)^2}{\sigma^2}\right\} \quad -\infty < x < \infty$$

Approximately, 68% of the area under the normal curve lies between the values of $\mu \pm \sigma$, about 95% ~~lies~~ of the area ~~between~~ lies between $\mu \pm 2\sigma$, and about 99.7% of the area lies between $\mu \pm 3\sigma$.

$$Z = \frac{X-\mu}{\sigma} \quad \text{(Standardized Normal variable)}$$

$$f(z) = \frac{1}{\sqrt{2\pi}} \exp\left\{\frac{-z^2}{2}\right\}$$



Let $X_1 \sim N(\mu_1, \sigma_1^2)$, $X_2 \sim N(\mu_2, \sigma_2^2)$

$$Y = aX_1 + bX_2$$

$$Y \sim N[a\mu_1 + b\mu_2, a^2\sigma_1^2 + b^2\sigma_2^2]$$

• Central Limit Theorem

Let $X_1, X_2, \ldots, X_n$ denote n independent random variables, all of which have the same PDF with mean $= \mu$ and variance $= \sigma^2$.

Let $\bar{X} = \frac{1}{n}\sum X_i$ (Sample mean).

Then as n increases indefinitely (i.e. $n \to \infty$)

$$\bar{X} \underset{n\to\infty}{\sim} N\left(\mu, \frac{\sigma^2}{n}\right)$$

That is, $\bar{X}$ approaches the normal distribution with mean μ and variance σ^2/n. This result holds true regardless of the form of PDF.

$$Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} = \frac{\sqrt{n}(\bar{X}-\mu)}{\sigma} \sim N(0,1)$$

- Third Moment : $E(X-\mu)^3 = 0$
- Fourth Moment : $E(X-\mu)^4 = 3\sigma^4$

All odd-powered moments about the mean value of a normally distributed variable are zero.

- Jarque-Bera (JB) test of Normality

$$JB = n\left[\frac{S^2}{6} + \frac{(K-3)^2}{24}\right]$$

$JB \sim \chi^2(2)$

★ The χ^2 (Chi-Square) Distribution

$Z_1, Z_2, \ldots, Z_n \rightarrow$ independent standardized Normal variables

$$Z = \sum_{i=1}^{n} Z_i^2 \sim \chi^2_K$$

$k \rightarrow$ degrees of freedom (df)

df ⇒ number of independent quantities in the previous sum

- χ^2 distribution is a skewed distribution, the degree of skewness depending on the df. For comparitively few df, the distribution is highly skewed to the right, but as the number of df increases, the distribution becomes increasingly symmetrical.

If $df > 100$,

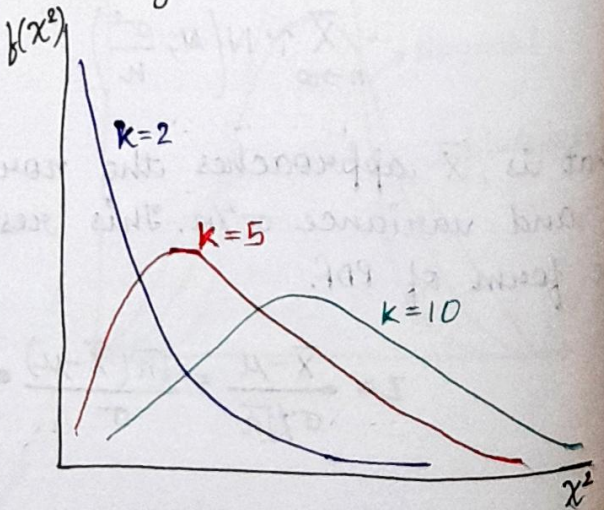
$\sqrt{2\chi^2} - \sqrt{2(k-1)} \rightarrow$ Standardized normal variable.

- $E(\chi^2_k)$ $\quad X \sim \chi^2_K$

$E(X) = k \quad V(X) = 2k.$

- $X_1, X_2 \Rightarrow \chi^2$ with $df = k_1$ & k_2.

$X_1 + X_2 \sim \chi^2_{k_1+k_2}$



★ Student's t-Distribution

$Z_1 \sim$ Standardized Normal Variable

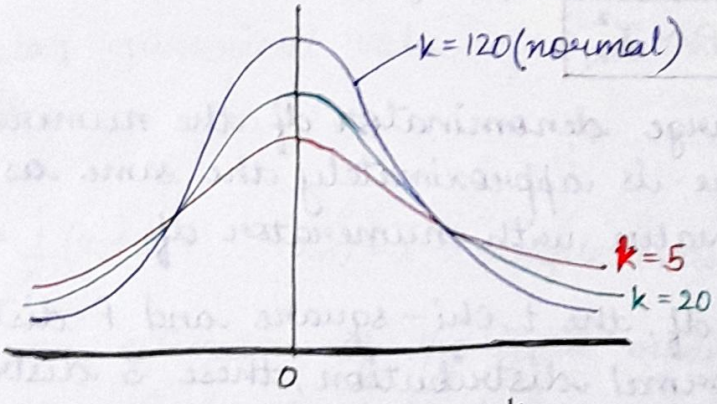
$Z_1 \sim N(0,1)$

$Z_2 \sim \chi^2_k$

$Z_1, Z_2 \Rightarrow$ independent

$$t = \frac{Z_1}{\sqrt{Z_2/k}} = \frac{\sqrt{k}\, Z_1}{\sqrt{Z_2}}$$

t-distribution is symmetrical, but it is flatter than the normal distribution. But as the df increases, the t-distribution approximates the normal distribution.



Mean = 0 Variance = $\frac{k}{k-2}$

★ The f distribution

$Z_1 \sim \chi^2_{k_1}$ $Z_2 \sim \chi^2_{k_2}$

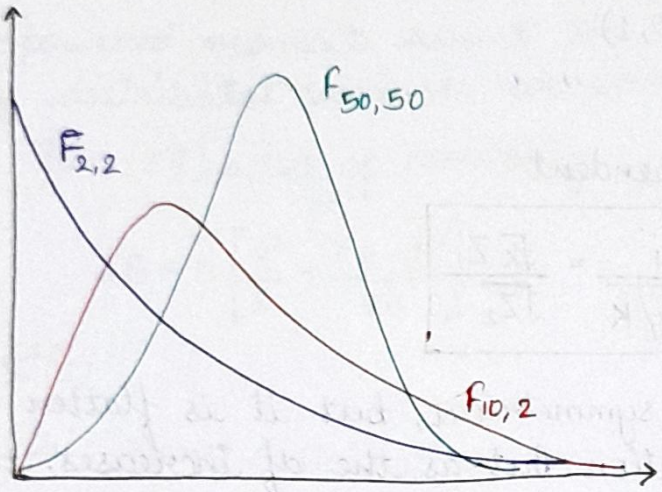
→ Independent.

$$F = \frac{Z_1/k_1}{Z_2/k_2} \qquad f_{k_1, k_2}$$

The f-distribution is skewed to the right. As k_1, k_2 become large, the f-distribution becomes normal.

Mean = $\frac{k_2}{k_2 - 2}$ $(k_2 > 2)$ Variance = $\frac{2k_2^2(k_1 + k_2 - 2)}{k_1(k_2-2)^2(k_2-4)}$ $(k_2 > 4)$

$$t_k^2 = F_{1,k}$$



If denominator df, k_2 is fairly large

$$k_1 F \sim \chi^2_{k_1}$$

i.e. for fairly large denominator df, the numerator df times the F value is approximately the same as a ~~chi~~ chi-square value with numerator df.

Since for large df, the t, chi-square and F-distribution approach the normal distribution, these 3 distributions are known as the distributions related to the normal distribution.

★ The Bernoulli Binomial Distribution

$$P(X=0) = 1-p$$
$$P(X=1) = p \text{ (Success)}$$

$$E(X) = p \qquad Var(X) = pq = p(1-p)$$

★ Binomial Distribution

$$f(X) = {}^nC_x\, p^x (1-p)^{n-x}$$

$X \Rightarrow$ no. of successes in n trials.