Date :
Page :

Review of data and statistics

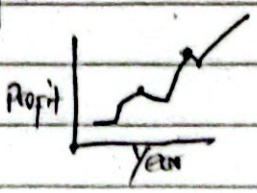
1. Classification of data
   → ① cross sectional / time series / panel data

① Cross sectional data
   → collecting the data in single frame of time
   → conducting a survey only once.

Respondent individuals	Variable 1 age	Variable 2 salary	. . . .	variable k gender

Purpose
   → identification of current status of statistical data.
   → comparision by analytical domain
      by age, gender, province etc
   → identify relationship betw^n variables
   → prediction / estimation
   → to make policy / strategy

② Time series
   → Collection of data over period of time
      Year/GDP , Year/remittance , Year/profit



Purpose
   → To observe trend
   → Forecasting the future
   → construction of cycle
   → study of seasonal impact

Living standard survey

Date :
Page :

(M) Panel data
→ combined of cross-sectional & time series

2. Categorical vs Numerical
Categorical
→ any question having alphabetical or grouping response
→ Freq / %
Numerical
→ any question having numeric response.
→ min/max/mean/sd.dev.

3. According to no of variables
Univariate
→ analysis of single variable
Bivariate
→ ... two ....
multivariate
→ ... three or more
→ dependance technique
↳ one dependent variable
↳ two or more independent variable.

$\bar{x}$ - sample
ω - population

Date: _____
Page: _____

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

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

Parameter:- Result/outcome of population data
Characteristic of population data

Population size → N
" mean → ω
" proportion → π (or P)
" variance → σ^2
" correlation → ρ
" regression coefficient → β_i

Sample:
Subset of the population.

Sampling :- collecting the information/data from sample.

Estimator/~~statistic~~ Statistic :- Result/outcome of sample data
Characteristic of sample data.

Sample size → n ($\hat{N}$)
" mean → $\bar{x}$ ($\hat{\omega}$)
" proportion → p ($\hat{P}$)

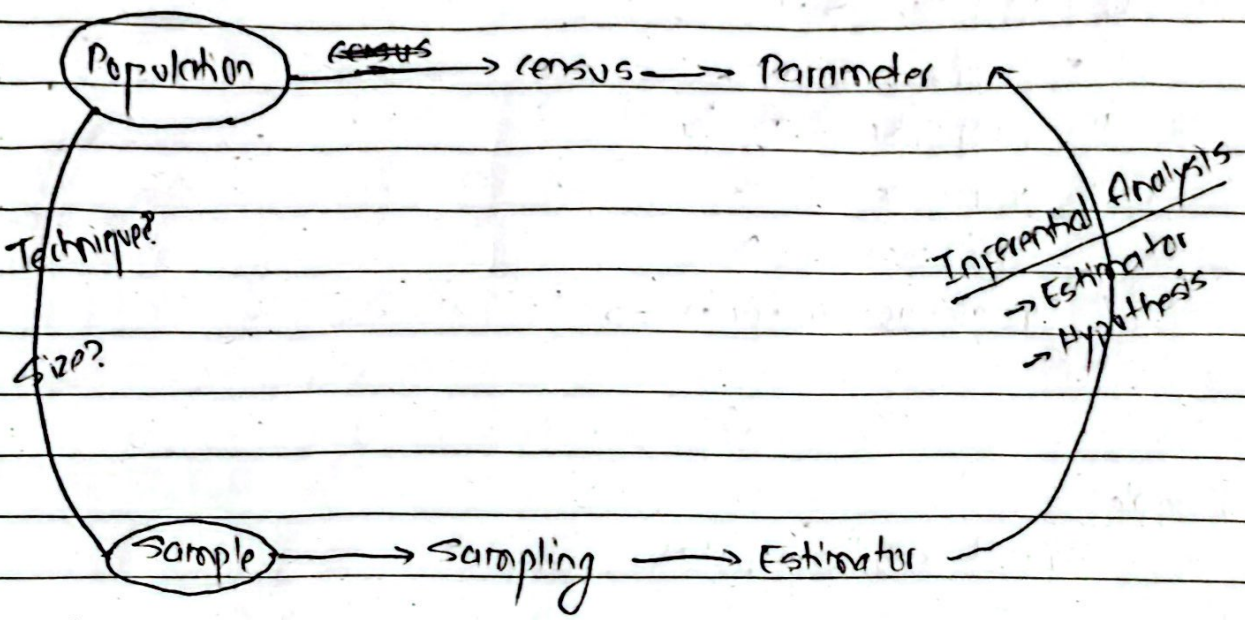
Econometrics
Wooldridge

Date:
Page

sample variance → s^2 ($\hat{\sigma}$)
" correlation → r ($\hat{\rho}$)
" regression coefficient → b_i ($\hat{\beta}_i$)

Sampling error → |Estimator - Parameter|
→ used to determine sample size

* Sample design



Quantitative → Quantitative research Date: _____
Page: _____

Research

Quantitative research

↓

Survey

↓

Questionaire (Instrument)

(Instrument should be valid
ie. should measure the what
research wants to measures
→ objectives of the research)

→ should be correct & objective
reliable (validity)

↓

answer/response → Correct
(data) (reliability)

↓

Data Management/Processing

↓

Result → Data analysis

↓

Knowledge

↓

Policy/Plan/strategy/Program.

Null and alternate are
→ mutually exclusive
→ exclusive (belongs to one)

Date:
Page:

Statistics

Statistics

Descriptive	Inferential
→ Describing the statuse Table / Chart Central values / freq / ... cross tabulation	→ Technique of giving conclusion about the population parameter on the base of sample data / estimator / statistic → Estimation → Hypothesis

* Hypothesis:

→ Pre-assumption / statement / guess about the population parameter
→ (outcome / result of the research)
With the help of:
→ Past ~~es~~ literature
→ Qualitative research / pilot

Null hypothesis (H_0): Statement of no difference or no relationship
'=', '≥', '≤' (equality)

Alternative hypothesis (H_1): opposite / compositive of null hypothesis.
statement of relationship / difference
≠ , >, <
↓ two , ↓ one

As $n \to N$
estimator → parameter

Critical value → point approach/value of the point
p-value → area approach

Date: _____
Page: _____

* Level of significance (α):
  Probability of rejecting the true statement
  1% | 5% | 10%

* Confidence level ($1-\alpha$)
  Probability of accepting the true statement
  99% | 95% | 90%

* Decision Criteria:
  → Critical value approach
  → If |test statistics| ≤ critical value (tab value)
    do not reject H_0

  → P-value approach
    If p-value < α
      reject H_0
    else
      do not reject H_0

Example:

Q1 Gender [2]
  Female-1
  Male -2

Q2. Post [1]
  Assistant - 1
  Officer - 2
  Manager - 3

Q3. Income 34,000

Q4 Expenditure 15000

Q1	Q2	Q3	Q4
2	1	34000	15000

Date:
Page

θ_1 → nominal
θ_2 → ordinal
θ_3 → ratio/numerical data

* Univariate
Case a: Categorical data (nominal/ordinal)
→ Frequency/% analysis
→ Graph
- Bar
- pie

Distribution by gender

Gender	Freq	%
Male		
female		
Total		

Case b: Numerical data
→ min/max/mean/std dev
→ one sample t-test
→ Graph
- → Histogram
- → Box-plot

* Bivariate analysis
Two variables at a time
Case a: If both are categorical/grouping
→ cross-tabulation betw two variables with freq and appropriate %.
→ Test
- → Chi square test

Date:
Page:

Gender \ Post	Assitare	office	Manager	
Female	70 (0.70)	20 (0.1)	10 (0.1)	100
Male	100 (0.5)	50 (0.2)	50 (0.2)	200
	170	70	60	300

Case b: If one is categorical and one is numerical
- → comparision of numerical by categorical
- → independent sample t-test (Two samples)
- → One way ANOVA (f-test) (More than two samples)

Gender	Average	Min	Max	st.dev
Female				
Male				
Total				

Comparision betw two means

Comparision of income by post

Post \ Salary	Average	Min	Max	st.dev
Assitare				
Office				
Manager				

Case c: If both are numerical
- → correlation (relationship betw two variables)
- → regression (we need dependent and independent variable)

Regression analysis

Identification of dependent and independent variable

no. of dependent variable → one

no of independent variable → one
→ two or more (multiple) ($x_1, x_2 \ldots x_n$)

Example :-

Price x	Sales volume y	Sales volume	Price	no. of ads	no. of outlets

→ Developing an equation

$Y = (\beta_0 + \beta_1 x + \epsilon)$ (Popⁿ)

$y = b_0 + b_1 x + e$ (sample)

$\hat{y} = b_0 + bx$

$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon$, (Popⁿ)

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

$\hat{y} = b_0 + b_1 x_1 + \ldots + b_p x_p$

Y = real value / actual / collected

$\hat{y}$ = estimated value / predicted

$y - \hat{y}$ = error / residuals

Date:
Page:

Objectives:

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

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 dependent variable.

$$\hat{y} = b_0 + b_1 x$$

$$b_1 = \frac{\nabla y}{\nabla x} = \frac{\text{amount change in } y}{\text{pa. unit change in } x}$$

$$\ln \hat{y} = b_0 + b_1 \ln x$$

$$b_1 = \frac{\nabla y/y}{\nabla x/x} = \frac{\text{\% change in } y}{\text{\% change in } x} \quad \text{[Elasticity]}$$

$$\ln \hat{y} = b_0 + b_1 x$$

$$b_1 = \frac{\nabla y/y}{\Delta x} = \frac{\text{\% change in } y}{\text{unit change in } x}$$

④ Computation of growth rate/amount per time period.

⑤ Proportion of ~~explaind variable~~ variation in Y (dependent variable) that is explained by independent variable.
[Coeff. of determination $- R^2$]

Price (x)	sales volume
	$b_1 = -0.75$
	$R^2 = 0.82$

82% of variations in sales is explaind by Price.

Date:
Page:

(vi) Policy/program/plan/decision

Simple regression analysis

dependent varinbe → one (y)
independent " → one (x)

Equations:

$$Y = \beta_0 + \beta_1 X + \epsilon \quad (popn)$$

$$y = b_0 + b_1 x + e \quad (simple)$$

By using least sq. method

$$b_1 = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{n\Sigma x^2 - (\Sigma x)^2} = \frac{cov(x,y)}{var(x)}$$

$$b_0 = \frac{\Sigma y}{n} - b_1 \frac{\Sigma x}{n}$$

$$\boxed{\hat{y} = b_0 + b_1 x}$$

Exaple:

x (income) in 1000	y (expenditure) in 1000	xy	x^2	y^2
15	10	150	225	100
18	12	216	324	144
20	12	240	400	144
20	17	340	400	289
25	17	425	625	289
29	18	522	841	324
127	86	1893	2815	1290

$\Sigma xy = 1893 \quad \Sigma x^2 = 2815 \quad \Sigma y^2 = 1290$

$\Sigma x = 127 \quad \Sigma y = 86$

Date: _____
Page: _____

$$b_1 = \frac{n\Sigma xy - \Sigma(x)\Sigma(y)}{n\Sigma x^2 - (\Sigma x)^2} \qquad e = y - \hat{y}$$

= observed value − estimated value

$$= \frac{6 \times 1893 - 127 \times 86}{6 \times 2815 - 127^2}$$

$$= \frac{436}{761} = 0.573$$

Standard error of estimate

$$SE(\hat{y}) = \sqrt{\frac{\Sigma e^2}{n-2}}$$

$$b_0 = \frac{\Sigma y}{n} - b_1 \frac{\Sigma x}{n}$$

$$= \frac{86}{6} - \left(\frac{436}{761}\right) \times \frac{127}{6}$$

$$= 2.206$$

The estimation equation

$$\hat{y} = 2.10 + 0.572$$

$$\hat{y} = 2.206 + 0.572x$$

$\hat{y} = 2.206 + 0.572x$	$(y - \hat{y})$ (error)	$(y - \hat{y})^2$
16.786	−0.786	0.617
12.502	−0.502	0.252
13.146	−1.691	2.709
13.146	3.354	11.249
16.506	0.494	0.244
18.794	−0.794	0.630
	0.12 ≈ 0	15.701

→ error terms should be randomly distributed

→ if there is pattern in sign of error it is auto-correlation

Date:
Page:

Standard error of estimate

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

$$= \sqrt{\frac{\Sigma(y-\hat{y})^2}{n-p-1}}$$

$p \rightarrow$ no. of independent variable

$$= \sqrt{MSE}$$

It measures the average variation/deviation of observed value of dependent variable (x) around its fitted equation ($\hat{y}$).

"cairl

"tidyverse
install

Date:
Page:

Multiple regression analysis

Example:

Sales volume	Price	no of shoes	no of orders
y	x_1	x_2	x_3

$P = 3$
$k = 3$
} independent variable

Note: If the graph of error follows certain pattern there is auto-correlation which means there is problem in data.

* In matrix form

y_i	x_{1i}	x_{2i}	...	x_{ki}
y_1	x_{11}	x_{21}	...	x_{k1}
y_2	x_{21}	x_{22}	...	x_{k2}
$\vdots$				

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki} + \epsilon_i$$

ie. $$y_1 = \beta_0 + \beta_1 x_{11} + \beta_2 x_{21} + \ldots + \beta_k x_{ki} + \epsilon_1$$

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} x_{21} \ldots x_{k1} \\ 1 & x_{22} x_{22} \ldots x_{k2} \\ \vdots & \vdots \\ 1 & x_{1n} x_{2n} \ldots x_{kn} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_k \end{bmatrix}$$

variance(x): $\frac{\Sigma (x-w)^2}{2}$

$= E(x-w)^2$

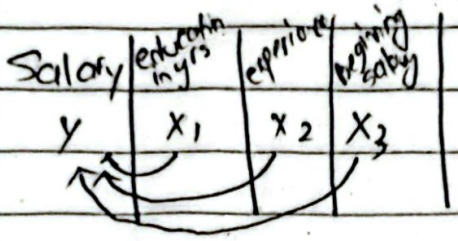
$= E[(x-w)(x-w)^t]$

Date: _____
Page: _____

ANOVA table:

Software → F-value & p-value
coefficient table
summary → coefficient table
↳ t-value & p-value

F-test → it is used for overall test
t-test → it is used for individual test



impact of x_1 on y, x_2 on y, x_3 on y → t-test (individual-b
$\beta_1 = 0$ $\beta_2 = 0$ $\beta_3 = 0$

impact of x_1, x_2 and x_3 on y → overall test
$\beta_1 = \beta_2 = \beta_3 = 0$

F-test ⇒ H_0 : $\beta_1 = \beta_2 = \beta_3 = 0$
there is no significant impact of all independent variables on dependent variable.

H_1 : At least $\beta_i \neq 0$
There is significant impact of at least one independent variable on dependent variable.

if p-value < α (0.05) → reject H_0
p-value ≥ α (0.05) → do not reject H_0

Date:
Page:

Format of ANOVA table

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

Format of coeff. table

Predictor	Regression coeff.	St. error	t-value ($t = \frac{b_i}{Sb_i}$)	P-value
constant	b_0	Sb_0		
x_1	b_1	Sb_1		
x_2	b_2	Sb_2		
⋮	⋮	⋮		
x_k	b_k	Sb_k		

From ANOVA table and coeff. table

- Developing and equation & prediction

$$\hat{y} = b_0 + b_1x_1 + b_2x_2 + \ldots + b_kx_k$$

- Computing multiple coeff. of determination

$$R^2 = \frac{\text{Explained variation}}{\text{Total variation}} = \frac{SSR}{TSS}$$

TSS = Total sum of squares (Total variation)

SSR = Sum of squares due to regression

SSE = Sum of squares due to error / residuals

$$TSS = SSR + SSE$$

$$R^2 = \frac{SSR}{TSS} = 1 - \frac{SSE}{TSS}$$

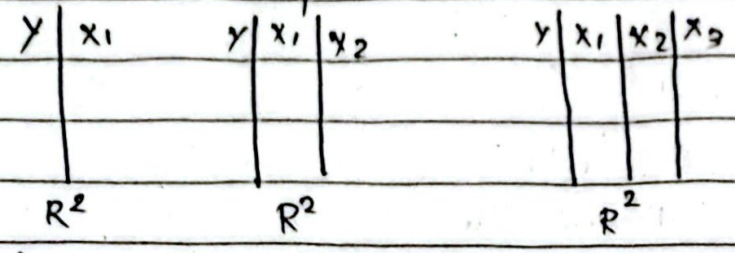
$$0 \le R^2 \le 1$$

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

81% of variations in dependent variable is explained by 'k' independent variables.

iii) Adjusted R^2

→ It is used for model selection.



Increment in independent variable increases R^2.

So we used adjusted R^2.

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

iv) Standard error of estimate

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

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

v) Overall test/godness of fit

Hypothesis testing for all regression coefficient simultaneously.

Date:
Page

Null hypothesis (H_0): $\beta_1 = \beta_2 = \ldots = \beta_k = 0$
There is no significant impact of all independent variable on dependent variable.

Alternative hypothesis (H_1): At least one $\beta_i \neq 0$
There is significant impact of at least one independent variable on dependent variable.

Test statistics

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

Cal $F = F$
Tab $F = F_{k, n-k-1}$ } at α%.

$k \rightarrow$ df. for numerator
$n-k-1 \rightarrow$ " " denominator

If cal $F \leq$ Tab F
we do not reject H_0
If cal $F >$ Tab F
We reject H_0

P-value approach
If p-value $< \alpha$
we reject H_0
If p-value $\geq \alpha$
we dont reject H_0

Date :
Page

Individual test (t-test) [coeff table]

Null hypothesis (H_0): $\beta_i = 0$
There is no significant impact of x_i on y

Alternative hypothesis (H_1): $\beta_i \neq 0$
There is significant impact of x_i on y

Test statistics

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

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
P-value $\geq \alpha \Rightarrow$ do not reject

Confidence interval estimate

$b_i \pm t_{n-k-1}, \alpha \; Sb_i$

- b_i ↓ coeff table
- t_{n-k-1}, α ↓ table ↓ statistical
- Sb_i ↓ coeff table.

Date:
Page:

Example:

By using following ANOVA table obtained from 30 obs.

Source	df	SS	MSS	F-value
Regression	4	300	?	?
Error	?	?	?	
Total	?	500		

a) Compute the given ANOVA table

b) Compute standard error of estimate & intrepret its meaning

c) Compute multiple coeff. of determination and intrepret its meaning

d) Compute Adj R^2

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

a) $n=30$

Source	df	SS	MSS	F-value
Regression	4	300 (SSR)	$300/4 = 75$	$75/8 = 9.37$
Error	$25(n-k-1)$	$200(TSS-SSR)$	$200/25 = 8$	
Total	$29(n-1)$	500 (TSS)		

b) Standard error of estimate$(S_{yx}) = \sqrt{MSE}$

$= \sqrt{8}$

$= 2.828$

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

c) Multiple coeff of determination

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

$$= \frac{300}{500} = 3/5 = 0.60$$

60% of variation in dependent variable is explained by 4 independent variables.

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

$$= 1 - \left\{ (1 - (3/5)^2 \frac{29}{25} \right\}$$

$$= 0.536$$

e) Null hypothesis (H_0): $\beta_1 = \ldots = \beta_4$

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

Alternative hypothesis (H_1): $\beta_i \neq 0 \quad i = 1 \ldots 4$

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

Test - statistics

$$F = \frac{MSR}{MSE} = 9.37$$

Cal $F = 9.37$

Tab $F = 2.759$

Cal F > tab F

So, we reject H_0