Jan 24, 2025

Review of data and statistics

Classification of data

1. Cross-sectional / Time Series / Panel Data

(Combined form of cross-sectional & Time Series)

Cross-sectional : Collecting the data in a single frame of time from large number of respondents.

eg: Employee satisfaction survey

→ Conducting a survey only once

Structure of data:

Respondent ↓	Variable 1	Variable 2	...	Variable K

Purposes of conducting cross-sectional study

* Identifying the current status of study area
  ↓
  central values, graphical representation
* Comparison of data by analytical domain.
  ↓
  province, age group, gender etc
* To identify the relationship among variables.
* Prediction and estimation.
* To make plans, policies and strategies.

Time Series

Collection of data over a period of time

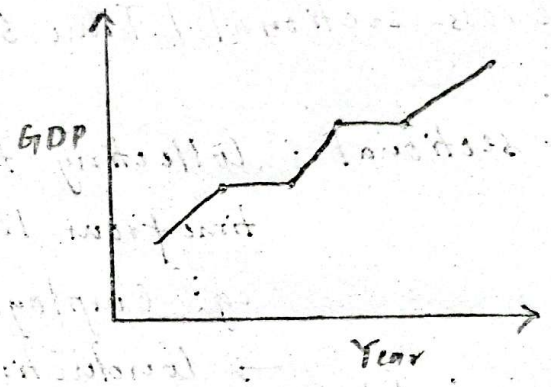
Used in macroeconomics

Examples

Year	GDP

Year Vs Remittance

Year Vs Profit



* Analysis of past trends
* Forecasting for the future
* Construction of business cycle
* Studying the seasonal impact

→ Regression in Time Series

Nepal living standard Survey by World Bank

2. Categorical vs Numerical

Categorical: Any question having alphabetic/grouping response

Gender □

- Female 1
- Male 2 ✓

use: Frequency/Percentage

Numerical: Any question having a numeric response

Salary $1200

use: Min, Max, Mean, Standard Dviation

3. According to number of variables

Univariate data

Bivariate data

Multivariate data $\xrightarrow[\text{technique}]{\text{dependence}}$ Dependent variable → one
Independent variable → multiple regression

Canonical correlation

Jan 25, 2025

Population : Totality of all items under the study
(study area — coverage)

Census: Complete enumeration
Collecting the information / data from each and every element of population

Parameter: Result / outcome of population data
Characteristics of population data

Population size → N
Population mean → μ
Population ~~property~~ proportion → Π (or P)
Population variance → σ^2
Population correlation → ρ
Population regression co-efficient → β_i

Sample: Subset of the population

Sampling: Collecting the information / data from sample.

Estimator / Statistic: Result or outcome of sample data.
Characteristic of sample size.
Sample size → n
Sample mean → $\bar{x}$ $(\hat{\mu})$

dist of all population items → frame

Sample proportion – p $(\hat{P})$

Sample variance → s^2 $(\hat{\sigma}^2)$

Sample correlation → r $(\hat{\rho})$

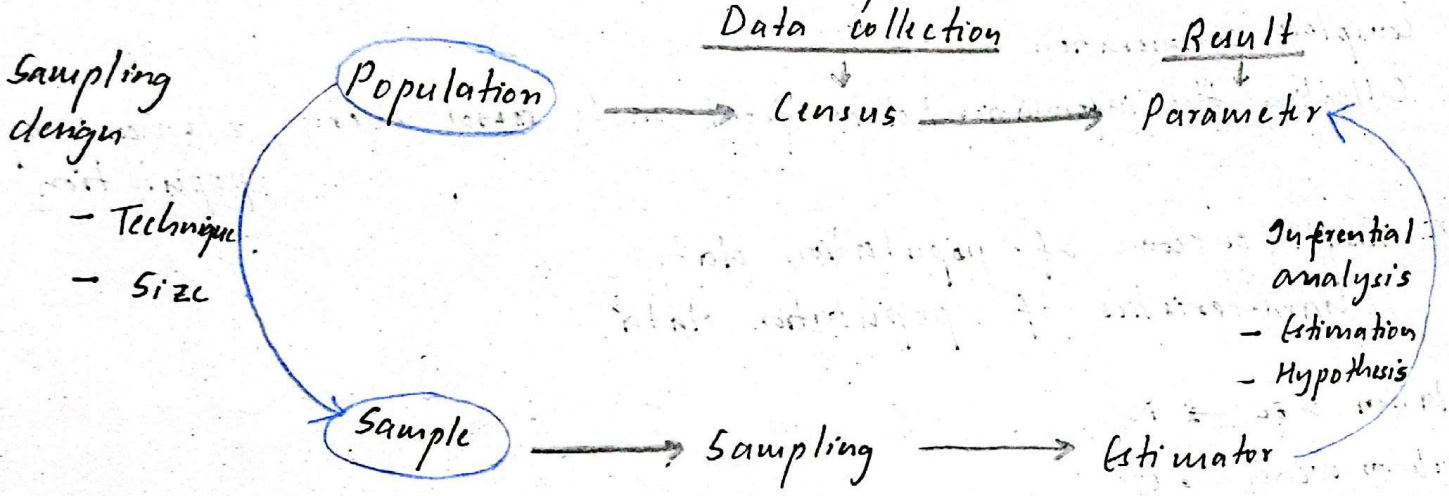
Sample regression coefficient → b_i $(\hat{\beta}_i)$

Sampling Error: Difference between Estimator and Parameter

= | Estimator – Parameter |

↓

Used to determine the sample size.



Research can be done in two ways:

Research

↓

Quantitative

↓

Survey

↓

Questionnaire (Instrument)

– should meet objective/ hypothesis

– (Should be valid)

– should measure what researcher wants to measure.

Statistics

– correct/ objective related (validity)

↓

Answer/response (data)

– correct (reliability)

↓

Data management/ Data Processing / Data Organization

↓

Result (Data Analysis)

↓

To gain Knowledge

↑

Policy / Program / Plans etc

Statistics is the collection, organization and analysis of data.

Statistics

Descriptive statistics

Describing the status
- Table / chart
- Central value
- Frequency / %
- Cross-Tabulation

Inferential statistics

Techniques of giving conclusion about the pop^n parameter on the basis of sample data / statistic / estimator.

Consists of
- Estimation
- Hypothesis

Hypothesis

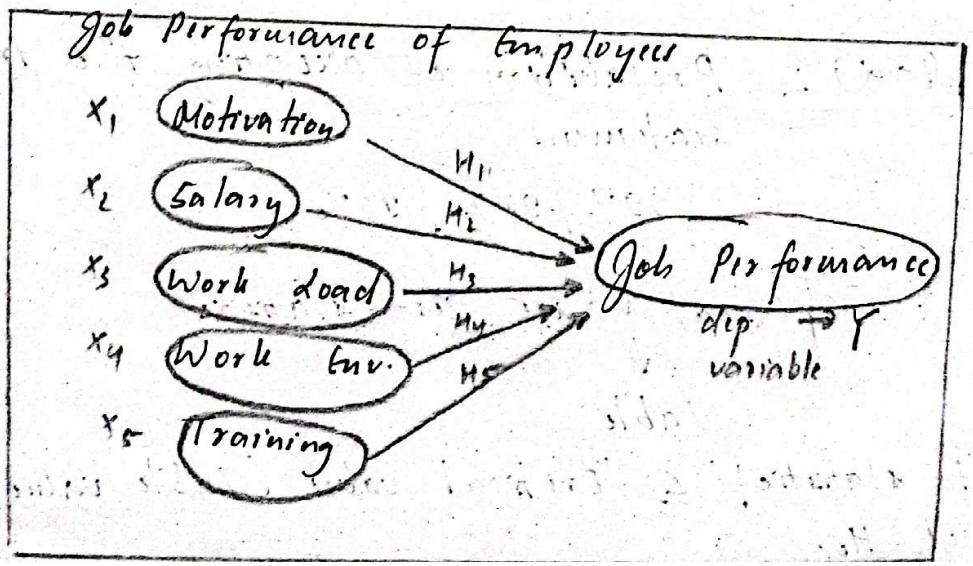
Pre-assumption / statement / guess about the population parameter. (outcome / result of the research)

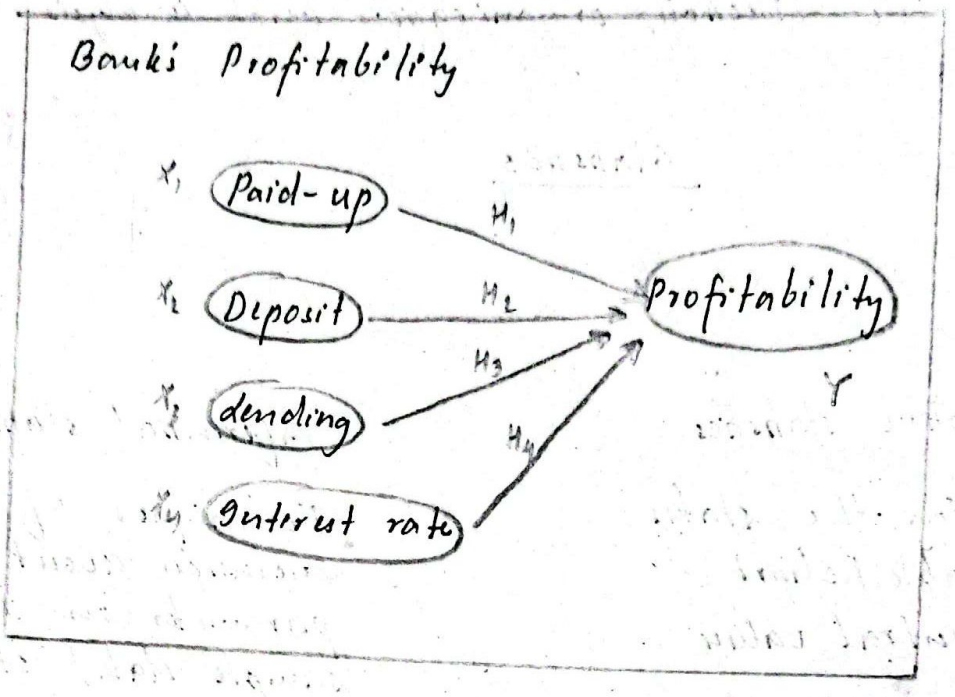
→ Set on the basis of Past literature

→ Conduction of Qualitative research / Pilot study

Null hypothesis (Ho): Statement of no difference or no relationship.

=, ≥, ≤ (equality)





Alternative hypothesis (H_1): Statement of relationship/difference.

$\neq$, $>$, $<$

↓ (Two tailed test)

(One Tailed Test)

$n \longrightarrow N$

Estimator → Parameter

Level of Significance (α): Probability of rejecting true statement:

1% | 5% | 10%

Confidence level ($1-\alpha$): Probability of accepting the true statement

99% | 95% | 90%

Decision Criteria: Critical value / Test statistic

↓

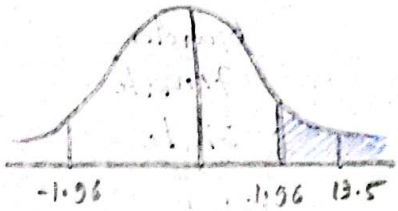
Table

If |test-statistic| $\leq$ critical value (Table value) do not reject H_0.

p-value approach

If p-value $< \alpha$ (reject H_0)

If p-value $\geq \alpha$ (do not reject H_0)



Test statistic
$t = 13.5$
Table (α) = 5%
1.96

Example

Q1 Gender

Female – 1
Male – 2
} Nominal

Q2 Post

Assistant 1
Officer 2
Manager 3
} Ordinal

Cross-Sectional

Q1	Q2	Q3	Q4
1	2	34000	15000

Q3 Income 34000
Q4 Expenditure 15000
} Ratio

Q5 _____

Q1 → Frequency /%.
Q2 → Frequency /%.
Q3 → Min, Max, Mean, SD
Q4 →

Testing or comparing

Q1, Q2 → χ^2 Test
Q1, Q3 → ANOVA Test.
Q3, Q4 → Z-test, t-test etc

Reference Book
Econometrics
– Wooldrige

Data analysis:

I. Univariate Analysis

Case a: Analysis of categorical data

Frequency / Percentage analysis
Graph / Visualization: Bar diagram or Pie chart

* Distribution by Gender

Gender	Frequency	%
Female		
Male		
Total		

Case b: Analysis of numerical data

Min / Max / Mean / St. deviation

Test: One sample t-test

Graph: Histogram / Box-plots

II. Bivariate Data Analysis

Analysis of two variables at a time

Case a: If both are categorical / grouping

Cross tabulation between two variables with frequency and appropriate %

+

Test → Chi-Square Test (χ^2 Test)

Gender \ Post	Assistant	Officer	Manager	
Female	70 (0.7)	20 (0.2)	10 (0.1)	100
Male	100 (0.5)	50 (0.25)	50 (0.25)	200
	170	70	60	300

+

Test: χ^2 Test

Case b: If one is categorical and another is numerical

Title: # Comparison of numerical by categorical

Gender	Avg	Min	Max	sd
Female				
Male				
Total				

Comparison between two means ⇒ Independent sample t-test

Title: # Comparison of Income by Post

Post	Avg	Min	Max	SD
Asst				
Officer				
Manager				
Total				

Comparison bt^n three means ⇒ One-way ANOVA (F-Test)

NOTE:

If categorical variable consists of exactly two sub-groups, we use independent sample t-test, else for more than two subgroups, we use f-test.

Case c: If both are numerical

Correlation : measures the relationship between variables.

Regression : Should separate dependent and independent variables.

Feb 1, 2025

Developing an equation

$Y = \beta_0 + \beta_1 X + \epsilon$ (Population)

$Y = b_0 + b_1 x + e$ (Sample)

$Y \rightarrow$ Observed value (Actual value)

$\hat{Y} \rightarrow$ Predicted value / Estimated value

$Y - \hat{Y} =$ error / residual

Objectives

① Prediction of dependent variable by using information of independent variables.

$$\hat{y} = \underset{b_0}{5.2} + \underset{b_1}{2.8} x_1 + \underset{b_2}{7.3} x_2$$

② Impact analysis

Impact of independent variables on dependent variables. (Slope / Coefficient)

③ Computation of amount / percentage change in dependent variable per unit / % change in independent variables. [meaning of slope]

Income	Expenditure
X	Y

$\hat{Y} = b_0 + b_1 X$

$b_1 = \Delta Y / \Delta X =$ $\dfrac{\text{amount change in } Y}{\text{percent change in } X}$

$\ln \hat{Y} = b_0 + b_1 \ln X$

$b_1 = \dfrac{\Delta Y/Y}{\Delta X/X} = \dfrac{\%\text{ change in } Y}{\%\text{ change in } X}$

↓

(Elasticity)

$\ln \hat{Y} = b_0 + b_1 X$

$b_1 = \dfrac{\Delta Y/Y}{\Delta X} = \dfrac{\%\text{ change}}{\text{unit change}}$

Year	GDP
X	Y
	$\ln \hat{Y} = b_0 + b_1 X$

~ Per annum

④ Computation of growth rate / amount per time period.

⑤ Proportion of ~~explained~~ variation in Y (dependent variable) that is explained by independent variables.

[Coefficient of determination → R^2]

Price	Sales Volume
	$b_1 = -7.5$
	$R^2 = 0.82$

⇒ 82% variation in sales volume is explained by price.

⑥ To make policies / programs / plans / decisions.

Simple Regression Analysis

dependent variable $\longrightarrow Y$

independent variables $\longrightarrow X$

Equation

$$Y = \beta_0 + \beta_1 X + \epsilon \quad \text{(Population)}$$

$$y = b_0 + b_1 X + e \quad \text{(Sample)}$$

By using least squares method.

$$b_1 = \frac{n\Sigma XY - (\Sigma X)(\Sigma Y)}{n\Sigma X^2 - (\Sigma X)^2}$$

$$= \frac{Cov(X,Y)}{Var(X)}$$

$$b_0 = \frac{\Sigma Y}{n} - b_1 \frac{\Sigma X}{n}$$

$$\hat{Y} = b_0 + b_1 X$$

Example

X (Income) (in '000)	Y (Exp in '000)	XY	X^2	Y^2	$\hat{Y} = 2.206 + 0.572X$	$(Y - \hat{Y})$	$(Y - \hat{Y})^2$
15	10	150	225	100	10.786	-0.786	0.6177
18	12	216	324	144	12.502	-0.502	0.2520
20	12	240	400	144	~~13.644~~ 13.646	-1.646	2.7093
20	17	340	400	289	13.646	3.354	11.2493
25	17	~~425 6.25~~	~~6.25~~	289	16.506	0.494	0.2440
29	18	522	841	324	18.794	-0.794	0.6304
127	86	1893	2815	1290		$\Sigma = 0.12 \approx 0$	$\Sigma(Y - \hat{Y})^2 = 15.701$

$\Sigma |e|$ (c)

$e = Y - \hat{Y}$ = Observed value - Estimated value

$$b_1 = \frac{6 * 1893 - 127 * 86}{6 * 2815 - (127)^2}$$

$$= \underline{0.572}$$

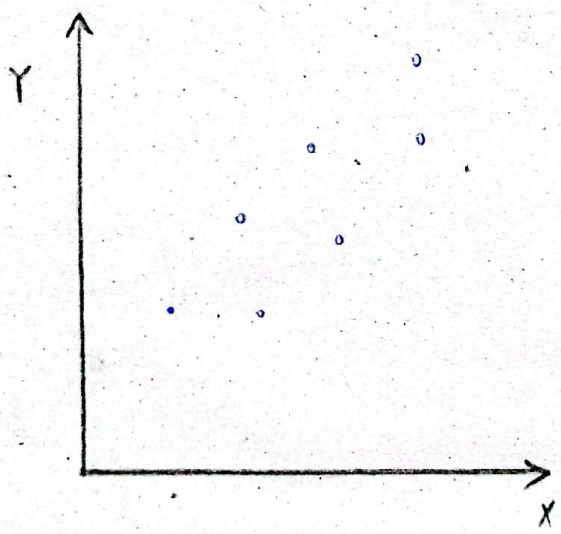
$$b_0 = \frac{\Sigma Y}{n} - b_1 \frac{\Sigma X}{n}$$

$$= \frac{86}{6} - 0.572 * \left(\frac{127}{6}\right)$$

$$= \underline{2.206}$$

The estimating equation $\hat{y} = 2.206 + 0.572 X$

(b_0 = 2.206, b_1 = 0.572)



* $E(e) = 0$
* Signs should not be in a pattern ($\Rightarrow$ Autocorrelation)
* Standard error of estimate $(\hat{y}) = \sqrt{\frac{\Sigma e^2}{n-2}}$

$SE(\hat{Y}) = S_{YX} = Sc(\hat{y}) = \sqrt{\frac{\Sigma (Y - \bar{Y})^2}{n-p-1}}$

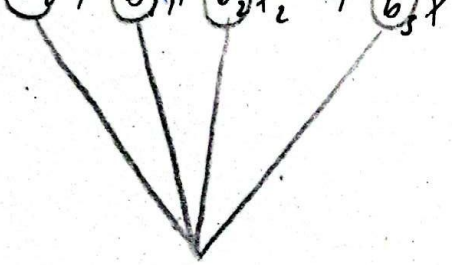
$p \rightarrow$ no. of independent variables

$= \sqrt{MSE}$

* It measures the average variation/ deviation of observed values of dependent variable (Y) around its fitted equation ($\hat{Y}$).

* We can use this for model selection.

$Y = b_0 + b_1 x_1 + b_2 x_2^2 + b_3 x$



linear

Feb 7, 2025

Multiple regression analysis

Important Packages
car, tidyverse, lmtest
> RStudio

Example

Sales Volume	Price	no. of stores	no. of ads
¥	¥	¥	¥
Y	X_1	X_2	X_3

$$Y = \beta_0 + \beta_1 X_1 + \beta_1 X_2 + \cdots + \beta_p X_p + \epsilon$$

where, β_0 = intercept
β_1 = parameter associated with x_1
β_2 = parameter associated with x_2 and so on.

For sample data it is written as

$$y = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p + e$$

(VVI)

Assumptions of CLRM (Classical linear Regression Model)

- Zero mean of ϵ_i $E(\epsilon_i) = 0$ for each i.
- Homoscedasticity Var $(\epsilon_i) = \sigma^2$ constant
  Homoscedasticity
- Non autocorrelation $Cov(\epsilon_i, \epsilon_j) = 0$ where $\epsilon_i \neq \epsilon_j$
- Normality: ϵ_i is normally distributed
- Non stochastic x_s, the values of x_s – variables are same in repeated samples.

- Zero covariance between ϵ_i and X variables.

$$Cov(\epsilon_i, X_{1i}) = Cov(\epsilon_i, X_{2i}) = 0$$

- No exact relationship exists between X variables i.e. Xs are not correlated (no multicollinearity).

Model specification and Assumption (in vector and matrix form)

- The general population regression model involving the dependent variable Y_i and the independent variables $X_{1i}, X_{2i}, X_{3i}, \dots, X_{Ki}$ is specified as:

Y_i	X_{1i}	X_{2i}	X_{3i}	...	X_{Ki}
Y_1					
Y_2					
⋮					
Y_n					

The multiple regression equation is written as:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \dots + \beta_K X_{Ki} + \epsilon_i$$

⋮

$$Y_n = \beta_0 + \beta_1 X_{1n} + \dots + \beta_K X_{Kn} + \epsilon_n$$

The system of linear equations can be written as:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & X_{21} & \cdots & X_{K1} \\ 1 & X_{12} & & & \vdots \\ \vdots & \vdots & & & \\ 1 & X_{1n} & & & X_{Kn} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_K \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

That is

$$Y = X\beta + \epsilon \qquad \text{——————} \quad (1)$$

Assumptions

1) Zero mean of ϵ

That is

$$E(\epsilon) = E\begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix} = \begin{bmatrix} E(\epsilon_1) \\ E(\epsilon_2) \\ \vdots \\ E(\epsilon_n) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

2) Constant variance of ϵ

$$Var(\epsilon) = E(\epsilon\epsilon')$$

$$= E\left[\begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix} [\epsilon_1 \ \epsilon_2 \cdots \epsilon_n]\right]$$

$$= E\begin{bmatrix} \epsilon_1^2 & \epsilon_1\epsilon_2 & \cdots & \epsilon_1\epsilon_n \\ \epsilon_2\epsilon_1 & \epsilon_2^2 & & \vdots \\ \vdots & & \ddots & \\ \epsilon_n\epsilon_1 & \cdots & \cdots & \epsilon_n^2 \end{bmatrix} = \begin{bmatrix} E(\epsilon_1^2) & E(\epsilon_1\epsilon_2) & \cdots & E(\epsilon_1\epsilon_n) \\ E(\epsilon_2\epsilon_1) & E(\epsilon_2^2) & \cdots & \\ \vdots & & & \\ E(\epsilon_n\epsilon_1) & \cdots & \cdots & E(\epsilon_n^2) \end{bmatrix}$$

$$= \begin{pmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & & \vdots \\ \vdots & & \ddots & \\ 0 & \cdots & & \sigma^2 \end{pmatrix}$$

$$\sigma^2 \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & & & \\ 0 & 0 & \cdots & 1 \end{pmatrix}$$

$\therefore$ $Var(\epsilon) = \sigma^2 I$ where I is $n \times n$ identity matrix.

3. Non stochastic xs: This implies that all explanatory variables are non stochastic and hence, independent of ϵs.

OLS Estimation (derive OLS for multiple regression model)

Given the population regression model

$$Y = X\beta + \epsilon$$

The sample regression model is

$$Y = X\hat{\beta} + e$$

Here, e is an estimate of ϵ.

And $e = Y - X\hat{\beta}$

The least squares estimators are obtained by minimizing the sum of squares and which is

$$\sum e_i^2 = e_1^2 + e_2^2 + \cdots + e_n^2$$

$$= (e_1 \quad e_2 \quad \cdots \quad e_n) \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix}$$

$$= e^t e$$

$$= (Y - X\hat{\beta})^t (Y - X\hat{\beta})$$

$$= (Y^t - \hat{\beta}^t X^t)(Y - X\hat{\beta})$$

$$= Y^t Y - Y^t X\hat{\beta} - \hat{\beta}^t X^t Y + \hat{\beta}^t X^t X\hat{\beta}$$

$$= Y^t Y - (\hat{\beta}^t X^t Y)^t - (\hat{\beta}^t X^t Y) + \hat{\beta}^t X^t \hat{\beta}$$

$$\therefore \sum e_i^2 = Y^t Y - 2\hat{\beta}^t X^t Y + \hat{\beta}^t X^t \hat{\beta} \quad \text{---------- (1)}$$

For least squares

$$\frac{\partial \Sigma \epsilon_i^2}{\partial \hat{\beta}} = 0$$

$$\Rightarrow \frac{\partial (Y^t Y - 2\hat{\beta}^t X^t Y + \hat{\beta}^t X^t X \hat{\beta})}{\partial \hat{\beta}} = 0$$

$$\Rightarrow -2X^t Y + 2X^t X \hat{\beta} = 0$$

$$\Rightarrow \boxed{(X^t X)\hat{\beta} = X^t Y} \quad \text{---------- (2)}$$

The equations ~~can be~~ contained in 2 are called OLS normal equations in the context of the general linear model.

Therefore,

$$\boxed{\hat{\beta} = (X^t X)^{-1} (X^t Y)} \quad \text{---------- (3)}$$

The vector contains estimators for all unknown parameters.

Software output and Interpretation

* format of ANOVA Table

Source	Dof	SS	MSS	F-Value	P-Value
Regression	k	SSR	MSR = SSR/k	F = MSR/MSE	
Residual or Error	n-k-1	SSE	MSE = SSE/(n-k-1)		
Total	n-1	TSS or SST			

SST: Total sum of squares

n = Sample size, k = no. of independent variables

SSR = explained sum of squares SSE = residual sum of squares

F- Test: Overall Test

t- test : Individual Test

Current Salary	Education (in years)	Experience	Beginning Salary
Y	X_1	X_2	X_3

Individual Test

$H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$ (Overall Test)

There is no significant impact of all independent variable on dependent variable.

H_1: There is a significant impact of at least one independent variable on dependent variable.

p-value $< \alpha$ (0.05) → Reject H_0

p-value $> \alpha$ (0.05) → Accept H_0

Feb 8, 2025

Format of ANOVA Table

Source	dof	Sum of Squares	MSS	F-value	p-value
Regression	k (or p)	SSR	$MSR = \frac{SSR}{k}$	$F = \frac{MSR}{MSE}$	
Error	$n-k-1$	SSE	$MSE = \frac{SSE}{n-k-1}$		
Total	$n-1$	TSS			

↑ Overall Test

Format of Coefficient-Table

Predictor	Regression Coeff.	St. Error	t-value	p-value
Constant	b_0	Sb_0	$t = \frac{b_i}{Sb_i}$	
x_1	b_1	Sb_1		
x_2	b_2	Sb_2		
⋮	⋮	⋮		
x_K	b_K	Sb_K		

↑ Individual Test

(VVI)

From ANOVA table and Coefficient table

✓① Developing an equation and prediction

$$\hat{Y} = b_0 + b_1 x_1 + b_2 x_2 + \dots + b_K x_K$$

✓② Computing multiple coefficient of determination

$$R^2 = \frac{\text{Explained Variation}}{\text{Total Variation}} \quad \begin{matrix} (SSR) \\ (TSS) \end{matrix}$$

TSS = Total Sum of Squares (Total Variation)

SSR = Sum of squares due to regression (Explained

Variation)

SSE = Sum of squares due to error

$$TSS = SSR + SSE$$

Thus,

$$R^2 = \frac{SSR}{TSS}$$

$$= 1 - \frac{SSE}{TSS}$$

$$0 \leq R^2 \leq 1$$

Meaning: Suppose $R^2 = 0.81$; $k = 5$

⇒ 81% of variation in dependent variable is explained by k independent variables.

③ Adj R^2

→ Used for model selection

Y	X_1

R^2

<

Y	X_1	X_2

R^2

<

Y	X_1	X_2	X_3

R^2

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

↓

AIC

VIC

④ Standard error of estimate

$SE(\hat{y}) = S_{yx} = \sqrt{MSE}$

It measures the average variation of observed values of dependent variables around its fitted equation.

⑤ Overall test / Goodness-of-fit

Hypothesis testing for all regression co-efficients simultaneously

Null hypothesis (H_0): $\beta_1 = \beta_2 = \dots = \beta_K = 0$

There is no significant impact of all independent (explanatory) variables on dependent variables.

Alternative hypothesis (H_1): at least one $\beta_i \neq 0$

There is a significant impact of at least one independent variable on dependent variable.

Test-statistic: $F = \frac{MSR}{MSE}$

Calc F = F

Tab F = $F_{K,\, n-K-1}$ at α%

K = dof for numerator

$n-k-1$ = dof for denominator

Decision: If calc $F \leq$ Tab F, we do not reject H_0.

If calc $F >$ Tab F, we reject H_0.

p-value approach (Usually for software)

If p-value $< \alpha$, we reject H_0

If p-value $\geq \alpha$, we do not reject H_0.

Individual test (t-test)

[Co-efficient table].

Null Hypothesis (H_0): $\beta_i = 0$ (i ranges from 1 to k)

There is no significant impact of X_i on Y.

Sales	Price	Ads
Y	X_1	X_2

Alternative Hypothesis (H_1): $\beta_i \neq 0$

There is a significant impact of an independent variable on dependent variable.

Test-Statistic:

$$t = \frac{b_i}{sb_i}$$

Decision: cal $t = |t|$

Tab $t = t_{n-k-1,\ \alpha\%}$

If cal $t \leq$ Tab t, we do not reject H_0.

If cal $t >$ Tab t, we reject H_0.

Software

p-value $< \alpha \Rightarrow$ reject H_0

p-value $\geq \alpha \Rightarrow$ do not reject H_0.

Confidence Interval Estimate

(Coeff table) $\leftarrow b_i \pm t_{n-k-1,\ \alpha}$ (Table → statistical) $sb_i \rightarrow$ Coefficient table

Example

By using following ANOVA table obtained from 30 observations.

Source	df	SS	MSS	F-value
Regression	4	300	MSR = 300/4 = 75	F = 75/8 = 9.375
Error	25	200	MSE = 200/25 = 8	
Total	29	500		

TSS = 500

(a) Complete the given ANOVA table

(b) Compute standard error of estimate & interpret its meaning.

(c) Compute multiple coefficient of determination & interpret its meaning

(d) Compute Adj R^2

(e) Set up null and alternative hypothesis and carry out F-Test.

Solution

(b) $SE(\hat{Y}) = \sqrt{MSE} = \sqrt{8} = 2.828$

The avg. variation of observed values of dependent variable around fitted eqn is **2.828**

(c) $R^2 = \frac{300}{500} = 0.60$ (60% of variation in dependent variable is explained by 4 independent variables)

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

$= 1 - \left(0.4 * \frac{29}{25}\right) = 0.536$

(c) $H_0: \beta_1 = \beta_2 = \cdots = \beta_K = 0$

$H_1: \beta_1 \neq \beta_2 = \cdots \beta_K \neq 0$

Critical value: $F_{tab} = F_{4, 25, 0.05} = 2.759$

Decision: Since $F_{calc}(9.375) > F_{tab}(2.759)$, we reject H_0.

- Mean Vector
- Correlation matrix
- Dispersion

Factor analysis, PCA, Clustering

Data reduction

February 14, 2025

Degrees of freedom = No. of obs - no. of unknown parameters (no. of restrictions)

1) Adj R^2

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

2) Standard Error

$$S_{yx} = \sqrt{MSE}$$

3.

4. Testing significance of individual regression coefficient

$H_0: \beta_i = 0$

$H_1: \beta_i \neq 0$

From software, we get the t-value

c/w

VVI

Q.

Solution

$n = 24$

i)

Predictor	Regression co-efficient b_i	Standard Error sb_i	t	p-value
Constant	6356.17	838.701	7.578	
X_1	0.56038	0.15811	3.5442	0.000
X_2	-31.2077	8.95905	-3.4833	~~0.025~~ ~~0.000~~
X_3	-327.503	149.169	-2.1955	0.001 ~~0.025~~
X_4	-113.895	16.2604	-7.0044	~~0.025~~ ~~0.001~~ 0.000
X_5	-621.458	147.828	-4.2039	0.000

ANOVA

Source	SS	df	MSS	F	p-value
Regression	12204000	5	2440800	F = 20.28	0.000
Residual	2166000	18	120333.333		
Total	14370000	23			

$F_{5,18,0.05} = 2.77$

Decision: Since F_{calc} (20.28) > F_{tab} (2.77), we reject H_0.

ii) Here given $\alpha = 0.05$

$H_0: \beta_3 = 0$

$H_1: \beta_3 \neq 0$

If p-value not given

Here, p-value (0.001) < α (0.05), we reject null hypothesis.

Decision: Thus, β_3 is statistically significant at 5% significance level.

iii)

$H_0: \beta_3 = 0$

$H_1: \beta_3 \neq 0$

Test-statistic: $t = \frac{b_3}{Sb_3} = -2.195$

Cal $t = |t| = 2.195$

Tut $t = t_{n-k-1}$ at α

$= t_{18}$ at 5%.

iii) 99% CI for x_1,

Regression co-efficient = $b_i \pm t_{n-k-1, \alpha} Sb_i$

[NOTE: Interval estimation → Two Tailed]

for square feet heated space,

$b_1 \pm t_{18, 0.01} Sb_1 =$ $\left(1-\alpha = 99\%, \ \alpha = 1\%\right)$

$$= 0.56038 \pm 2.878 * 0.15811$$

$$= 0.56038 \pm 0.45504$$

$$= 0.10534 - 1.01542$$

$$= \boxed{0.1053 \leq \beta_1 \leq 1.0154}$$

↓

Parameter

iv) Standard Error of Estimate

$$SE(\hat{y}) = S_{yx} = \sqrt{MSE}$$

$$= \sqrt{120333.333}$$

$$= 346.890$$

Interpretation: The average variation of observed values of dependent variable around its fitted equation is 346.890.

v) $$R^2 = \frac{SSR}{TSS} = \frac{12204000}{14370000}$$

$$= 0.849$$ (KWh/month)

Int: 84.9% of variation in dependent variable is explained by five independent variables. ($x_1, x_2, \ldots, x_5$)

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

$$= 1 - \left\{ (1 - 0.849) * \frac{23}{18} \right\}$$

$$= 1 - 0.1929$$

$$= \boxed{0.8070}$$ (same)

Interpretation: After adjusting for given degrees of freedom.

vi)

$$Y = 6356.17 + 0.56038 X_1 - 31.2077 X_2 - 327.503 X_3 - 113.895 X_4 - 621.458 X_5$$

for $X_1 = 1295$, $X_2 = 18$, $X_3 = 5$, $X_4 = 3$ and $X_5 = 1$,

$$\hat{Y} = 6356.17 + 0.56038 * 1295 - 31.2077 * 18 - 327.503 * 5 - 113.895 * 3 - 621.458 * 1$$

$$= \boxed{3919.4655}$$

vii)

H_0: $\beta_1 = \beta_2 = \dots = \beta_5 = 0$

H_1: $\beta_i \neq 0$ where $i \in \{1, 2, 3, \dots, 5\}$ — At least one

Critical value:

Calculated F

$$F_{calculated} = \frac{MSR}{MSE}$$

$$= 20.28$$

Critical value: $F_{k,\ n-k-1,\ \alpha} = F_{5,\ 18,\ 0.05} = 2.77$

Decision: Since F_{calc} (20.28) > F_{tab} (2.77), we reject null hypothesis. Thus, there is a significant impact of at least one independent variable on

dependent variable

Feb 15, 2025

BLUE Properties
(But linear Unbiased Estimate)

- Unbiasedness
- Linearity
- Best ness (minimum variance)
- Consistency

Regression Coefficient → β_i → Parameter
(Slope)
→ b_i → Estimator
($\hat{\beta}_i$)

BLUE Properties

① The estimators are linear, that is, they are linear functions of the dependent variable Y. Linear estimators are easy to understand and deal with compared to nonlinear estimators.

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \dots \hat{\beta}_n X_n$$

Linear

Curve, but still linear (due to linear β)

② The estimators are unbiased, that is, in repeated applications of the method, on average, the estimators are equal to their true values.

Example $N = 3$

Variance $= \sigma^2$

2	4
6	

$n = 2$	$\bar{X}$
(2, 4)	3
(4, 6)	5
(2, 6)	4

mean of $\bar{X} = \frac{3+5+4}{3} = 4$

$\sigma^2/n = 1$

Different values for sample means.

$N = 35$

$10 \longrightarrow {}^{35}C_{10}$

$\begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_1 \end{bmatrix}$

Thus, $E(\bar{x}) = \mu$

mean = β_1

* Distribution of all possible sample <u>means</u> and sample <u>variance</u> is known as <u>sampling distribution</u>.

③ • In the class of linear unbiased estimators, OLS estimators have minimum variance. As a result, the true parameter values can be estimated with least possible uncertainty; an unbiased estimator with the least variance is called an efficient estimator.

$35 \rightarrow 10 \longrightarrow {}^{10}C_5$

<u>* Unbiasedness</u>

Given the general linear model

$$\underline{Y} = \underline{X\beta} + \underline{\epsilon}$$

And we know that:

$$\hat{\beta} = (X^T X)^{-1} (X^T Y)$$

"Now substituting $\underline{Y} = X\underline{\beta} + \underline{\epsilon}$ in the equation:

we get,

$$\hat{\beta} = (X^T X)^{-1} X^T (X\beta + \epsilon)$$

$$= (X^T X)^{-1} (X^T X \beta) + (X^T X)^{-1} X^T \epsilon$$

$$= \beta + (X^T X)^{-1} X^T \epsilon$$

Taking expectation on both sides:

$$E(\hat{\beta}) = E(\beta) + (X^T X)^{-1} X^T E(\epsilon)$$

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

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

Bestness

$$\hat{\beta} = \beta + (X^T X)^{-1} X^T \epsilon \quad \text{——— (1)}$$

or, $\hat{\beta} - \beta = (X^T X)^{-1} X^T \epsilon$

The variance of $\hat{\beta}$ is

$$\text{Var}(\hat{\beta}) = E[(\hat{\beta} - \beta)(\hat{\beta} - \beta)^T]$$

$$= E[(X^T X)^{-1} X^T \epsilon \, \epsilon^T X (X^T X)^{-1}]$$

$$= (X^T X)^{-1} X^T E(\epsilon \epsilon^T) X (X^T X)^{-1}$$

$$= (X^T X)^{-1} X^T \sigma^2 X (X^T X)^{-1}$$

$$= \sigma^2 (X^T X)^{-1} (X^T X) (X^T X)^{-1}$$

$$= \sigma^2$$

Measuring goodness-of-fit in multiple regression analysis

$$R^2 = \frac{\text{Explained Variation}}{\text{Total Variation}} = \frac{ESS}{TSS}$$

Adj R^2 penalizes for adding additional explanatory variables in the model so that Adj R^2 values of models having

different number of explanatory variables become comparable.

Other model selection criteria

- Akaike Information Criteria (AIC)

$$AIC = -2 * \log(L) + 2 * k$$

- Schwarz Bayesian Criteria (SBC) (BIC)

$$BIC = -2 * \log(L) + k * \ln(n)$$

- Hannan-Quinn Criteria (HQC)

$$HQ = -2 * \log(L) + 2 * k * \ln(\ln(n))$$

We look for minimum values of these indices while selecting a model.

Multicollinearity

- Multicollinearity refers to a situation in which there is high inter correlations among the explanatory variables of a multiple regression model. It arises only in the context of multiple regressions.

- ANOVA

  F-value ← significant

  t | p → $y \geq 0.05$ (insignificant)

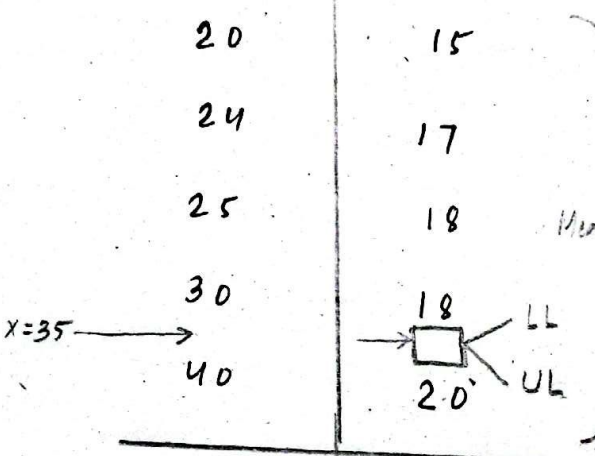
- A good number of the estimated coefficients can be statistically insignificant. This occurs when variances and hence standard errors of the estimated coefficients are large. This is possible when there is little variation in explanatory variables or high inter correlations among the explanatory variables or both.

February 21, 2025

Confidence Interval Estimate and Prediction Interval Estimate for mean and individual response of Y:

Example: Simple Regression Equation

X (Income) in '000	Y (Expenditure) in '000	XY	X^2	Y^2
20	15	300	400	225
24	17	408	576	289
25	18	450	625	324
30	18	540	900	324
40	20	800	1600	400
$\Sigma X = 139$	$\Sigma Y = 88$	$\Sigma XY = 2498$	$\Sigma = 4101$	$\Sigma = 1562$



$$b_1 = \frac{n\Sigma XY - (\Sigma X)(\Sigma Y)}{n\Sigma X^2 - (\Sigma X)^2} = \frac{5 \times 2498 - 139 \times 88}{5 \times 4101 - (139)^2}$$

$$= \boxed{0.2151}$$

$$b_0 = \frac{\Sigma Y}{n} - \frac{b_1 \Sigma X}{n} = \frac{88}{5} - \frac{0.2151 * 139}{5}$$

$$= \boxed{11.623}$$

The estimating equation / regression equation / best line / line of best fit is

$$\hat{Y} = 11.623 + 0.2151 X$$

(b_0 : 11.623, b_1 : 0.2151)

when $x = 35$ (in '000)

$$\hat{Y} = 11.623 + 0.2151 * 35$$

$$= 19.1515 \text{ (in '000)}$$

---

hat matrix element (h_i)

$$h_i = \frac{1}{n} + \frac{(x^* - \bar{x})^2}{\Sigma (x - \bar{x})^2}$$

$$= \frac{1}{n} + \frac{(x^* - \bar{x})^2}{\Sigma x^2 - n\bar{x}^2}$$

$$\bar{x} = \frac{\Sigma x}{n} = \frac{139}{5} = \underline{27.8}$$

Now, $h_i = \frac{1}{5} + \frac{(35 - 27.8)^2}{4101 - 5*(27.8)^2}$

$$= \underline{0.4189}$$

Confidence interval estimate for mean response of Y (Always two-tailed)

$$\hat{Y} \pm t_{n-2,\alpha} \, S_{yx} \sqrt{h_i}$$

$$LL \leq \mu_{Y|x} \leq UL$$

Prediction interval estimate for individual response of

$\hat{Y} \pm t_{n-2,\alpha} S_{YX} \sqrt{1+h_i}$

$LL \leq Y_i \leq UL$

$$S_{YX} = \sqrt{\frac{\Sigma (Y - \bar{Y})^2}{n-2}}$$

$$= \sqrt{\frac{\Sigma Y^2 - b_0 \Sigma Y - b_1 \Sigma XY}{n-2}}$$

Now, $S_{YX} = \sqrt{\frac{1562 - 11.623 \times 88 - 0.2151 \times 2498}{5-2}}$

$= 0.8378$

for mean response of Y,

Now, $\hat{Y} \pm t_{n-2,\alpha} S_{YX} \sqrt{h_i}$

$19.1515 \pm t_{3, 0.05} * 0.8378 \sqrt{0.4189}$

$= 19.1515 \pm 3.182 * 0.5422$

$= 19.1515 \pm 1.7254$

$= 17.4261 - 20.8769$

$\therefore 17.4261 \leq \mu_{YX} \leq 20.8769$

for individual response of Y,

$\hat{Y} \pm t_{n-2,\alpha} S_{YX} \sqrt{1+h_i}$

$= 19.1515 \pm 3.182 * 0.8378 * \sqrt{1+0.4189}$

$= 19.1515 \pm 3.1757$

$\therefore 16.3242 \leq Y_i \leq 22.3256$

In Multiple regression analysis,

Confidence interval estimate for mean response of Y is

$$\hat{Y} \pm t_{n-p-1,\alpha} \, S_{YX} \sqrt{hi}$$

$$LL \leq \mu_{YX} \leq UL$$

Prediction interval for individual response of Y is

$$\hat{Y} \pm t_{n-p-1,\alpha} \, S_{YX} \sqrt{1+hi}$$

where, p = no. of explanatory (independent variables.

EXAM

EXAMPLE

Suppose, $b_0 = 33$, $b_1 = 2.15$, $b_2 = 7.15$, $b_3 = 0.008$

$Sb_0 = 10.5$, $Sb_1 = 0.18$, $Sb_2 = 1.28$, $Sb_3 = 0.001$

$n = 30$, $\Sigma (Y - \bar{Y})^2 = 20.18$; $hi = 0.131$

a) Fit a regression line

b) Predict Y when $X_1 = 10$, $X_2 = 5$, $X_3 = 7$ and also construct 95% confidence interval estimate for mean response of Y.

c) Also construct 95% prediction interval estimate for individual ruponse of Y.

d) Test the significance of impact of X_2 on Y

e) Construct 95% CI estimate for regression coefficient of X_2.

Solution

Here, $p = 3$

a) Regression line,

$$\hat{Y} = 33 + 2.15 X_1 + 7.15 X_2 + 0.008 X_3$$

b) when $X_1 = 10$, $X_2 = 5$, $X_3 = 7$,

$$\hat{Y} = 33 + 2.15 * 10 + 7.15 * 5 + 0.008 * 7$$

$$= 90.306$$

$$S_{YX} = \sqrt{\frac{\Sigma (Y - \bar{Y})^2}{n - p - 1}} = \sqrt{\frac{20.18}{30 - 3 - 1}} = 0.880$$

Now, 95% CI estimate is:

$$\hat{Y} \pm t_{n-p-1, \alpha} \, S_{YX} \sqrt{h_i}$$

$$= 90.306 \pm t_{26, 0.05} * 0.880 * \sqrt{0.131}$$

$$= 90.306 \pm 2.056 * 0.880 * \sqrt{0.131}$$

$$= 90.306 \pm 0.654$$

$$\therefore 89.652 \leq \mu_{YX} \leq 90.960$$

c) PI estimate,

$$\hat{Y} \pm t_{n-p-1, \alpha} \, S_{YX} \sqrt{1 + h_i}$$

$= 90.306 \pm 2.056 * 0.880 * \sqrt{1 + 0.131}$

$= 90.306 \pm 1.924$

$\therefore \quad 88.382 \leq Y_i \leq 92.23$

d) $H_0: \beta_2 = 0$

$H_1: \beta_2 \neq 0$

Cal t, $|t| = \frac{b_2}{sb_2} = \frac{7.15}{1.28} = 5.585$

Table Test $t = t_{n-p-1, \alpha}$

$= t_{26, 0.05}$

$= 2.056$

Decision: Since cal $t >$ Table t, we reject H_0.
Thus, there is a significant impact of X_2 on Y.

e) For X_2,
Regression co-efficient $= b_i \pm t_{n-p-1; \alpha} \, sb_i$

for $i = 2$,

Regression coefficient $= b_2 \pm t_{26, 0.05} \, sb_2$

$= 7.15 \pm 2.056 * 1.28$

$= 7.15 \pm 2.63$

$\therefore$ $\boxed{4.52 \leq \beta_2 \leq 9.781}$

$\therefore$ $4.52 \leq \beta_2 \leq 9.781$

March 8, 2025

Multicollinearity

Multicollinearity refers to a situation where there are high inter correlations among the explanatory variables of a multiple regression model. Multicollinearity problem arises only in the context of multiple regressions, it is considered as a problem because when the explanatory variables are highly correlated, most of their variation is common so that there is little variation unique to each variable.

Consequences of multicollinearity

- Perfect multicollinearity
- Under perfect ~~not~~ multicollinearity it is not possible to compute the values of ols estimates.
- Imperfect multicollinearity

B L U E

↓ ↓ ↓

But linear Unbiased Estimate

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2 + \ldots + b_p X_p$$

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2 + (b_3 X_3) + b_4 X_4 + b_5 X_5$$

(Excluded)

Imperfect multicollinearity

→ Correlation between two independent variables → High

→ Most of the independent variables → Insignificant

F-test → reject

t-test → do not reject (insignificant).

#Detection of multicollinearity

1. Correlation matrix ⟶ Gives idea ; Not utilized for final decision
2. F-statistic and t-statistic
3. Klen's rule of thump
4. Variance Inflation Factor (VIF)
5. Tolerance method (TOL)

- VIF = 1 (no multicollinearity)
- 1 < VIF < 5 (less multicollinearity)
- 5 ≤ VIF ≤ 10 (Moderate multicollinearity)
- VIF > 10 (High multicollinearity)

Two-way Dendrogram

Correlation matrix →

	Y	X_1	X_2	X_3	X_4
Y	(1)				
X_1		1			
X_2					
X_3					
X_4					(1)

Remedial Technique

1) By increasing sample size
2) Transformation of variables → Using literature review.
   Eg: ln transformation, Inverse transformation etc
3) Using extraneous estimates
4) Dropping the variables.
   - We have to drop the variable having highest VIF (>10) at a time. Again run the regression of remaining variables check whether there is VIF > 10 or not and repeat the same process until we get VIF < 10.

Cobb - Douglas Production

- To illustrate the CD function, we use Cobb data on output (as measured by value added, in thousands of dollars), labor input (worker hours, in thousands), and capital input (capital expenditure, in thousands of ~~data~~ dollars) for the US manufacturing sector.
- The data is cross - sectional, covering 50 states and Washington, DC, for the year 2005.

log(Labor) 0.47

The interpretation of the coefficient of log(Labor) of about 0.47 is that if we increase the labor input by 1% on average, output goes up by about 0.47%.

#AUTOCORRELATION & HETEROSKEDASTICITY

Autocorrelation

- One of the assumptions of the Classical linear Regression Model (CLRM) is that the disturbance term of the model is independent.

$$Cov(\epsilon_t, \epsilon_s) = E(\epsilon_t, \epsilon_s) = 0 \text{ for } t \neq s$$

This feature of regression disturbance is known as serial independence or non autocorrelation. It implies that the value of disturbance term in one period is not correlated with its value in another period.

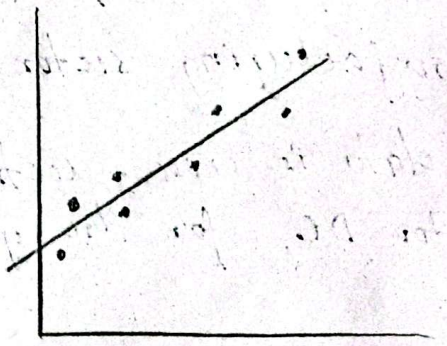
In time series, the disturbance term at period t may be related with the disturbance term at t-1, t-2, ..., and t+1, t+2, ... and so on. In that case, $Cov(\epsilon_t, \epsilon_s) \neq 0$ for $t \neq s$ and we say that the disturbances are autocorrelated.

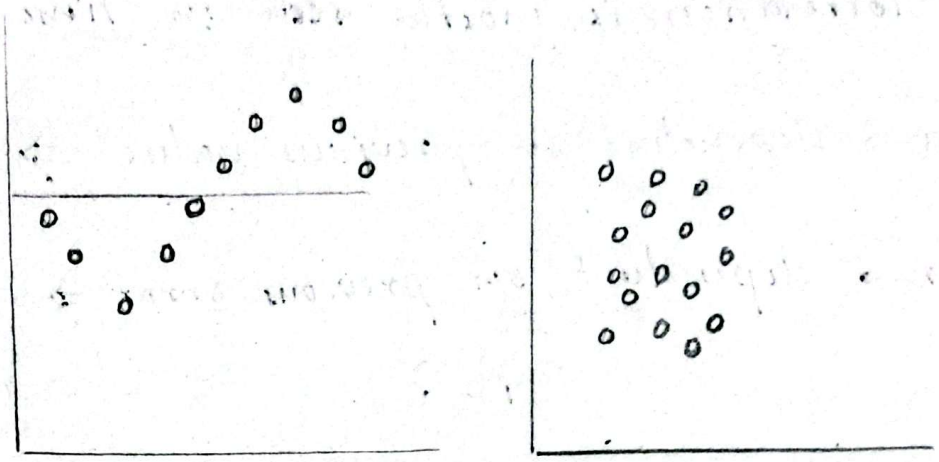
$Y - \hat{Y} = e$

$E(e_i) = 0$

$Var(e_i) = Constant$

$Cov(e_i, e_j) = 0$





Pattern Visible

No pattern / Random

Important $\Rightarrow$ Assumptions + BLUE Properties

> Bestness: Minimum Variance

$\rightarrow$ Impact on Bestness

> Impact on test / Hypothesis testing

Y	X_1	X_2	X_3	...	X_p	$\hat{Y}$	$Y - \hat{Y}$
							+
							+
							−
							−
							+
							−
							+
							−
							−
							−
							+

This is preferable

Pattern observed

However, in real data, multicollinearity and auto correlation are preferred / useful.

* Autocorrelation is mostly seen in <u>time-series data</u>.

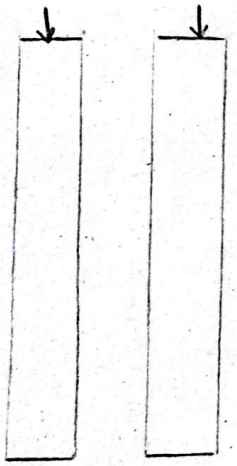
data → dependent on previous value ⇒ Autoregression

error → dependent on previous error ⇒ Autocorrelation

$Y - \hat{Y}$

ARIMA — Autoregression Integrated Moving Average
(Dependency on Previous Error)

lag-1 Correlation



Equilibrium

$$Y_t = \underbrace{X_t + X_{t-1} + X_{t-1}}_{\text{Distributed lag}}$$

ARDL Model

$$\underbrace{Y_t = Y_{t-1} + Y_{t-2} + X_t + X_{t-1}}_{\text{ARDL model}}$$

Specification of Autocorrelation Relationship

Mostly used in Time Series

$$\epsilon_t = \rho \epsilon_{t-1} + u_t \quad \text{——— (1)}$$

$\downarrow$

White noise

$E(u_t) = 0$

$Var(u_t) = E(u_t^2) = \sigma_u^2$ for all t.

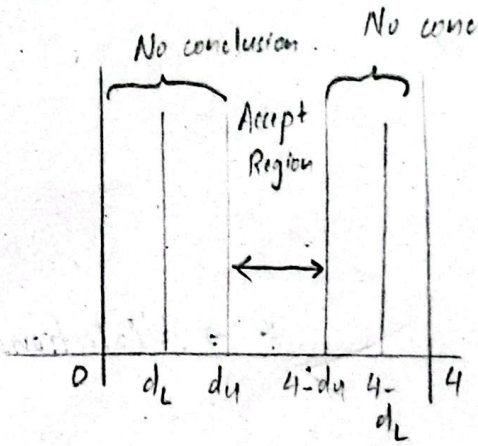
u_t is normally distributed

$E(u_t u_{t-1}) = 0$

$\Downarrow$

No autocorrelation

Durbin-Watson Test $\longrightarrow$ Autocorrelation Test



Further test $\Longrightarrow$ LM Test

$$\epsilon_t = \rho \epsilon_{t-1} + v_t$$

$$\epsilon_{t-1} = \rho \epsilon_{t-2} + v_{t-1}$$

$\vdots$

$$\epsilon_t = u_t + \rho v_{t-1} + \rho^2 u_{t-2} + \rho^3 u_{t-3} + \ldots \quad \text{——— (2)}$$

This shows that under the first-order autoregressive scheme the

effect of past disturbance wears of gradually as $|\rho| < 1$.

where, ρ represents the correlation between ϵ_t and ϵ_{t-1} (first order auto-correlation), ρ^2 is correlation coefficient between ϵ_t and ϵ_{t-2} (second order autocorrelation) and so on.

- $\rho = 0 \Rightarrow$ no autocorrelation
- strength of autocorrelation becomes high as ρ approaches to unity (+1).
- strength of autocorrelation becomes high again as ρ approaches to -1.

Mean

$$E(\epsilon_t) = 0$$

*Variance of ϵ_t

$$Var(\epsilon_t) = \frac{\sigma_u^2}{1-\rho^2} = \sigma_\epsilon^2$$

> * → Violations

*Covariance of ϵ_t and ϵ_{t-1}

$$Cov(\epsilon_t \epsilon_{t-1}) = \rho \sigma_\epsilon^2$$

Assumptions
+
BLUE Properties

March 14, 2015

Consequences of autocorrelation are:

i) The OLS estimators are still unbiased and consistent.

ii) The OLS estimators are no longer minimum variance or best estimators. Hence, they are not efficient and BLUE

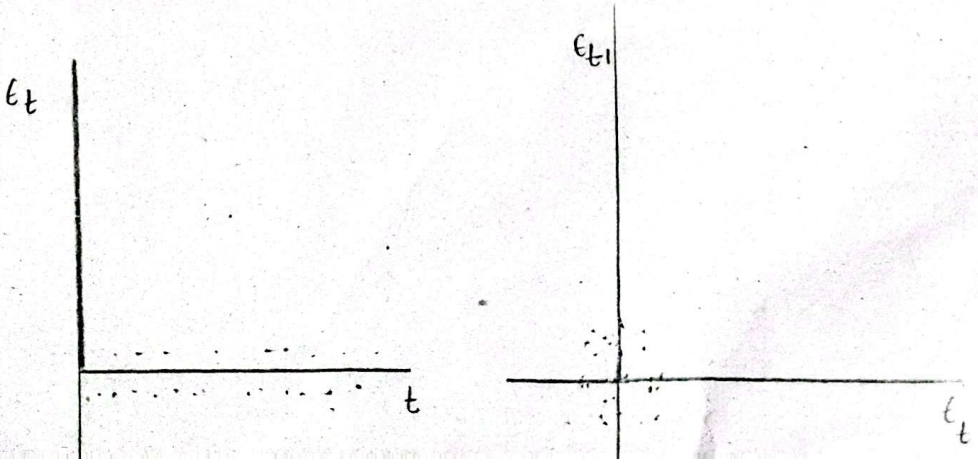
iii) If we disregard the problem of autocorrelation and believe that all assumptions are valid, following problems will arise

* The estimated variance of disturbance term will be under estimate of its true variance.
* The standard error of the estimated slope coefficient will be much smaller if it is computed with usual OLS formula. This will provide spurious (wrong) impression about the statistical significance.
* The usual t and F test will become invalid.

Detection (of autocorrelation)

- Graphical Method (Residual Plot)
  - Graph of ϵ_t against time
  - Graph of ϵ_t against ϵ_{t-1}

i) $\rho = 0$



(IMP) Durbin-Watson Test (D-W Test)

1 Numerical

This is the simplest and most widely used test for autocorrelation. It is based on the following assumptions:

a. The regression model includes a constant or intercept term.

b. We are examining presence of first order autocorrelation.

c. The regression model does not include a lagged dependent variable as an explanatory variable. Now, to understand how the test is performed, consider the model:

$$Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + \cdots + \beta_K X_{Kt} + \epsilon_t$$

where $\epsilon_t = \rho \epsilon_{t-1} + v_t \quad |\rho| < 1$

Null hypothesis (H_0): $\rho = 0$

or, $H_0: \rho \geq 0$

or, $H_0: \rho \leq 0$

There is no significant autocorrelation.

Alternative hypothesis (H_1): $\rho \neq 0$

or, $H_1: \rho < 0$

or, $H_1: \rho > 0$

There is a significant autocorrelation.

Test-statistic

$$\text{D-W statistic} = \frac{\sum_{i=2}^{n} (e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2}$$

For tabulated value,

Given, n, α,

k' (or p) = number of independent variables

= 1 (if it is not given)

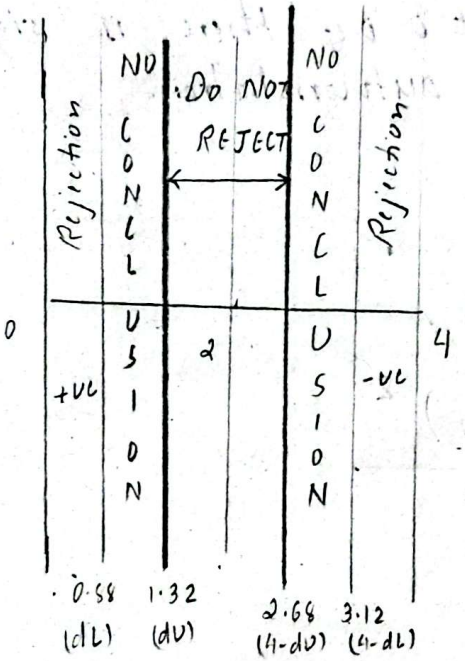
Example

$\alpha = 0.05$

	k = 1		k = 2		k = 3	
n	dL	dU	dL	dU	dL	dU
10	0.88	1.32				

n = 10
k = 1
α = 0.05

Decision Criteria



Example: By using following information

Period	Error / Residual (e_i)	e_{i-1}	$(e_i - e_{i-1})^2$	e_i^2
1	+5	—	—	25
2	+4	+5	1	16
3	+3	+4	1	9
4	+2	+3	1	4
5	+1	+2	1	1
6	-1	+1	4	1
7	-2	-1	1	4
8	-3	-2	1	9
9	-4	-3	1	16
10	-5	-4	1	25
			$\sum_{i=2}^{n}(e_i - e_{i-1})^2 = 12$	$\sum e_i^2 = 110$

At $\alpha = 0.05$, test the presence of autocorrelation.

$$\sum_{i=1}^{10} e_i^2 = 110$$

Null Hypothesis (H_0): $\rho = 0$ i.e. there is no significant autocorrelation.

Alternative Hypothesis (H_1): $\rho \neq 0$ i.e. there is significant autocorrelation.

Test-statistic:

$$\text{D-W statistic} = \frac{\sum_{i=2}^{n} (e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2}$$

$$= \frac{12}{110}$$

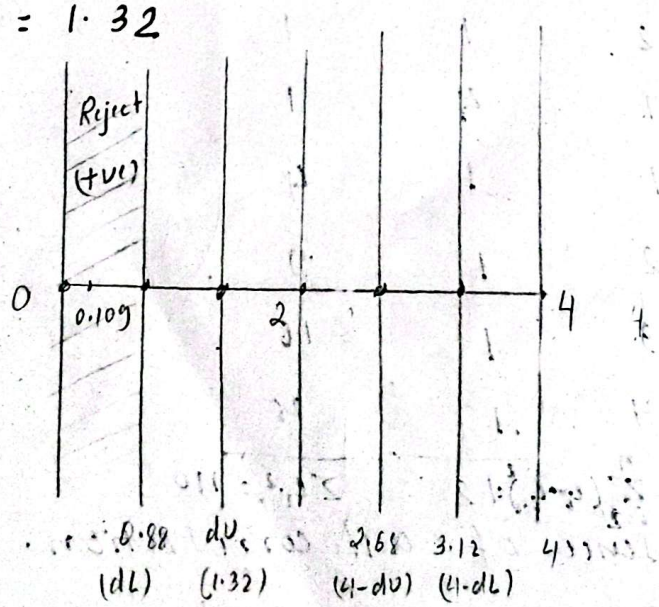
$$= 0.109$$

From Table

$n = 10$, $\alpha = 0.05$, $k = 1$

$dL = 0.88$

$dU = 1.32$



Decision:
Here, $0 < DW = 0.109 < dL$ (0.88), we reject H_0

i.e. There is a significant autocorrelation.

Limitations

1) It can not be used for testing higher order autocorrelation.
2) This test is biased

Breusch - Godfrey Lagrange Multiplier Test (LM test)

Null Hypothesis (H_0): $\rho_1 = \rho_2 = \rho_3 = \dots = \rho_p = 0$ [no autocorrelation]
(second order autocorrelation, third order autocorrelation)

Alternative Hypothesis (H_1): At least one ρ is not zero

$$\rho_i \neq 0 \quad ; \quad i \in \{1, 2, \dots, p\}$$

(IMP)

Steps

1) Estimate eq^n (1) by OLS and estimate $\hat{\epsilon}_t = e_t$
2) Run the auxiliary regression of e_t on $x_{1t}, x_{2t}, \dots, x_{kt}$, $e_{t-1}, e_{t-2}, \dots e_p$

3) To test the validity of above null hypothesis compute LM statistic as follows.

$$LM = (n-p)R^2$$

where R^2 is coefficient of determination of auxiliary equation.

n is number of observations

p is order of autocorrelation.

The LM statistic follows a Chi-Square distribution with degree of freedom p.

March 15, 2025 (EXAM)

HETEROSKEDASTICITY

For the sample of two variable model

$$Y_i = \alpha + \beta X_i + \epsilon_i \quad \text{——— (1)}$$

We assumed the variances of disturbance term ϵ_i is constant for all observations

i.e. $Var(\epsilon_i) = E(\epsilon_i^2) = \sigma^2$ (constant for all i)

——— (2)

- This feature of disturbances term of the regression model is known as <u>homoskedasticity</u>.

However, it is quite common in regression analysis to have cases where the variance of disturbance term becomes variable rather than ~~the~~ remaining constant.

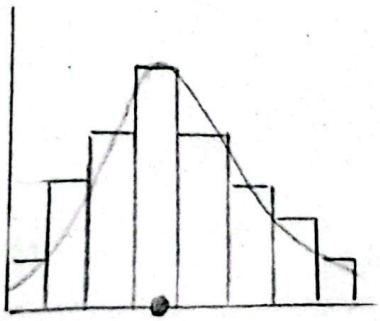
- In this situation the disturbance is said to be heteroskedasticity.

- i.e. $\text{Var}(\epsilon_i) = E(\epsilon_i^2) = \sigma_i^2$ ——— (3)
- which means that the variance of disturbance term can change for every different observation in the sample $i = 1, 2, \ldots, n$.

Marketing Research
- Naresh Malhotra

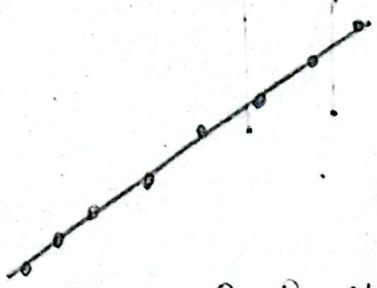
Y	X_1 ...	X_K	$\hat{Y}$	e ($Y - \hat{Y}$)

e
↓
Normality Test



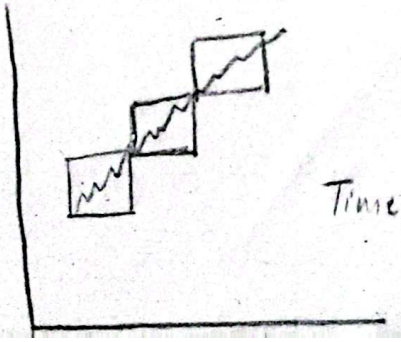
Histogram

Mean:
Median:
Mode

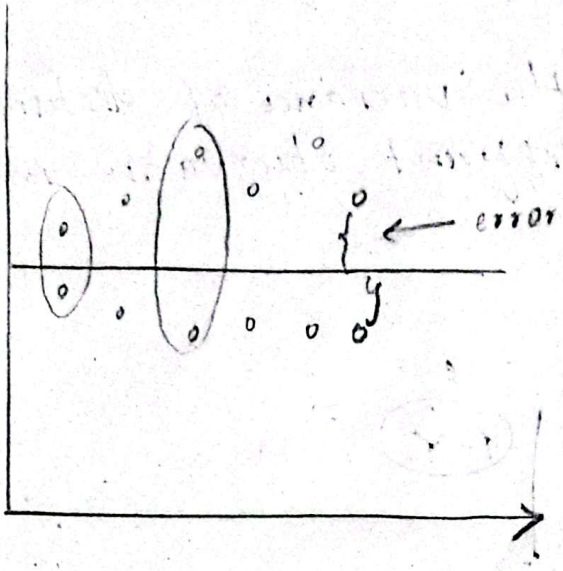


Q - Q plot

$e_i \sim N(0, \sigma^2)$
↓
Variance should be constant
↓ Violation
leads to heteroskedasticity.



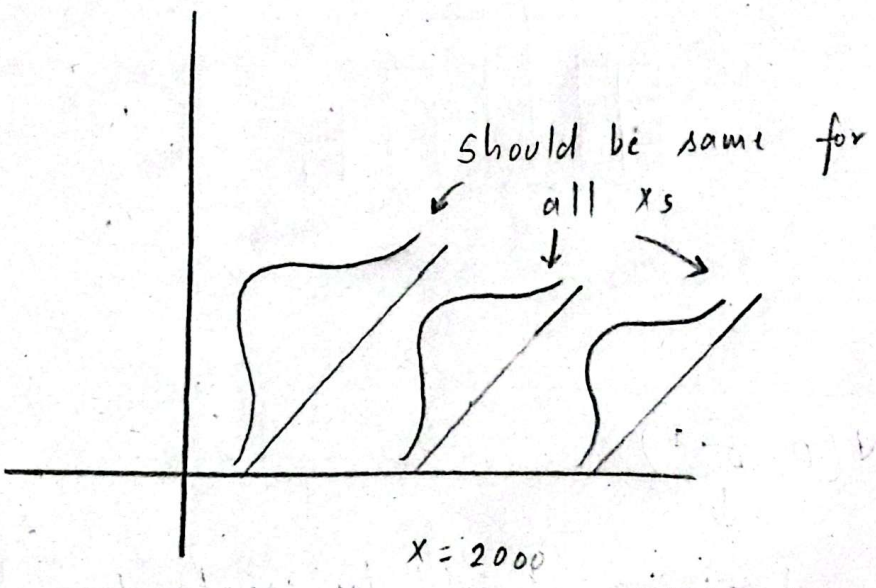
Time Series Data



Income | Expenditure

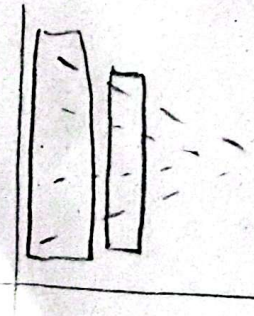
X	Y	$\hat{Y}$	e
20			
20			
20			
20			

$\hat{Y} = \beta_0 + \beta_1 X$



$Var(\epsilon_i^2) = \sigma^2 \longrightarrow$ Homoskedasticity

$Var(\epsilon_i^2) = \sigma_i^2 \longrightarrow$ Heteroskedasticity



like Flashlight
(Heteroskedasticity)

Sources of Heteroskedasticity

i) When we are dealing with micro-economic or cross-section data, we are very likely to have a heteroskedasticity problem.

ii) Presence of outliers

iii) If some relevant variables have been mistakenly omitted.

iv) Inclusion of explanatory variables in the model whose distributions are skewed.

v) Heteroskedasticity may also arise due to incorrect data transformation.

Consequences

Unbiasedness

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

This shows that $\hat{\beta}$ remains unbiased when the disturbance term of the model ϵ_i is heteroskedasticity.

Bestness

Thus $Var(\hat{\beta})$ | heteroskedasticity $>$ $Var(\hat{\beta})$ | homoskedasticity

So, there is no longer minimum variance and hence not best estimator.

$\hat{\beta}$ is unbiased but not the best.

Consistence

$\hat{\beta}$ is consistent when the disturbance term is

- The OLS estimators continue to remain unbiased and consistent under heteroskedasticity.
- Heteroskedasticity increases the variances of the distributions of estimator of β thereby turning the OLS estimators inefficient (not best).

Heteroskedasticity also affects the variance of OLS and their standard error. In fact, the presence of ~~heteroskedasticity~~ heteroskedasticity, in general causes the OLS method to understimate the variances and hence standard error of the estimators. As a sequence, we have higher that expected values of t and F statistic.

As the OLS estimators are unbiased under heteroskedasticity

Detection Techniques

Graphical method:

Plotting square ~~do~~ of residuals (ϵ_i^2) against explanatory variable (X_i) to which it is suspected the disturbance a variance is related. Since ϵ_i^2 is unknown, its proxy measure $\epsilon_i^2 = e_i^2$

Breusch-Pagan - Godfrey Test (EXAM) + Steps

They developed a dagrange Multiplier (LM) test to examine the presence of heteroskedasticity in data:

Considering the model.

$$Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_K X_{Ki} + \epsilon_i \quad \text{— (1)}$$

And suppose that

$$\text{Var}(\epsilon_i) = \sigma_i^2 = f(\gamma_0 + \gamma_1 Z_{1i} + \cdots + \gamma_r Z_{ri})$$

Goldfeld – Quandt Test

Goldfeld - Quandt Test (1965) proposed a test of heteroskedasticity that may be applied when one of the explanatory variables is suspected to be the heteroskedasticity culprit.

- We assume $\sigma_i^2 \propto x_i$ and also $\epsilon_i \sim$ Normal distribution.

The hypothesis are:

Null Hypothesis (H_0): $\text{Var}(\epsilon_i | x_i) = \sigma^2$, a constant (Homoskedasticity)

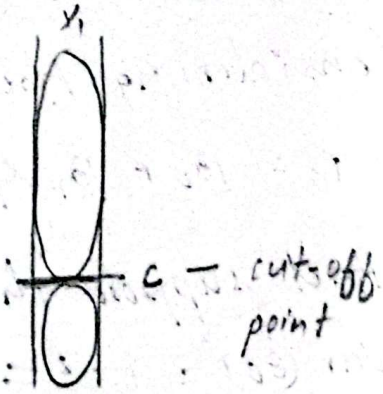
Alternative Hypothesis (H_1): $\text{Var}(\epsilon_i | x_i) = \sigma_i^2$, a variable (Heteroskedasticit

Steps

1) Identify the variable to which the variance of disturbance term is suspected to be related.

2) Sort the raw data in ascending order (starting with lowest and going to be highest) of the values of x_i.

3) Cut out some central observations (c), breaking data into



limitations

1) Determination of appropriate value of c.
2) Difficulty in identifying x- variable

#Remedial Techniques

- log Transformation
- Weighted least Squares Method

Generalized least Squares (GLS)

Consider the model:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \epsilon_i \quad \text{——— (1)}$$

And suppose that there is heteroskedasticity

Packages

car ⟶ qq plot
lmtest ⟶ for BP Test

April 4, 2025

[Regression Analysis]

Dummy Variable

$$b_1 = \frac{n\Sigma XY - \Sigma X \Sigma Y}{n\Sigma X^2 - (\Sigma X)^2}$$

$$= \frac{8 \times 141 - 4 \times 236}{8 \times 4 - (4)^2}$$

$$= 11.5$$

$$b_0 = \bar{Y} - b_1 \bar{X}$$

$$= \frac{236}{8} - 11.5 \times \frac{4}{8}$$

$$= 29.5 - 5.75$$

$$= 23.75$$

$$\hat{Y} = 23.75 + 11.5X$$

Ex 1

Salary (000)	Gender	XY	X^2
30	1	30	1
18	0	0	0
20	0	0	0
34	1	34	1
40	1	40	1
25	0	0	0
32	0	0	0
37	1	37	1
$\Sigma Y = 236$	$\Sigma X = 4$	$\Sigma_{XY} = 141$	$\Sigma X^2 = 4$

1 Male
0 Female

$\bar{Y} = 29.5$
$\bar{X} = 0.5$

Ex 2

Y Salary (000)	Gender (X)	XY	X^2
30	0	0	0
18	1	18	1
20	1	20	1
34	0	0	0
40	0	0	0
25	1	25	1
32	1	32	1
37	0	0	0

$\Sigma X = 4$ ↓ $\Sigma XY = 95$

1 Female
0 Male

$$\hat{Y} = 35.25 - 11.5X$$

$$b_1 = \frac{8 \times 95 - 4 \times 236}{8 \times 4 - (4)^2}$$

$$= -11.5$$

$$b_0 = \bar{Y} - b_1 \bar{X}$$

$$= 29.5 - (-11.5) * \frac{4}{8}$$

$$= 29.5 + 5.75$$

$$= 35.25$$

$\Sigma X^2 = 4$

for 1st model,

$\hat{Y} = 23.75 + 11.5X$

when $X = 1$,

$\hat{Y} = 23.75 + 11.5(1) = \underline{35.25}$ ('000) Average salary of male.

when $X = 0$,

$\hat{Y} = 23.75 + 11.5(0) = \underline{23.75}$ ('000). Average salary of female.

Binary number, Dichotomous variable

Dummy variable: i) Any variable having two options, denoting them by 1 ~~and~~ or 0, is known as dummy variable.

ii) We can include dummy variable as an independent variable in regression analysis.

In second model,

$\hat{Y} = 35.25 - 11.5X$

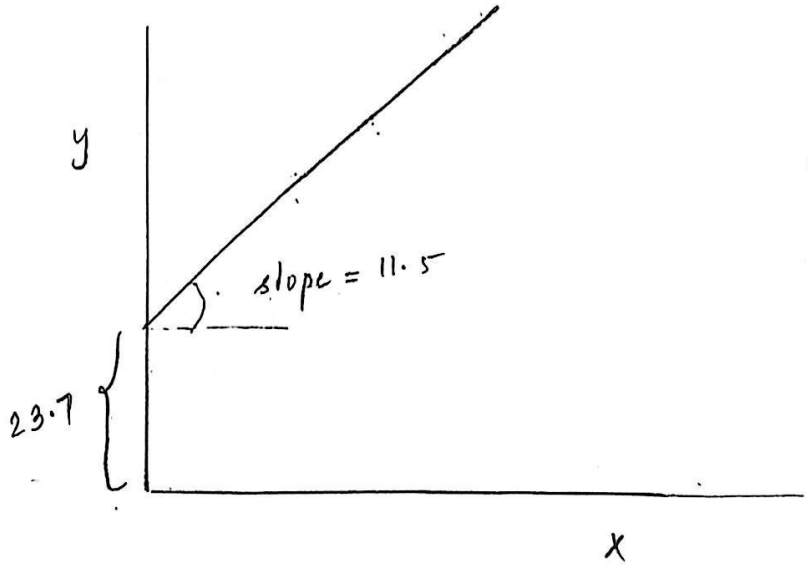
when $X = 1$,

$\hat{Y} = 35.25 - 11.5$

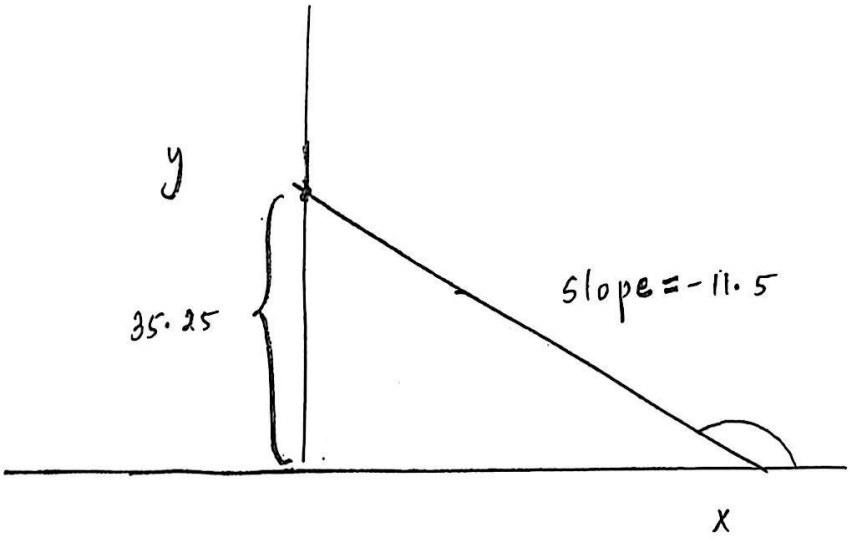
$= 23.75$

when $X = 0$,

$\hat{Y} = 35.25 - 11.5(0) = 35.25$



$\hat{y} = 23.75 + 11.5x$



$\hat{y} = 35.25 - 11.5x$

$H_0: \beta_1 = 0$

$H_1: \beta_1 \neq 0$

$t = \frac{b_1}{sb_1}$ reject $\Rightarrow$. There is a significant impact of gender on salary.

$\Rightarrow$ Salary discrimination between male and female.

Eg: I am satisfied with the mileage.

Toyota – 1

Kia – 0

While encoding,

0 → Reference category

Q. Salary ——————

Post

Assistant [1]

Officer [2]

Manager [3]

Y Salary	X Post
Doesn't give correct results!!!	1
	2
	3

(Y) Salary	X_1 Assistant	X_2 Officer	X_3 Manager
	1	0	0
	0	1	0
	0	0	1
	1	0	0

$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2 + b_3 X_3$ we can not get the result / outcome.

(Perfect Multicollinearity)

↓

Dummy variable trap.

$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2$ ⟶ (left variable is reference category)

⟶ compared against Manager.

If categorical variable consists of k number of options / groups / categories, we have to make (k-1) dummy variables.

Share price (index)	Sun	Mon	Tues	Wed

Dummy Variable dependent — Can't be estimated using OLS

↓ We use

logistic / Probit

Logit / Probit

Multinomial Logistic Regression

Demographic Health Survey ← DHS → Worldwide Health data

Nepal MICS

↓

Multiple Indicators Cluster Survey

April 5, 2025

Carter and Hue

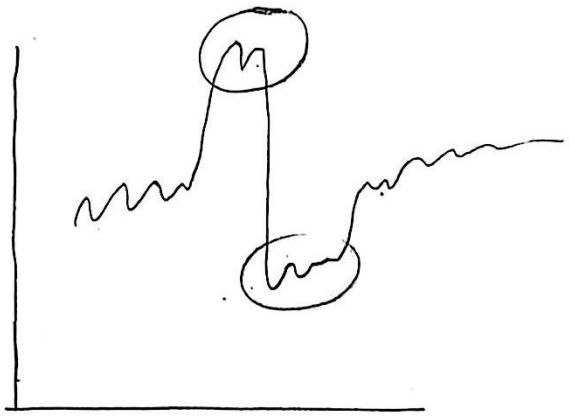
R Practicals

Categorical regression

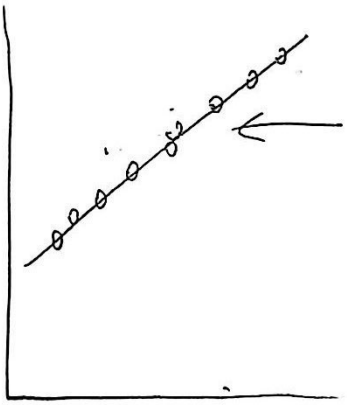
Autocorrelation

Heteroskedasticity

Dummy Variable example (Wage)



QQ-Plot



April 12, 2025 (Regression)

Econometrics Hilkast Book

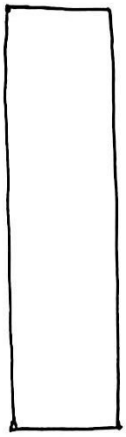
Categorical Data Analysis

Dependent	Independent (Explanatory)
↓ one	Multi-collinearity: $X_1, X_2, \ldots, X_K$ ↓ numerical / category ↓ dummy
Regression / Econometrics	

error

$Y - \hat{Y}$

e



autocorrelation + heteroskedastic

ARIMA: Auto-regression Integrated Moving Average

Dependent	Independent (Explanatory)
Categorical ↓ Logistic / Probit	
Discriminant	
Multinomial	
Ordinal	

dogistic / Probit

↓

Dependent Variable

↓

Categorical / Binary (0 / 1)

Dependent

↓

Categorical

→ Discriminant (Multi-variate analysis)

Dependent Variable

↓

Ranking (Ordinal)

↓

Ordinal (or multinomial)

GDP	Remittance
Y_t	X_t

Cointegration

Y_t = Y_{t-1} Y_{t-2} X_t X_{t-1} X_{t-2}

AR — Auto-Regression

DL — Distributed Lag

Distributed lag Model

AR - DR model

Logistic Regression

Models with dummy dependent variable. Three important approaches:

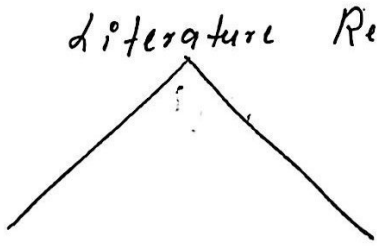
Linear Probability Model (LPM)

Logit Model

Probit Model

Example

Factors affecting for poverty in Nepal



Reading

- Topic selection
- Report writing

Writing in-text

- Related Topic
  - Text book
  - Edited book
  - Articles
  - Journals
  - Thesis
  - Opinions
  - Meanings
  - Findings
  - Frameworks

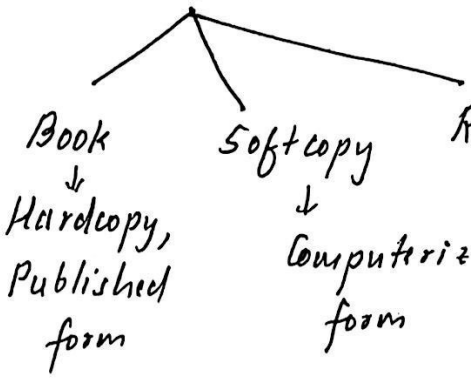
Citation (in-text, reference)

World Bank - "Poverty and Inequality"
Understanding Poverty

dependent variable	Independent variable
↓	
1	
0	

Factors affecting employment status in Nepal: NLFS

Factors affecting poverty in Nepal:
NLSS



Computerized form:
Internet - Free downloadable
Online - Restricted
Offline -
Pendrive, CDs etc

Logistic Regression

→ Binary logistic regression is a form of regression which is used when the dependent variable is a true or forced dichotomy and the independent variables are of any type.

Employment Status : Not in labor force
Employed
Unemployed

→ dogistic regression can be used to predict a categorical dependent variable on the basis of continuous and/or categorical independent variables; to determine the effect size of the independent variables on dependent variable

→ dogistic regression applies maximum likelihood estimation after transforming the dependent into a logit variable.

→ A logit is the natural log of the odds

No multicollinearity

Y	X_1 X_2 ...
1	
0	
0	
1	

(IMP)

Specification of the model $\pi_i = P(1) = P_i$ success →

$$P_i = P(Y_i = 1) = F(Z_i) = \frac{1}{1+e^{-Z_i}}$$

where, P_i is the probability

$$1 - P_i = 1 - \frac{1}{1+e^{-Z_i}} = \frac{1+e^{-Z_i}-1}{1+e^{-Z_i}} = \frac{e^{-Z_i}}{1+e^{-Z_i}}$$

$$= \frac{1}{\frac{1+e^{-Z_i}}{e^{-Z_i}}} = \frac{1}{1+e^{Z_i}}$$

$$\frac{P_i}{1-P_i} = \frac{\frac{1}{1+e^{-z_i}}}{\frac{1}{1+e^{z_i}}} = \frac{1+e^{z_i}}{1+e^{-z_i}} = \frac{1+e^{z_i}}{1+\frac{1}{e^{z_i}}}$$

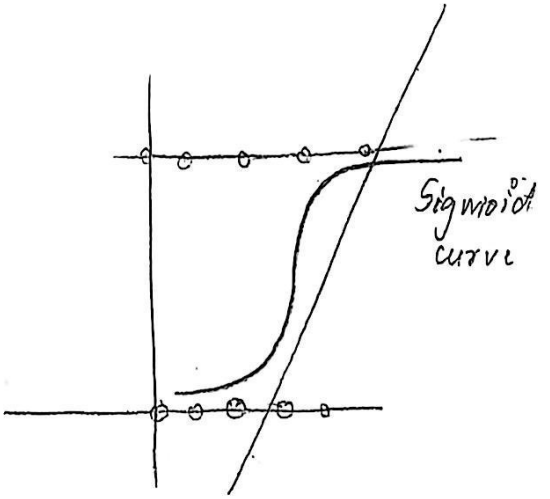
$$= \frac{1+e^{z_i}}{\frac{1+e^{z_i}}{e^{z_i}}}$$

$$= e^{z_i}$$

Now,

$$\ln\left(\frac{P_i}{1-P_i}\right) = Z_i = \alpha + \beta X_i \quad \ldots\ldots\ldots (2)$$

Hence, $\frac{P_i}{1-P_i}$ is called odds ratio in favor of the event occuring and $\ln\left(\frac{P_i}{1-P_i}\right)$ is the log odds ratio (also called logit of P).



#Logistic Regression has many analogies to OLS regression:

Psuedo-R^2

Numerical
Y
↑
R^2
for categorical (high).
Psuedo R^2 = 0.1 0
(Relatively significant)

Assumptions

- Models the natural log of odds (logits) of success probability.
- Associated dependent variable : Dichotomous (0/1)
- Probability distribution : Binomial
- link : logit

Assumption

- Non autocorrelation
- linear relationship between predictors and log odds of dependent variable.
-

#Measuring goodness-of-fit

* Effron's R^2

$$\text{Effron's } R^2 = 1 - \frac{n}{n_1 n_2} \Sigma (Y - \hat{Y})^2$$

* Macfadden's Pseudo R^2

$$\text{Pseudo } R^2 = 1 - \frac{\ln L}{\ln L_0}$$

For individual Test

Wald's Test $= \left(\frac{b_i}{S b_i}\right)^2 = t^2$

Estimating overall ~~test~~ significance of regression

$$LR = 2 \ln \frac{L}{L_0} = 2(\ln L - \ln L_0)$$

$LR \sim \chi^2$ with k degrees of freedom (explanatory variables)

Odds: $\frac{P(S)}{P(F)} = \frac{P_i}{1 - P_i}$

odds Ratio: Ratio of two odds

Log odds: Also called logit odds $= \log\left(\frac{P_i}{1 - P_i}\right)$

e^{b} (b: slope) $\Rightarrow$ Odds ratio for an independent variable

Smoking data (Logistic Regression)

Smoker Y/N	age	income	Education	price of cigarettes
	***		***	(at 10%)
	— sign		— sign	— sign

Negative slope $\Rightarrow$ odds ratio < 1

Positive slope $\Rightarrow$ odds ratio > 1

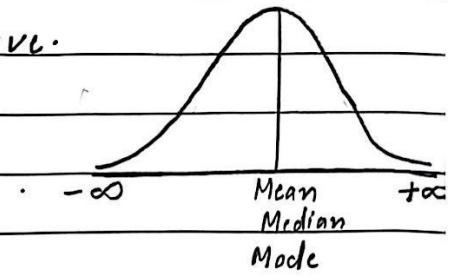
April 25, 2025 — Regression Analysis

Checking Normality

ⓔ → Normal Distribution

Normal distribution

It is a bell-shaped curve.



Properties

* Symmetric
* Mean = Median = Mode
* Asymptotic curve

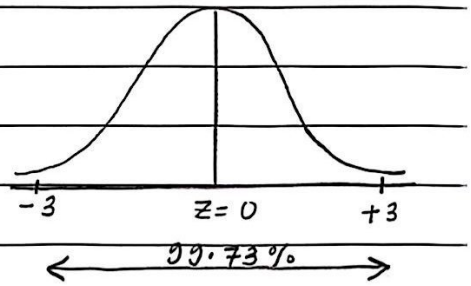
Probability function

Probability density function (pdf)

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

Standard normal variate

$$z = \frac{x-\mu}{\sigma}$$



Data

↓

Numerical

Σ Discrete → Integer (whole nos) → Counting processes (no. of colleges / books / students / successes / defectives / calls)

∫ Continuous → Decimal (fractional) → measuring processes (price, wt, salary, eps)

$\mu = 55$

$\sigma = 6$

$P(50 < X < 60) = ?$

$$\int_{50}^{60} f(x)\,dx$$

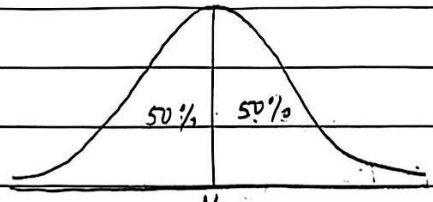
$$= \int_{50}^{60} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx$$

Detection

i) Mathematical method

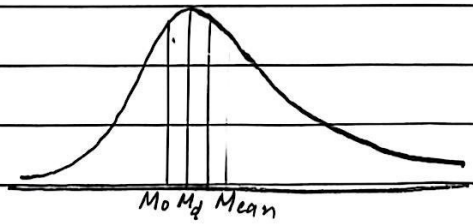
Mean ≠ Median ≠ Mode

Normal ⟹ Symmetric



Mean
Median
Mode

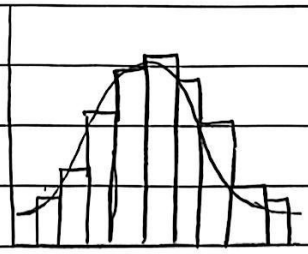
Skewed



ii) Graphical method

a. Histogram

b. Frequency curve
(Density curve)

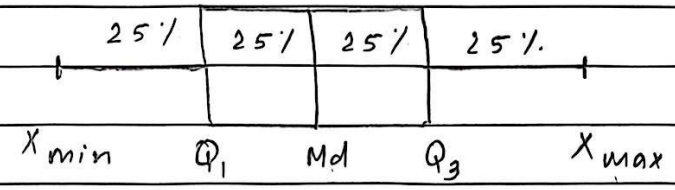


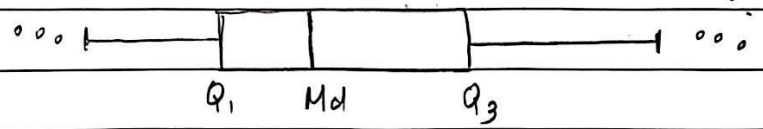
Histogram
Frequency curve

c. Box-and-Whisker plot

Five number summary

x_{min}	Q_1	Md	Q_3	x_{max}

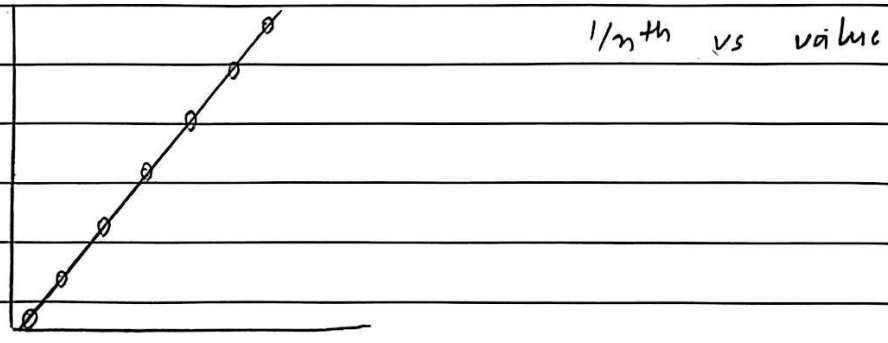


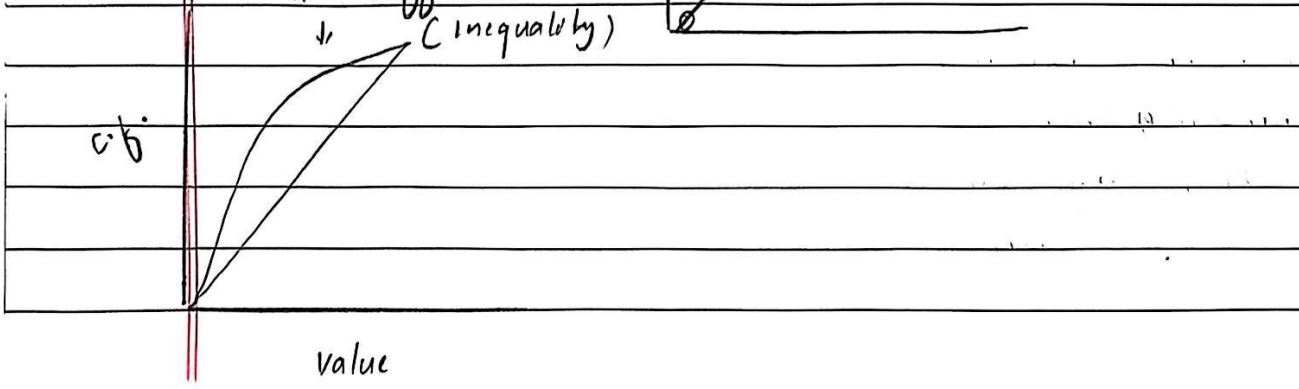


Lower limit $= Q_1 - 1.5(Q_3 - Q_1)$

Upper limit $= Q_3 + 1.5(Q_3 - Q_1)$

d. Q-Q Plot (Quantile-Quantile Plot)





Testing

→ Kolmogorov Smirnov Test } non-parametric

→ Shapiro-Wilks Test

Null Hypothesis (H_0): Data are normally distributed

Alternative Hypothesis (H_1): Data are not normally distributed.

p-value < α (0.05) ⇒ we reject H_0.

p-value > α (0.05) ⇒ we do not reject H_0.

Impact of violation of normality

Almost all parametric tests become invalid.

t-test
z-test
f-test
→ no longer valid

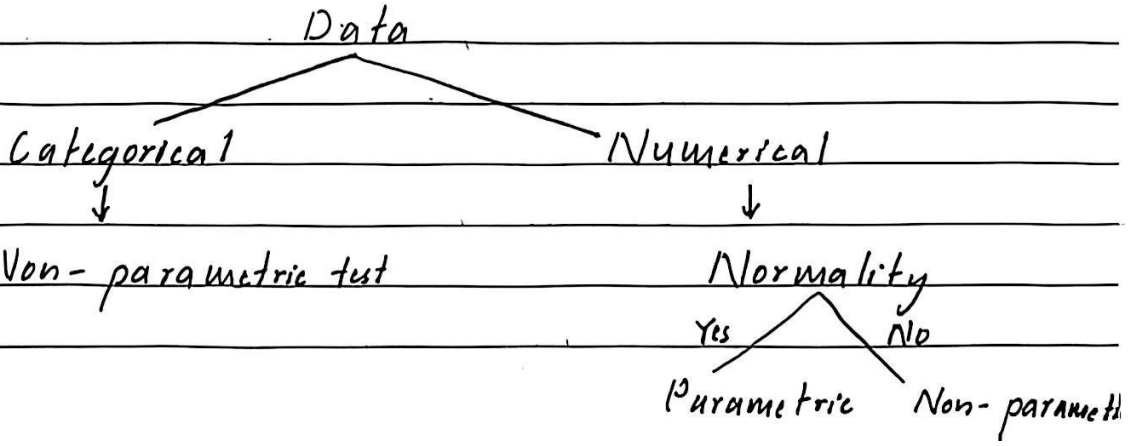
Regression:

errors → normally distributed

↓

violated

tests are not valid



Popular parametric test

Single mean → t- test (z- test)
Two means → t- test
Three or more means → F- test
Correlation test → t- test

Regression
Individual regression coefficient : t- test

$H_0: \beta_i = 0$
$H_1: \beta_i \neq 0$

Overall test : F- test

$H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0$
$H_1: \beta_i \neq 0$ (at least one)

$$F = \frac{MSR}{MSE}$$

Cal F
Tab F
Decision.

Popular non-parametric test

Chi- Square test for independence
(Significant association between two categorical variables)

Profession

Brand of car

Goodness-of-fit tests.

Kolmogorov-Smirnov Test (Normality)

Mann-Whitney Test

To compare two groups (alternative technique of two independent sample t-test)

$\mu_1 = \mu_2$

Kruskal-Wallis test.

To compare three or more groups (Alternative technique of F-Test or one way ANOVA)

$[\mu_1 = \mu_2 = \mu_3 = \mu_4]$

April 25, 2025

MARKOV CHAIN

April 26, 2025

Stepwise Regression

Model Selection Criteria

Variable Selection (inclusion of the variable in final model)

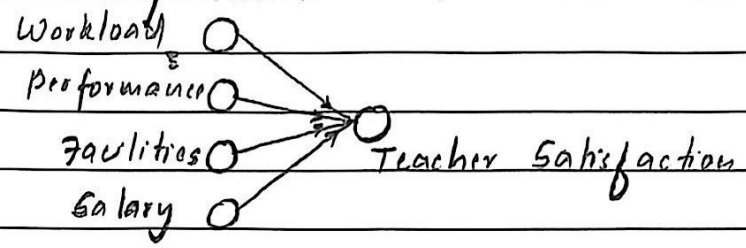
↓

Literature Review (Objectives / Hypothesis / Conceptual framework) → Topic, Objectives, Hypothesis

↓ Variable Selection

Model-1

Equation



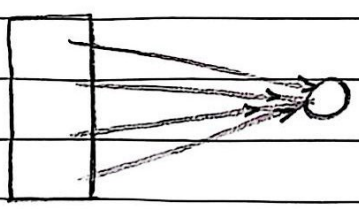
Model-2

⋮

Model-3

↓

Own Conceptual framework



Literature Review

↓

Variable Selection

→ Bivariate analysis (one ind and one dependent)
significant
insignificant

→ Multiple regression
logit
ordinal
multinomial
→ VIF → Omitted

Only significant in bivariate

Stepwise Regression

Y	X_1	X_2	X_3	X_4	~~X_5~~	X_6
Exp	Income	Family size	No. of child	no. of elder		Rural / Urban R-1 U-0

Forward selection

Y | One variable

↓

Result

R^2

Y | Two variables *

↓

Result

R^2

* Use significantly impacti variable

→ Standard Stepwise Regression selection
Combines attributes of both forward, and backward elimination

Mallow's Cp

$$C_p = RSS_p / \hat{\sigma}^2 - n + 2(P+1)$$

RSS, MSE, No. of predictor variables, Sample size

lowest C_p is the best

Models
Define logistic model.

Which test is used for individual predictors in LR?
Logistic Regression
Test for Individual Predictors

Individual Test

Exam

$H_0: \beta_i = 0$
$H_1: \beta_i \neq 0$

Test-statistic:
Wald Statistic ← $W = \left(\frac{b_i}{s\,b_i}\right)^2$

$$= \left(\frac{\hat{\beta}_i}{s\,\hat{\beta}_i}\right)^2$$

↓

p-value $< \alpha \Rightarrow$ reject H_0
$> \alpha \Rightarrow$ do not reject H_0

TPR

FPR

Summarizing Predictive Power: Classification Model.

$$P(\hat{Y} = 1 / Y = 1)$$

		Predicted Y	Predicted N
Actual	Y	O	
	N		O