Econometrics – 18KP3EC11
UNIT – III, IV, V

Dr. S. Manonmani
Assistant professor
Department of Economics
KNGAC for Women
Thanjavur.

UNIT - III

- Simple Linear Regression Model and Multiple Linear Regression Model
- Meaning – Specification of model-assumptions of SLRM – Stochastic and Non-stochastic- OLS – Specification, Estimation, Evaluation and application – Gauss – Markov theorem – problems – Multiple linear Regression model - Meaning

- **Simple Linear Regression Model**
- **Introduction**

Economic Theory is concerned with the relations between variables, eg. Cost function, demand function, production function.

The entire body of economic theory can be regarded as a collection of relations among variables.

- **Relations between variables**

A systematic study of economics is possible only because the different parts of an economy are inter-related.

These relations either deterministic or stochastic.

A relation between X and Y which are characterised as Y = f(X) will be deterministic, if for each value of X there is only one corresponding value of Y.

A relation between X and Y is said to be stochastic , if each value of X there is a whole probability of distribution of values of Y. i.e. for any given value of X , the Y variable may assume some specific value or fall within some specific interval.

We will see how to establish the relationship. For eg., if we have a statement that consumption depends on income. This can be expressed as,

$$C = f(Y)$$

It shows that a change in Y causes a change in C and a change in C due to a change in Y or there is a cause and effect relation ship between the two (Y is cause and C is effect). This form of relationship is called an implicit functional form.

the functional form does not serve our purpose fully. So it is to be specified. The specific form will describe the intercept, slope and curveture etc.

- **Linear relationship between two variables**

The linear relationship will be,

$$Y = \alpha + \beta X$$

Econometrics concerned with testing the above form of relation and with estimating parameters α and β.

- **Simple Linear Regression model**
- **Meaning and specification of the model**

A simple linear regression model (SLRM) includes only two variables with linear relationship between them. It is the form,

$$Y = \alpha + \beta X + u$$

Where,

Y is endogenous or dependent variable

X is exogenous or explanatory variable

u is error or random or stochastic variable

α is constant regression parameter or intercept on Y axis (Y when X = 0)

β is slope regression parameter, which measures the rate of change or marginal value $\Delta Y / \Delta X$. It is the effect of a unit change in X on Y.

In this model $(\alpha + \beta X)$ is called systematic component.

The error or stochastic term 'u' includes,

1. Influence of variables omitted in the model.
2. Errors due to aggregation of macro variables.
3. Errors due to measurement of variables
4. Errors due to misspecification of the model.

If $Y = \alpha + \beta X + u$ represents demand function then Y is quantity demanded and X is the price level of the commodity. According to Marshall's law of demand for normal goods, α (maximum demand) is positive and β is negative (due to inverse relation between price and demand).

Estimation of SLRM

In order to estimate the model Ordinary Least Squares (OLS) method is used due to the following reasons.

OLS estimators are,

i. BLUE
ii. Simple to understand and easy to calculate
iii. Provides more reliable results in a wide range of problems.

For a sample of i = 1, 2, 3, …. n if i refers to the item.

True relation is $Y_i = \alpha + \beta X_i + u_i$ and

True regression is $E(Y_i) = \alpha + \beta X_i$

Estimated relation is $Y_i = \hat{\alpha} + \hat{\beta} X_i + e_i$

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

The method of least squares is based on the stochasting and non-stochasting assumptions

Assumptions of SLRM

Stochastic Assumptions

These are the assumptions about the random term u. They are;

i. u_i is a real random variable for all i = 1,2,3…n

ii. Mean value of u_i is zero even though in individual cases it may have positive or negative values. $E(u_i) = 0$ for all i = 1,2,3….n

iii. Variance of u_i is constant for every u_i.

$Var(u_i) = E(u_i^2)\ \sigma_i^2$, for all i = 1,2,3….n

This assumption states that the error terms are homoscedastic.

iv. The term u_i has a normal distribution.

ui is $N(0, \sigma_i^2)$ for all i = 1,2,3,…n

v. The random terms u_i and u_j are not correlated which means that there exists non-auto correlated error terms

$E(u_i\ u_j) = 0$ for all $i \neq j$

vi. ui is independent of the of the explanatory variables.

$E(uX) = 0$ or X's are purely exogenous.

Non – stochastic Assumptions

These are the assumptions about the systematic component ($\alpha + \beta X$) of the model

i. The relationship being estimated is just identified.
ii. The relationship is correctly specified and there is no misspecification or specification error.
iii. The explanatory variables are measured without errors, which means that there is no measurement error.
iv. The macro variables are correctly aggregated.
v. The explanatory variables are not perfectly linearly correlated, which implies that there is no multi – collinearity problem.

OLS (Ordinary least squares)

According to this method of estimation the sum of squares of deviations of the actual values should be minimised.

$\Sigma e_i^2 = \Sigma (y - \hat{Y}_i)^2 = \Sigma [Y_i – (\hat{\alpha} + \beta X_i)]^2$ is minimised

Differentiating with respect to $\hat{\alpha}$ and $\hat{\beta}$ respectively

$\frac{\partial}{\partial\hat{\alpha}}(\Sigma e_i^2) = -2\,\Sigma(Y_i - \hat{\alpha} - \hat{\beta}X_i) = 0$ and

$\frac{\partial}{\partial\hat{\beta}}(\Sigma e_i^2) = -2\,\Sigma X_i(Y_i - \hat{\alpha} - \hat{\beta}X_i) = 0$

On rearranging these two, we get the following two equations

$\Sigma y_i = n\hat{\alpha} + \hat{\beta}\,\Sigma X_i$ --------- 1

$\Sigma X_i Y_i = \hat{\alpha}\Sigma X_i + \hat{\beta}\,\Sigma X_i^2$ ------- 2

Solving these two equations we get

$\hat{\alpha} = \overline{Y} - \hat{\beta}\,\overline{X}$ and

$\hat{\beta} = \frac{\Sigma x_i y_i}{\Sigma x_i 2}$

Where $x_i = X_i - \overline{X}$, $y_i = Y_i - \overline{Y}$ and $x_i^2 = \Sigma x_i - n\overline{X}^2$

Properties of OLS estimators (Gauss – Markov Theorem)

In a SLRM, $Y_i = \alpha + \beta X_i + u_i$ the OLS estimators are optimal (BLUE)

1. Linearity: The OLS estimator is a linear function of the sample values.
2. Unbiasedness: Expected value of the estimator is equal to is actual value.
3. Minimum Variance: OLS estimators have minimum variance

Thus $\hat{\alpha}$ and $\hat{\beta}$ are BLUE or OLS estimators are Best, Linear, Unbiased Estimators.

Evaluation of SLRM

Economic criteria: The signs and magnitude of the estimated parameters are to be verified for economic theory.

Statistical criteria: Using standard error values, the statistical tests of significance are to be conducted to evaluate the precision of estimates. If necessary the the confidence intervals are to be worked out.

When the sample is large ($n \geq 30$),

$$Z = \frac{|\hat{\beta} - \beta_0|}{SE(\hat{\beta})} \quad (H_0: \beta = \beta_0) \quad \text{and}$$

$$Z = \frac{|\hat{\beta}|}{SE(\hat{\beta})} \quad (H_0: \beta = 0)$$

When the sample is small ($n<30$)

$$t_{n-2} = \frac{|\hat{\beta} - \beta_0|}{SE(\hat{\beta})} \quad (H_0: \beta = \beta_0) \quad \text{and}$$

$$t_{n-2} = \frac{|\hat{\beta}|}{SE(\beta)} \quad (H_0: \beta = 0)$$

with degrees of freedom ($n - 2$). Comparing the calculated values, if they are greater than the respective table values the H_0 is rejected at fixed level of significance. So the estimated parameters are concluded to be significant.

Economic criteria: Using R^2 (Coefficient of determination) value, its significance is tested to assess the forecasting ability of the model.

In case of small and large samples

$$F_{(p-1,\ n-p)} = \frac{\frac{R^2}{p-1}}{\frac{1-R^2}{n-p}} \quad (H_0: R^2 = 0)$$

$$F_{(1,\ n-2)} = \frac{R^2}{1-R^2}\sqrt{n-2} \quad (H_0: R^2 = 0)$$

When n = sample size, p = number of estimated parameters with d.f. = (1, n-2)

If $F > F_{5\%}$ then H_1 is accepted at 5% level of significance and vice versa.

For confidence limits for β is

$\hat{\beta} \pm$ Table value of test statistic $(S.E(\hat{\beta}))$

Application of SLRM

In the conventional format the reports of regression analysis are presented with the estimated equation, the standard error values of parameters and the value of R^2.

For example,

In demand analysis, the estimated linear demand function is,

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

$\hat{Y} = 45.86^{**} - 3.45^{**}X$ with R2 = 0.934**

With S.E $(\hat{\alpha}) = 8.25$ and S.E $(\hat{\beta}) = 1.25$, n= 25

** denotes significant at 1% level of significance.

R^2 value indicates the goodness of fit and forecasting ability while the significance of the parameters indicates the precision of estimators. It shall be derived from the fitted linear demand function that price effect is negative. It shows inverse relation between price and demand, for one unit increase in price, demand is found to fall by 3.45 units on an average. When the commodity becomes free, the maximum demand for it becomes 45.86. For this commodity 93.4% of change in demand is due to the linear influence of its price.

- 1. Estimate the linear cost function and evaluate it.

Year	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015
Total Cost (000$) (Y)	150	140	160	170	150	162	185	165	190	185
Quantity Produced (000 units)	40	42	48	55	65	79	88	100	120	140

Specification: Linear cost function is

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

Yi = Total Cost (000$)

Xi = Total quantity produced (000 units)

ui = Error term

α = Total fixed cost (Constant regression Parameter)

β = Marginal cost (slope regression parameter)

Year	X	Y	x (X - $\bar{X}$)	Y (Y - $\bar{Y}$)	xy	x^2	y^2
2006	40	150	- 37.7	-15.7	591.89	1421.29	246.49
2007	42	140	-35.7	-25.7	917.49	1274.49	660.49
2008	48	160	-29.7	-5.7	169.29	882.09	32.49
2009	55	170	-22.7	-4.3	-97.61	515.29	18.49
2010	65	150	-12.7	-15.7	199.39	161.29	246.49
2011	79	162	1.3	-3.7	-4.81	1.69	13.69
2012	88	185	10.3	19.3	198.79	106.09	372.49
2013	100	165	22.3	-0.7	-15.61	497.29	0.49
2014	120	190	42.3	24.3	1027.89	1789.29	590.49
2015	140	185	62.3	19.3	1202.39	3881.29	372.49
	777	1657			4189.1	10530.1	2554.07

$$\bar{X} = \frac{\Sigma x}{n} = \frac{777}{10} = 77.7$$

$$\bar{Y} = \frac{\Sigma y}{n} = \frac{1657}{10} = 165.7$$

Estimation: Using OLS method of Estimation

$$\hat{\beta} = \frac{\Sigma xy}{\Sigma x^2} = \frac{4189.1}{10530.1} = 0.398$$

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X} = 165.7 - 0.398(77.7) = 134.775$$

Estimated cost function is $\hat{Y}$ = 134.775 + 0.398 X

Evaluation : Using Gauss – Markov theorem

$$r^2 = \frac{\hat{\beta}(\Sigma xy)}{\Sigma y^2} = \frac{(0.398)(4189.1)}{2554.07} = 0.653$$

$$\Sigma e^2 = (1 - r^2)(\Sigma y^2) = 1 - 0.653)(2554.07) = 886.26$$

$$SE(\hat{\beta}) = \sqrt{(\frac{\Sigma e^2}{n-2})(\frac{1}{\Sigma x^2})} = \sqrt{(\frac{886.26}{8})(\frac{1}{10530.1})} = 0.103$$

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

$\Sigma X^2 = \Sigma x^2 + n\bar{X}^2$ = 10530.1+10(77.7)^2 = 70903

$$= \sqrt{(\frac{886.26}{8})(\frac{70903}{10(10530.1})}$$

$$= \sqrt{74.59} = 8.6368$$

Tests of Significance

For $\hat{\beta}$

$$t_{n-2} = \frac{\hat{\beta}}{SE(\hat{\beta})} \quad (H_0: \beta = 0)$$

$$t_8 = \frac{0.398}{0.103} = 3.864$$

For $\hat{\alpha}$

$$t_{n-2} = \frac{\hat{\alpha}}{SE(\hat{\alpha})} \quad (H_0: \alpha = 0)$$

$$t_8 = \frac{134.775}{8.6368} = 15.60$$

For degrees of freedom 8 (one sided test) $t_{5\%} = 1.860$ and $t_{1\%} = 2.896$

Here, for $\hat{\beta}$ calculated 't' value is greater than table value at 1% level of significance

Therefore, $\hat{\beta}$ is significant at 1% level.

For $\hat{\alpha}$ calculated 't' value is greater than table value at 1% and 5% levels of significance.

Therefore $\hat{\alpha}$ is significant at 1% level.

For R^2

$$F_{(p-1,\, n-p)} = \frac{\frac{R^2}{p-1}}{\frac{1-R^2}{n-p}} \qquad F_{(1,\, 8)} = \frac{\frac{0.653}{1}}{\frac{0.347}{8}} = \frac{0.653}{0.0433} = 15.055$$

For degrees of freedom (1, 8) $F_{5\%} = 5.32$ and $F_{1\%} = 11.26$

calculated F value is greater than the table value at 1% level

Therefore, R2 is significant at 1% level.

The fitted cost function is

$\bar{Y}$ = 134.775** + 0.398** X   with R2 = 0.653**

SE($\hat{\alpha}$) = 8.638   SE ($\hat{\beta}$) = 0.103   n = 10

Note: ** indicates significant at 1% level

Application: The constant regression parameter is significant at 1 per cent level. It's positive value implies that total fixed cost (Y) is 134.775 when the quantity produced is zero and hence it is the short run cost function.

The slope regression parameter is significant at 1% level. Its positive value indicates that for every 1000 units increase in production (X) there is an increase of $398 in total cost. So, marginal cost is constant.

R2 is significant at 1% level. Its value implies that 65.3% of change in cost is due to the linear influence of production and 34.7% of the change due to the influence of other variable.

SLRM $Y = \alpha + \beta X + u$

1. $\bar{X} = \frac{\Sigma X}{n}$
2. $\bar{Y} = \frac{\Sigma Y}{n}$
3. $\hat{\beta} = \frac{\Sigma xy}{\Sigma x^2}$
4. $\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}$
5. $r^2 = \frac{\hat{\beta}(\Sigma xy)}{\Sigma y^2}$
6. $\Sigma e^2 = (1 - r^2)(\Sigma y^2)$
7. $SE(\hat{\beta}) = \sqrt{\left(\frac{\Sigma e^2}{n-2}\right)\left(\frac{1}{\Sigma x^2}\right)}$
8. $SE(\hat{\alpha}) = \sqrt{\left(\frac{\Sigma e^2}{n-2}\right)\left(\frac{\Sigma X^2}{n\Sigma x^2}\right)}$
9. $\Sigma X^2 = \Sigma x^2 + n\bar{X}^2$

Tests of Significance

For $\hat{\beta}$

$$t_{n-2} = \frac{\hat{\beta}}{SE(\hat{\beta})} \quad (H_0: \beta = 0)$$

For $\hat{\alpha}$

$$t_{n-2} = \frac{\hat{\alpha}}{SE(\hat{\alpha})} \quad (H_0: \alpha = 0)$$

For R^2

$$F_{(p-1,\, n-p)} = \frac{\frac{R^2}{p-1}}{\frac{1-R^2}{n-p}}$$

Where p is number of parameters

CRITICAL VALUES OF STUDENT'S t – DISTRIBUTION

d.f	ρ-Level of significance for Two-sided test					d.f
	0.20	0.10	0.05	0.02	0.01	
	ρ-Level of significance for One-sided test					
	0.10	0.05	0.025	0.01	0.005	
1	3.078	6.314	12.706	31.821	63.657	1
2	1.886	2.920	4.303	6.965	9.925	2
3	1.638	2.353	3.182	4.541	5.841	3
4	1.533	2.132	2.776	3.747	4.604	4
5	1.476	2.015	2.571	3.365	4.032	5
6	1.440	1.943	2.447	3.143	3.707	6
7	1.415	1.895	2.365	2.998	3.499	7
8	1.397	1.860	2.306	2.896	3.355	8
9	1.383	1.833	2.262	2.821	3.250	9
10	1.372	1.812	2.228	2.764	3.169	10
11	1.363	1.796	2.201	2.718	3.106	11
12	1.356	1.782	2.179	2.681	3.055	12
13	1.350	1.771	2.160	2.650	3.012	13
14	1.345	1.761	2.145	2.624	2.977	14
15	1.341	1.753	2.731	2.602	2.947	15
16	1.337	1.746	2.120	2.583	2.921	16
17	1.333	1.740	2.110	2.567	2.898	17
18	1.330	1.734	2.101	2.552	2.878	18
19	1.328	1.729	2.093	2.539	2.861	19
20	1.325	1.725	2.086	2.528	2.845	20
21	1.323	1.721	2.080	2.518	2.831	21
22	1.321	1.717	2.074	2.508	2.819	22
23	1.319	1.714	2.069	2.500	2.807	23
24	1.318	1.711	2.064	2.492	2.797	24
25	1.316	1.708	2.060	2.485	2.787	25
26	1.315	1.706	2.056	2.479	2.779	26
27	1.314	1.703	2.052	2.473	2.771	27
28	1.313	1.701	2.048	2.467	2.763	28
29	1.311	1.699	2.045	2.462	2.756	29
Infinity	1.282	1.645	1.960	2.326	2.576	Infinity

5 % CRITICAL VALUES OF F – DISTRIBUTION

v1 / v2	1	2	3	4	5	6	8	12	24	∞
1	161.4	199.5	215.7	224.6	230.2	234.0	238.9	243.9	249.1	243.3
2	18.51	19.00	19.16	19.25	19.30	19.33	19.37	19.41	19.45	19.50
3	10.13	9.55	9.28	9.12	9.01	8.94	8.85	8.74	8.64	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.04	5.91	5.77	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.82	4.68	4.53	4.36
6	5.99	5.14	4.76	4.53	4.39	4.28	4.15	4.00	3.84	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.73	3.57	3.41	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.44	3.28	3.12	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.23	3.07	2.90	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.07	2.91	2.74	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	2.95	2.79	2.61	2.40
12	4.75	3.88	3.49	3.26	3.11	3.00	2.85	2.69	2.51	2.30
13	4.67	3.80	3.41	3.18	3.02	2.92	2.77	2.60	2.42	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.70	2.53	2.35	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.64	2.48	2.29	2.07
16	4.49	3.63	3.24	3.01	2.85	2.74	2.59	2.42	2.24	2.01
17	4.45	3.59	3.20	2.96	2.81	2.70	2.55	2.38	2.19	1.96
18	4.41	3.55	3.16	2.93	2.77	2.66	2.51	2.34	2.15	1.92
19	4.38	3.52	3.13	2.90	2.74	2.63	2.48	2.31	2.11	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.45	2.28	2.08	1.84
21	4.32	3.47	3.07	2.84	2.68	2.57	2.42	2.25	2.05	1.81
22	4.30	3.44	3.05	2.82	2.66	2.55	2.40	2.23	2.03	1.78
23	4.28	3.42	3.03	2.80	2.64	2.53	2.38	2.20	2.01	1.76
24	4.26	3.40	3.01	2.78	2.62	2.51	2.36	2.18	1.98	1.73
25	4.24	3.38	2.99	2.76	2.60	2.49	2.34	2.16	1.96	1.71
26	4.22	3.37	2.98	2.74	2.59	2.47	2.32	2.15	1.95	1.69
27	4.21	3.35	2.96	2.73	2.57	2.46	2.31	2.13	1.93	1.67
28	4.20	3.34	2.95	2.71	2.56	2.45	2.29	2.12	1.91	1.65
29	4.18	3.33	2.93	2.70	2.54	2.43	2.28	2.10	1.90	1.64
30	4.17	3.32	2.92	2.69	2.53	2.42	2.27	2.09	1.89	1.62
40	4.08	3.23	2.84	2.61	2.45	2.34	2.18	2.00	1.79	1.51
60	4.00	3.15	2.76	2.52	2.37	2.25	2.10	1.92	1.70	1.39
∞	3.84	2.99	2.60	2.37	2.21	2.10	1.94	1.75	1.52	1.00

1 % CRITICAL VALUES OF F – DISTRIBUTION

v_2 \ v_1	1	2	3	4	5	6	8	12	24	∞
1	4052	4999	5403	5625	5764	5859	6106	6106	6235	6366
2	98.50	99.00	99.17	99.25	99.30	99.33	99.42	99.42	99.46	99.50
3	34.12	30.82	29.46	28.71	28.24	27.91	27.05	27.05	26.60	26.13
4	21.20	18.00	16.69	15.98	15.52	15.21	14.37	14.37	13.93	13.45
5	16.26	13.27	12.06	11.39	10.97	10.67	9.89	9.89	9.47	9.02
6	13.75	10.92	9.78	9.15	8.75	8.47	7.72	7.72	7.31	6.88
7	12.25	9.55	8.45	7.85	7.46	7.19	6.47	6.47	6.07	5.65
8	11.26	8.65	7.59	7.01	6.63	6.37	5.67	5.67	5.28	4.86
9	10.56	8.02	6.99	6.42	6.06	5.80	5.11	5.11	4.73	4.31
10	10.04	7.56	6.55	5.99	5.64	5.39	4.71	4.71	4.33	3.91
11	9.65	7.21	6.22	5.87	5.32	5.07	4.40	4.40	4.02	3.60
12	9.33	6.93	5.95	5.41	5.06	4.82	4.16	4.16	3.78	3.36
13	9.07	6.70	5.74	5.21	4.86	4.62	3.96	3.96	3.59	3.17
14	8.86	6.51	5.56	5.04	4.69	4.46	3.80	3.80	3.43	3.00
15	8.68	6.36	5.42	4.89	4.56	4.32	3.67	3.67	3.29	2.87
16	8.53	6.23	5.29	4.77	4.44	4.20	3.55	3.55	3.18	2.75
17	8.40	6.11	5.18	4.67	4.34	4.10	3.46	3.46	3.08	2.65
18	8.29	6.01	5.09	4.58	4.25	4.01	3.37	3.37	3.00	2.57
19	8.18	5.93	5.01	4.50	4.17	3.94	3.30	3.30	3.92	2.49
20	8.10	5.85	4.94	4.43	4.10	3.87	3.23	3.23	2.86	2.42
21	8.02	5.78	4.87	4.37	4.04	3.81	3.17	3.17	2.80	2.36
22	7.95	5.72	4.82	4.31	3.99	3.76	3.12	3.12	2.75	2.31
23	7.88	5.66	4.76	4.26	3.94	3.71	3.07	3.07	2.70	2.26
24	7.82	5.61	4.72	4.22	3.90	3.67	3.03	3.03	2.66	2.21
25	7.77	5.57	4.68	4.18	3.85	3.63	2.99	2.99	2.62	2.17
26	7.72	5.53	4.64	4.14	3.82	3.59	2.96	2.96	2.58	2.10
27	7.68	5.49	4.60	4.11	3.78	3.56	2.93	2.93	2.45	2.13
28	7.64	5.45	4.57	4.07	3.75	3.53	2.90	2.90	2.52	2.06
29	7.60	5.42	4.54	4.04	3.73	3.50	2.87	2.87	2.49	2.03
30	7.56	5.39	4.51	4.02	3.70	3.47	2.84	2.84	2.47	2.01
40	7.31	5.18	4.31	3.83	3.51	3.29	2.66	2.66	2.29	1.80
60	7.08	4.98	4.13	3.65	3.34	3.12	2.50	2.50	2.12	1.60
∞	6.64	4.60	3.78	3.32	3.02	2.80	2.18	2.18	1.79	1.00

1. Estimate linear supply function for the following data and evaluate it.

Supply (Y) in Million Tons	12	10	8	9	12	15	11	10	8	2
Price (X) in Rs. Per ton	5	10	7	8	12	14	10	8	5	2

Specification

Linear Supply function is

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

Y_i = Supply in million tons

X_i = Price Rs. Per ton

u_i = Error term

α = Constant regression Parameter

β = slope regression parameter

X	Y	X $(X-\bar{X})$	Y $(Y-\bar{Y})$	xy	x^2	y^2
5	12	- 3.1	2.3	- 7.13	9.61	5.29
10	10	1.9	0.3	0.57	3.61	0.09
7	8	- 1.1	- 1.7	1.87	1.21	2.89
8	9	- 0.1	- 0.7	0.07	0.01	0.49
12	12	3.9	2.3	8.97	15.21	5.29
14	15	5.9	5.3	31.27	34.81	28.09
10	11	1.9	1.3	2.47	3.61	1.69
8	10	- 0.1	0.3	- 0.03	0.01	0.09
5	8	- 3.1	- 1.7	5.27	9.61	2.89
2	2	- 6.1	- 7.7	46.97	37.21	59.29
$\Sigma X = 81$	$\Sigma Y = 97$			90.3	114.81	106.1

$\bar{X} = \frac{\Sigma X}{n} = \frac{81}{10} = 8.1$

$\bar{Y} = \frac{\Sigma Y}{n} = \frac{97}{10} = 9.7$

Estimation: Using OLS method of Estimation

$$\hat{\beta} = \frac{\Sigma xy}{\Sigma x^2} = \frac{90.3}{114.81} = 0.787$$

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X} = 9.7 - 0.787(8.1) = 3.33$$

Estimated supply function is $\hat{Y} = 3.33 + 0.787X$

Evaluation : Using Gauss – Markov theorem

$$r^2 = \frac{\hat{\beta}(\Sigma xy)}{\Sigma y^2} = \frac{(0.787)(90.3)}{106.1} = 0.67$$

$$\Sigma e^2 = (1 - r^2)(\Sigma y^2) = (1 - 0.67)(106.1) = 35.013$$

$$SE(\hat{\beta}) = \sqrt{\left(\frac{\Sigma e^2}{n-2}\right)\left(\frac{1}{\Sigma x^2}\right)} = \sqrt{\left(\frac{35.013}{8}\right)\left(\frac{1}{114.81}\right)} = 0.195$$

$$SE(\hat{\alpha}) = \sqrt{\left(\frac{\Sigma e^2}{n-2}\right)\left(\frac{\Sigma X^2}{n\Sigma x^2}\right)} \qquad \Sigma X^2 = \Sigma x^2 + n\bar{X}^2 = 114.81 + 10(8.1)^2 = 770.91$$

$$= \sqrt{\left(\frac{35.013}{8}\right)\left(\frac{770.91}{10(114.81)}\right)}$$

$$= 1.71$$

Tests of Significance

For $\hat{\beta}$

$$t_{n-2} = \frac{\hat{\beta}}{SE(\hat{\beta})} \quad (H_0: \beta = 0)$$

$$t_8 = \frac{0.787}{0.195} = 4.036$$

For $\hat{\alpha}$

$$t_{n-2} = \frac{\hat{\alpha}}{SE(\hat{\alpha})} \quad (H_0: \alpha = 0)$$

$$t_8 = \frac{3.33}{1.71} = 1.95$$

For degrees of freedom 8 (one sided test) $t_{5\%} = 1.860$ and $t_{1\%} = 2.896$

Here, for $\hat{\beta}$ calculated 't' value is greater than table value at 1% level of significance

Therefore, $\hat{\beta}$ is significant at 1% level.

For $\hat{\alpha}$ calculated 't' value is greater than table value at 5% level of significance.

Therefore $\hat{\alpha}$ is significant .

For R^2

$$F_{(p-1,\, n-p)} = \frac{\frac{R^2}{p-1}}{\frac{1-R^2}{n-p}} \qquad F_{(1,\, 8)} = \frac{\frac{0.67}{1}}{\frac{0.33}{8}} = \frac{0.67}{0.041} = 16.34$$

For degrees of freedom (1, 8) $F_{5\%} = 5.32$ and $F_{1\%} = 11.26$

calculated F value is greater than the table value at 1% level

Therefore, R^2 is significant at 1% level.

Fitted linear supply function is

$\bar{Y}$ = 3.33* + 0.787** X   with R^2 = 0.67**

SE($\widehat{\alpha}$) = 1.71   SE ($\widehat{\beta}$) = 0.195   n = 10

Note: NS denotes not significant

** denotes significant at 1% level

* denotes significant at 5% level

Application:

Price effect (β) = 0.787 implies that for every 1 rupee increase in the price of the commodity per ton, its supply is expected to increase by 0.787 million tons.

R^2 = 0.67 implies that 67% of change in supply is due to the linear influence of price. Since the estimated parameters and R^2 are significant, the estimated linear supply function is a good fit.

2. The following table indicates the price and quantity demanded of the productover a six year period

Year	2012	2013	2014	2015	2016	2017
Quantity (in Thousand Yards)	8	3	4	7	8	0
Price (in $ per Yard)	2	4	3	1	3	5

Estimate the demand function and evaluate it. And estimate the average elasticity of demand.

$\Sigma X = 18$
$\Sigma Y = 30$
$\overline{X} = 3$
$\overline{Y} = 5$
$\Sigma xy = - 19$
$\Sigma x^2 = 10$
$\Sigma y^2 = 52$
$\hat{\beta} = - 1.9$
$\hat{\alpha} = 5- (- 1.9)(3) = 5 – (-5.7) = 5 + 5.7 = 10.7$
Estimated demand function
$\hat{Y} = 10.7 – 1.9X$
$R^2 = 0.69$
$\Sigma e^2 = 16.12$
$SE(\hat{\beta}) = 0.635$
$\Sigma X^2 = 64$
$SE (\hat{\alpha}) = 2.07$
Test of significance
For $\hat{\beta}$ $t_4 = 2.99$
For $\hat{\alpha}$ $t_4 = 5.17$
For degrees of freedom 4 table value $t_{5\%} = 2.132$ and $t_{1\%} = 3.747$
For $\hat{\beta}$ the calculated t value is greater than the table value at 5% level of significance
Therefore $\hat{\beta}$ is significant at 5% level.
For $\hat{\alpha}$ the calculated value is greater than the table value at 5% as well as 1% level of significance
Therefore $\hat{\alpha}$ is significant at 1% level.
For R^2 $F_{(1, 4)} = 8.90$
For degrees of freedom (1, 4) , $F_{5\%} = 7.71$ and $F_{1\%} = 21.20$
F calculated value is greater than F table value at 5% level of significance . Therefore R^2 is significant at 5% level.

4. Fit a linear demand function for the following data. Also estimate e_p and interpret your answer.

Price	20	24	28	32	36	40
Demand	80	72	65	60	54	45

5. Estimate a linear cost function for the following data.

Production	40	42	48	55	65	79	88	100	120	140
Cost	150	140	160	170	150	162	185	165	190	185

Fitted Demand Function is

$\hat{Y}$ = 10.7** - 1.9*X with R^2 = 0.69

SE ($\hat{\alpha}$) = 2.07 SE ($\hat{\beta}$) = 0.635 n= 6

The price effect (β) = - 1.09 implies that for every one rupee increase in price per yard, its demand is expected to decrease by 1.09 thousand yards. The negative β value shows that there is inverse relationship between price and quantity demanded. Therefore the theory of demand also verified.

R^2 = 0.69 implies that 69% of change in demand is due to linear influence of price. Since the estimated parameters and R^2 are significant, the estimated simple linear demand function is a good fit.

The average elasticity of demand is

$$E_{YX} = \hat{\beta}\left(\frac{\bar{X}}{\bar{Y}}\right)$$

$$= -1.09\left(\frac{3}{5}\right)$$

$$= -0.654$$

6. Obtain 95% confidence interval for the parameters of the linear model when the values are in usual notation.

$\Sigma X = 301$ $\Sigma XY = 11459$ $\Sigma Y = 266$ $\Sigma X^2 = 12971$

$n = 7$ $\Sigma Y^2 = 10136$

Answer

Linear model is

$Y = \alpha + \beta X + u$

$\bar{X} = \frac{\Sigma X}{n} = \frac{301}{7} = 43$ $\bar{Y} = \frac{\Sigma Y}{n} = \frac{266}{7} = 38$

$\Sigma xy = \Sigma XY - n\bar{X}\bar{Y}$

$= 11459 - 7(43)(38)$

$= 21$

$\Sigma x^2 = \Sigma X^2 - n\bar{X}^2$

$= 12971 - 7(43)^2$

$= 28$

$\Sigma y^2 = \Sigma Y^2 - n\bar{Y}^2$

$= 10136 - 7(38)^2$

$= 28$

$\hat{\beta} = \frac{\Sigma xy}{\Sigma x^2} = \frac{21}{28} = 0.75$

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

$= 38 - 0.75(43)$

$= 5.75$

$R^2 = \frac{\hat{\beta}(\Sigma xy)}{\Sigma y^2} = \frac{0.75(21)}{28} = 0.56$

Estimated SLRM is $\hat{Y} = 5.75 + 0.75X$ with $R^2 = 0.56$

$\Sigma e^2 = (1 - R^2)\Sigma y^2$

$= (1 - 0.56) 28 = 12.32$

$SE(\hat{\beta}) = \sqrt{\left(\frac{\Sigma e^2}{n-2}\right)\left(\frac{1}{\Sigma x^2}\right)} = \sqrt{\left(\frac{12.32}{5}\right)\left(\frac{1}{28}\right)} = 0.296$

$SE(\hat{\alpha}) = \sqrt{\left(\frac{\Sigma e^2}{n-2}\right)\left(\frac{\Sigma X^2}{n\Sigma x^2}\right)}$

$= \sqrt{\left(\frac{12.32}{5}\right)\left(\frac{12971}{7(28)}\right)}$

$= 12.77$

95% confidence interval for $\hat{\beta}$

$\hat{\beta} \pm T_{0.05}\, SE(\hat{\beta}) = 0.75 \pm 2.015\,(0.296)$

$= 0.75 + 0.596,\ 0.75 - 0.596$

$= 1.346,\ 0.154$

95% confidence interval for $\hat{\alpha}$

$\hat{\alpha} \pm T_{0.05}\,(SE(\hat{\alpha})) = 5.75 \pm 2.015(12.77)$

$= 5.75 + 25.732,\ 5.75 - 25.732$

$= 31.48,\ -19.982$

7. Obtain 95% and 99% confidence limits for parameters for the following intermediate results.

$\Sigma X = 150$ $\Sigma Y = 35$ $\Sigma XY = 900$ $\Sigma X^2 = 5500$

$\Sigma Y^2 = 255$ $n = 5$

Answer

SLRM is

$Y = \alpha + \beta X + u$

$\bar{X} = \frac{\Sigma X}{n} = \frac{150}{5} = 30$ $\bar{Y} = \frac{\Sigma Y}{n} = \frac{35}{5} = 7$

$\Sigma xy = \Sigma XY - n\bar{X}\bar{Y} = 900 - 5(30)(7) = -150$

$\Sigma x^2 = \Sigma X^2 - n\,\bar{X}^2$

$= 5500 - 5(30)^2$

$= 1000$

$\Sigma y^2 = \Sigma Y^2 - n\,\bar{Y}^2$

$= 255 - 5(7)^2$

$= 10$

$\hat{\beta} = \frac{\Sigma xy}{\Sigma x^2} = \frac{-150}{1000} = -0.15$

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

$= 7 - (-0.15)(30)$ $= 7 + 4.5$ $= 11.5$

$R^2 = \frac{\hat{\beta}(\Sigma xy)}{\Sigma y^2} = \frac{-0.15(-150)}{10} = 2.25$

Estimated SLRM is $\hat{Y} = 11.5 - 0.15X$ with $R^2 = 2.25$

$\Sigma e^2 = (1 - R^2)\,\Sigma y^2$

$= (1 - 2.25)\,10$ $= -12.5$

$SE(\hat{\beta}) = \sqrt{\left(\frac{\Sigma e^2}{n-2}\right)\left(\frac{1}{\Sigma x^2}\right)} = \sqrt{\left(\frac{-12.5}{3}\right)\left(\frac{1}{1000}\right)} = -0.063$

$SE(\hat{\alpha}) = \sqrt{\left(\frac{\Sigma e^2}{n-2}\right)\left(\frac{\Sigma X^2}{n\Sigma x^2}\right)}$

$= \sqrt{\left(\frac{-12.5}{3}\right)\left(\frac{5500}{5(1000)}\right)}$ $= -2.14$

95% confidence interval for $\hat{\beta}$

$\hat{\beta} \pm T_{0.05}\,SE(\hat{\beta})$ $= -0.15 \pm 2.353\,(-0.063)$

$= -0.15 + (-0.148),\ -0.15 - (-0.148)$

$= -0.298,\ -0.002$

95% confidence interval for $\hat{\alpha}$

$\hat{\alpha} \pm T_{0.05}\,(SE(\hat{\alpha}))$ $= 11.5 \pm 2.353(-2.14)$

$= 11.5 \pm (-5.035)$

$= 11.5 + (-5.035),\ 11.5 - (-5.035)$

$= 11.5 - 5.035,\ 11.5 + 5.035$

$= 6.465,\ 16.535$

99% confidence interval for $\hat{\beta}$

$\hat{\beta} \pm T_{0.01}\,SE(\hat{\beta})$ $= -0.15 \pm 4.541(-0.063)$

$= -0.15 \pm (-0.286)$

$= -0.15 + (-0.286),\ -0.15 - (-0.286)$

$= -0.15 - 0.286,\ -0.15 + 0.286$

$= -0.436,\ 0.136$

99% confidence limits for $\hat{\alpha}$

$\hat{\alpha} \pm T_{0.01}\,(SE(\hat{\alpha}))$ $= 11.5 \pm 4.541\,(-2.14)$

$= 11.5 \pm (-9.72)$

$= 11.5 + (-9.72),\ 11.5 - (-9.72)$

$= 11.5 - 9.72,\ 11.5 + 9.72$

$= 1.78,\ 21.22$

8. Estimate and evaluate linear consumption function for n = 18, $\bar{C}$ = 32, $\bar{Y}$ = 75
(C = consumption Y = Income) ΣC^2 = 6293,
ΣY^2 =11390, ΣCY = 4064. Interpret your results.
Answer
Consumption depends on Income
Dependent variable = Consumption (C)
Independent variable = Income (Y)

Specification

Linear consumption function is
C = α + βY + u
Where, C = Consumption
Y = Income
u = Error term
α = Constant regression parameter
β = Slope regression parameter

Estimation (Y = C, X =Y)

$\hat{\beta} = \frac{\Sigma yc}{\Sigma y^2}$

$\Sigma yc = \Sigma YC - n(\bar{Y})(\bar{C})$
= 4064 – 18(75)(32)
= - 39136

$\Sigma y^2 = \Sigma Y^2 - n\,\bar{Y}^2$
$= 11390 - 18(75)^2$
= - 89860

$\Sigma c^2 = \Sigma C^2 - n\,\bar{C}^2$
$= 6293 - 18(32)^2$
= - 12139

$\hat{\beta} = \frac{\Sigma yc}{\Sigma y^2} = \frac{-39136}{-89860} = 0.435$

$\hat{\alpha} = \bar{C} - \hat{\beta}\bar{Y}$
= 32 – 0.435(75)
= - 0.625

Estimated linear consumption function is
$\hat{C} = -0.625 + 0.435Y$

Evaluation

$r^2 = \frac{\hat{\beta}(\Sigma cy)}{\Sigma c^2} = \frac{0.435(-39136)}{-12139} = 1.402$

$\Sigma e^2 = (1 - r^2)(\Sigma c^2) = (1 - 1.402)(-12139) = 4879.8$

$SE(\hat{\beta}) = \sqrt{\left(\frac{\Sigma e^2}{n-2}\right)\left(\frac{1}{\Sigma y^2}\right)} = \sqrt{\left(\frac{4879.8}{16}\right)\left(\frac{1}{-89860}\right)} = -0.174$

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

$= \sqrt{\left(\frac{4879.8}{16}\right)\left(\frac{11390}{18(-89860)}\right)} = -1.465$

Tests of Significance

For $\hat{\beta}$

$$t_{n-2} = \frac{\hat{\beta}}{SE(\hat{\beta})} \quad (H_0: \beta = 0)$$

$$t_{16} = \frac{0.435}{0.174} = 2.5$$

For $\hat{\alpha}$

$$t_{n-2} = \frac{\hat{\alpha}}{SE(\hat{\alpha})} \quad (H_0: \alpha = 0)$$

$$t_{16} = \frac{0.625}{1.465} = 0.427$$

For degrees of freedom 16 (one sided test) $t_{5\%}$ =1.746 and $t_{1\%}$ = 2.583

Here, for $\hat{\beta}$ calculated 't' value is greater than table value at 5% level of significance

Therefore, $\hat{\beta}$ is significant at 5% level.

For $\hat{\alpha}$ calculated 't' value is lesser than the table value at5% and 1% level of significance

Therefore $\hat{\alpha}$ is not significant .

For R^2

$$F_{(p-1,\, n-p)} = \frac{\frac{R^2}{p-1}}{\frac{1-R^2}{n-p}} \qquad F_{(1,\, 16)} = \frac{\frac{1.402}{1}}{\frac{1-1.402}{16}} = \frac{1.402}{-0.025} = 56.08$$

For degrees of freedom (1, 16) $F_{5\%}$ =4.49 and $F_{1\%}$ =8.53

calculated F value is greater than table value at 1% and 5% level of significance

Therefore, R^2 is significant at 1% level

Fitted linear consumption function is

$\hat{C} = 0.625^{NS} + 0.435^{*} X$ with R^2 =1.402**

SE($\hat{\alpha}$) = 1.465 SE($\hat{\beta}$) = 0.174 n = 18

Note: NS denotes not significant , * denotes Significant at 5%, ** denotes significant at 1%

Application

The constant regression parameter is not significant . Its positive value implies that the total consumption C is 0.625 when the income is zero.

The slope regression parameter is significant at 5% level. Its positive value implies that if income increases consumption will also increase. If 1 rupee increase in income will increase the consumption by Rs. 0.435 .

R^2 = 1.402 implies that 140.2% change in consumption is due to the linear influence of income.

9. Estimate and evaluate food demand function(Y) on national product (X) for n = 25,
$\Sigma X = 450$, $\Sigma Y = 325$, $\Sigma x^2 = 394.4$, $\Sigma y^2 = 21.6$,
$\Sigma xy = 80.20$.

Answer

Specification

The linear food demand function is

$Y = \alpha + \beta X + u$

Y = Food demand

X = National Product

α = constant regression parameter

β = slope regression parameter

u = error term

Estimation

$\hat{\beta} = \frac{\Sigma xy}{\Sigma x^2} = \frac{80.20}{394.4} = 0.203$

$\bar{X} = \frac{\Sigma X}{n} = \frac{450}{25} = 18$ $\bar{Y} = \frac{\Sigma Y}{n} = \frac{325}{25} = 13$

$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X} = 13 - 0.203\,(18) = 9.346$

Estimated food demand function is

$\hat{Y} = 9.346 + 0.203X$

Evaluation

$R^2 = \frac{\hat{\beta}(\Sigma xy)}{\Sigma y^2} = \frac{0.203\,(80.20)}{21.6} = 0.754$

$\Sigma e^2 = (1 - R^2)\,\Sigma y^2$
$= (1 - 0.754)\,(21.6) = 5.314$

$SE(\hat{\beta}) = \sqrt{\left(\frac{\Sigma e^2}{n-2}\right)\left(\frac{1}{\Sigma x^2}\right)} = \sqrt{\left(\frac{5.314}{23}\right)\left(\frac{1}{394.4}\right)} = 0.022$

$SE(\hat{\alpha}) = \sqrt{\left(\frac{\Sigma e^2}{n-2}\right)\left(\frac{\Sigma X^2}{n\Sigma x^2}\right)}$ $\Sigma X^2 = \Sigma x^2 + n\bar{X}^2 = 394.4 + 25(18)^2 = 8494.4$

$= \sqrt{\left(\frac{5.314}{23}\right)\left(\frac{8494.4}{25(394.4)}\right)} = 0.446$

Tests of Significance

For $\hat{\beta}$

$$t_{n-2} = \frac{\hat{\beta}}{SE(\hat{\beta})} \quad (H_0: \beta = 0)$$

$$t_{23} = \frac{0.203}{0.02} = 10.15$$

For $\hat{\alpha}$

$$t_{n-2} = \frac{\hat{\alpha}}{SE(\hat{\alpha})} \quad (H_0: \alpha = 0)$$

$$t_{23} = \frac{9.346}{0.446} = 20.96$$

For degrees of freedom 23 (one sided test) $t_{5\%} = 1.714$ and $t_{1\%} = 2.500$

Here, for $\hat{\beta}$ calculated 't' value is greater than table value at 1% level of significance

Therefore, $\hat{\beta}$ is significant at 1% level.

For $\hat{\alpha}$ calculated 't' value is greater than table value at 1% level of significance.

Therefore $\hat{\alpha}$ is significant at 1% level.

For R^2

$$F_{(p-1,\, n-p)} = \frac{\frac{R^2}{p-1}}{\frac{1-R^2}{n-p}} \qquad F_{(1,\, 23)} = \frac{\frac{0.754}{1}}{\frac{1-0.754}{23}} = \frac{0.75}{0.011} = 68.54$$

For degrees of freedom (1, 23) $F_{5\%} = 4.28$ and $F_{1\%} = 7.88$

calculated F value is greater than the table value at 1% level

Therefore, R^2 is significant at 1% level.

Fitted Food Demand Function is

$\hat{Y} = 9.346^{**} + 0.203^{**}X$ with $R^2 = 0.754^{**}$

$SE(\hat{\alpha}) = 0.446$ $SE(\hat{\beta}) = 0.02$ $n = 25$

** denotes significant at 1%

Application

The constant regression parameter is significant at 1% level. It's positive value implies that the total food demand is 9.346 when the national product zero.

The slope regression parameter is significant at 1% level. It's positive value implies that for every 100 unit increase in national production there is an increase of 20.3 units of demand.

R^2 is significant at 1% level. It's value implies that 75.4% change in food demand is due to the linear influence of national product. The remaining 24.6 % of change is due to the influence of other factors.

Multiple Linear Regression Model (MLRM)

A multiple linear regression model (MLRM) includes many explanatory variables. Consider a multiple linear consumption function with two explanatory variables of the form;

$$Y_i = \beta_0 + \beta_{1i} X_{1i} + \beta_{2i} X_{2i} + u_i$$

Where

Y_i = Consumption expenditure for i^{th} household

X_{1i} = Consumer's disposable income

X_{2i} = Wealth of i^{th} household

u_i = Error term

β_0 = constant regression parameter

β_1 = Slope or partial regression parameter

β_2 = slope or partial regression parameter

β_1 and β_2 provide changes in the dependent variable Y for a unit change in the respective explanatory variable.

According to keynesian Law of consumption,

β_0 is positive which is the minimum subsistence level of consumption

β_1 is positive due to direct relation between income and consumption

β_2 is also positive due to direct relation between wealth and consumption.

UNIT - IV

SPECIFICATION IN OLS ESTIMATION

Auto Correlation- Multi-collinearity – heterocedasticity – causes and consequences

MULTI – COLLINEARITY

Meaning: The term multi collinearity is used to denote the presence of perfect linear relationship between the explanatory variables of the multiple linear regression model.

For application of least squares, there is an important condition, that the explanatory variables are not perfectly correlated (there is no correlation between explanatory variables r_{x1} ; $r_{x2} \neq 1$)

If there is perfect linear correlation between the explanatory variables, the computed parameters will not be correct one.(because by the method of least squares the condition for least squares in those variables should not correlated) On the other hand the explanatory variables are not perfectly correlated then the variables are called Orthogonal (r_{x1} ; $r_{x2} \neq 0$)

Sources or Causes of Multi-collinearity

1. In time series data significant secular trend leads to multi collinearity. It is due to the fact that many economic variables tend to move together over time. For example population, production, laour force, income, investment and price.
2. It arises due to the use of lagged value of the explanatory variables in the multiple linear regression model $C_1 = f(Y_t, Y_{t-1})$. It occurs in the distributed log model (Auto Regressive model).

   Thus multi collinearity is a serious problem in time series data and it is quite frequent in cross section data also.

Consequences of Multi collinearity

If there is perfect linear relationship between explanatory variables, that is $r_{x1x2} = 1$, then

1. The estimates of the co-efficients are indeterminate.
2. The standard errors or variances of estimates become infinitely large.

I. Take two explanatory variables model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u$$

Where X_1 and X_2 are perfectly correlated with the exact relation $X_1 = aX_2$

The formula for the $\hat{\beta}_1$ and $\hat{\beta}_2$ are

$$\beta_1 = \frac{(\Sigma x_1 y)(\Sigma x^2{}_2) - (\Sigma x_2 y)(\Sigma x_1 x_2)}{(\Sigma x_1{}^2)(\Sigma x_2{}^2) - (\Sigma x_1 x_2)^2}$$

$$\beta_2 = \frac{(\Sigma x_2 y)(\Sigma x^2{}_1) - (\Sigma x_1 y)(\Sigma x_1 x_2)}{(\Sigma x_1{}^2)(\Sigma x_2{}^2) - (\Sigma x_1 x_2)^2}$$

Substituting ax_2 for x_1

$$\beta_1 = \frac{a(\Sigma x_2 y)(\Sigma x^2{}_2) - a(\Sigma x_2 y)(\Sigma x_2 x_2)}{a^2(\Sigma x_2{}^2)(\Sigma x_2{}^2) - a^2(\Sigma x_2 x_2)^2} = 0$$

$$\beta_2 = \frac{a(\Sigma x_2 y)(\Sigma x^2{}_2) - a(\Sigma x_2 y)(\Sigma x_2 x_2)}{a^2(\Sigma x_2{}^2)(\Sigma x_2{}^2) - a^2(\Sigma x_2 x_2)^2} = 0$$

Therefore, co-efficients are indeterminate

2. Take $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u$

If X_1 and X_2 are perfectly correlated ($x_1 = a x_2$) then the variables of β_1 and β_2 will be

Variance of $(\hat{\beta}_1) = [\frac{\Sigma e_i^2}{n-3}] [\frac{\Sigma x_2^2}{(\Sigma x_1^2)(\Sigma x_2^2) - (\Sigma x_1 x_2)^2}]$

Substituting ax2 for x1

Variance of $(\hat{\beta}_1) = [\frac{\Sigma ei^2}{n-3}] [\frac{\Sigma x_2^2}{a^2(\Sigma x_2^2)(\Sigma x_2^2) - a^2(\Sigma x_2 x_2)^2}]$

$= [\frac{\Sigma ei^2}{n-3}] [\frac{\Sigma x_2^2}{0}] = \infty$

Therefore estimators have large standard errors

3. Estimates of co-efficients will not be significant due to large standard error values, so they are not statistically reliable. Investigators may drop some variables from the model due to their insignificance.

Tests for Multi-collinearity

to test the presence of multi-collinearity in the model the following tests may be conducted.

1. Zero-order correlation matrix (crude method)

Higher zero order correlation along with low $\bar{R}2$ are considered as the indication of presence of collinearity among the explanatory variables.

2. Frisch's Confluence Analysis or Bunch map analysis.

3. Farrar-Glauber Test

Solutions for Multi-collinearity

1. Increase the size of the sample
2. Introduce additional equations
3. Use of extraneous information (extraneous information may be available from economic theories and past studies)
4. Pooling time series and cross section data.

Auto-correlation

Meaning: Auto-correlation refers to the relationship between the succesive values of the disturbance term (u). It occurs when the assumption of $E(u_i u_j)=0$ is violated. Thus in the presence of auto-correlation $Cov(u_i u_j)=E(u_i u_j)\neq 0$

Or $E(u_t u_{t-1}) \neq 0$. Which implies that the values of u in any particular period is correlated with its own proceeding value.

Sources of Auto-Correlation

Auto correlation usually does not arise in cross section data. According to Tintner, it is a lag correlation between two different series.

1. Auto correlation arises in the time series sample, which exhibits a significant long term movement over time.
2. Cyclical fluctuations, which impose regularity among successive observations of the variables over time cause auto correlation.
3. It arises due to specification bias. Which occurs due to exclusion of true variables from the regression equation.
4. In case of Auto regressive model one of the explanatory variables is lagged value of the endogenous variable of the form $C_t = f(Y_t, C_{t-1})$ leads to auto correlation.

Consequences of Auto-Correlation

1. OLS estimators are unbiased and linear
2. The variance of OLS estimators are undeterminated.
3. The OLS estimators will not be the best estimators
4. The predictions will not be efficient.

Tests for Auto-Correlation

1. Graphic method: By regressing the model, derive et values, plot the points in a graph sheet. If the scatter follows a sloping upwards line, then the presence of possitive correlation is shown. On the other hand, if the scatter shows an up and down movement, then the presence of negative auto correlation is shown. If the scatter does not show any of these pattern, then there is no presence of auto correlation.
2. Von-Neuman Ratio Test
3. Durbin – Watson Test

Solution for Auto-Correlation

In order to delete the effects of auto-correlation the given model is to be transformed for the purpose of obtaining the best linear unbiased estimators of the parameters.

HETEROSCEDASTICITY

Meaning: In a simple linear regression model, the OLS estimators are derived by assuming that $E(u_i^2) = \sigma^2 u$ for all I = 1,2,.....n. That is, the error terms are assumed to have homoscedastic variances. In some cases the disturbance terms do not have same variance. Such situation of non-homogenity of variance is called heteroscedasticity.

$E(u_i^2) = Var(u_i) \neq \sigma^2 u$ or $Var(u_i) = f(X_i)\, \sigma^2 u = \sigma^2 u_i$ for I = 1,2,3....n, when $Y_i = \alpha + \beta X_i + u_i$

Sources and types of Heteroscedasticity

Heteroscedasticity occurs only in cross section data with the following types

i. As X_i value increases, Var (u_i) also increases. So var (u_i) = $K^2X_i^2$

ii. As X_i value increases, var(u_i) is maximum in the middle value. So Var(u_i) = f(X_i)

iii. As X_i increases, Var(u_i)decreases, So, Var(u_i) = $K^2(1/X_i^2)$.

For eg. Consider family budget study of consumption

$C_i = \alpha + \beta Y_i + u_i$ where C_i is household consumption and Y_i is disposable income of household. At low level of income, variations in consumption are not possible as it implies starvation at limited money income. So, consumption patterns are more similar at lower income levels than at higher levels. This shows type (a) heteroscedasticity.

In a cross section study of C-D production function

$X_i = AL_i^{\alpha}K_i^{\beta}e^{ui}$ the level of output is influenced by technology and economies of scale. So there will be much of variation among large units than among small units with type (a) heteroscedasticity.

Consequences of Heteroscedasticity

In the simple linear regression model,

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

1. The OLS estimators are still linear function of Y_i.
2. The OLS estimators are still unbiased.
3. Variances of OLS estimators will not correct.
4. OLS estimators will not be the best estimators.

Tests for Heteroscedasticity

1. The spearman Rank Correlation Test: First regress Y on X to obtain the residuals $e = Y - \hat{Y}$. Next arrange X values in ascending order and compute rank correlation between X and e. High and significant correlation indicates the presence of heteroscedasticity.
2. The Goldfeld and Quandt test
3. The Park test
4. The Glejser test

Solution for Heteroscedasticity

After testing for presence of a particular form of heteroscedasticity, suitable method of transformation of the given model is followed.

UNIT-V
PROBLEMS IN OLS ESTIMATION

Specification Problems and bias-errors in variables – dummy variables –lag models.

SPECIFICATION ERROR

Meaning: The specification of a model consists of formulation of the regression equation and of assumptions concerning the variables. The most common specification errors are those resulting from,

1. Omission of relevant explanatory variable from the function.
2. Inclusion of irrelevant explanatory variable in the function.
3. Omission of some equations from the model.
4. Incorrect mathematical form of the function.

Sources

Specification error arises due to

1. Limited knowledge of the variables.
2. Non-availability of required data and
3. Imperfection of economic theory.

Consequences

1. Case of omission of relevant explanatory variable

Consider the true function $Y = \beta_0+\beta_1X_1+\beta_2X_2+u$, whereas $Y= \beta_0+\beta_1X_1+u$ is estimated.

i) OLS estimator of β_1 will be unbiased and consistent if the omitted explanatory variable X_2 is not correlated with the variable X_1 in the estimated function. The estimator of intercept will be biased and inconsistent.

ii) OLS estimator of β_1 will be biased and inconsistent if the omitted explanatory variable is related with the other explanatory variables in the estimated model.

iii) The variance of β_1 will always have an upward bias if the omitted explanatory variable is related with the other explanatory variables in the estimated model.

iv) Test of significance of estimators would not lead to correct conclusions.

2. Case of inclusion of irrelevant explanatory variable

Consider the true function $Y = \beta_0+\beta_1X_1+u$, whereas $Y=\beta_0+\beta_1X_1+\beta_2X_2+u$ is estimated.

i) OLS estimator of β_1 will be consistent if the included explanatory variable X_2 is not correlated with the variable X_1 in the estimated function.

ii) Otherwise if the included explanatory variable X_2 is correlated with variable X_1 in the estimated function, β_1 will be an inefficient estimator with greater variance.

iii) Larger variance reduces the precision of the estimates and the confidence intervals become wider

3. Case of incorrect functional form

Suppose the true marginal cost function is $MC=\beta_0+\beta_1X_1+\beta_2X_2^2+u$, whereas the linear function $MC=\beta_0+\beta_1X_1+u$ is estimated. Therefore OLS estimator will be biased.

Solution

Consult economic theory and past studies to avoid mis-specification.

MEASUREMENT ERROR

Meaning: Measurement errors refers to the errors in the measurement of either the dependant variable or the explanatory variables or both. It is different from equation error.

Sources

All economic data are subject to some errors of measurement due to

1. Errors in sampling, errors in extrapolation of sample results and aggregation in most of the published data.
2. Use of variable different in content such as GNP in place of disposable income due to non-availability of income variable for estimation of Keynesian consumption function.
3. Use of inappropriate price indices for the purpose of deflating current values to arrive at constant price values in time series data.
4. Use of indices as explanatory variables for the regression model.

Consequences

1. Errors of measurement in endogenous variable

In the SLRM $Yi^* = \alpha+\beta X_i+w_i$ where $w_i=u_i+v_i$ will satisfy the usual assumptions if both u_i and v_i satisfy them and cov $(u_i, v_i) = 0$. Here u_i is equation error and v_i is measurement error.

i) $\hat{\beta}$ will be an unbised estimator

ii) The variances of estimated parameters are larger than in the case when there are no measurement errors.

iii) The estimators will be inefficient and so they will not be BLUE.

2. Errors in measurement of explanatory variable

In the SLRM $Yi^* = \alpha+\beta X_i^*+u_i$ suppose X_i^* is the true value and $X_i = X_i^* +v_i$ is the observed value.

So the model becomes $Yi = \alpha+\beta X_i+w_i$, where $w_i = u_i - \beta v_i$.

i) OLS estimators will be biased

ii) OLS estimators will be inconsistent

So the errors of measurement pose agreat threat to the validity of the standard interpretation of the estimators.

Solutions

1. **Inverse least squares method**: This method is appropriate only when errors of measurement are found in the explanatory variable but not in the dependent variable.
2. **Wald's Two Group Method**: This method is applied when there are errors in the measurement of either X or Y or both.
3. Bartlett's Three Group Method: This method is applied when there are errors in the measurement of either X or Y or both.
4. Weighted Regression Method: This method is applied when there are errors in the measurement of either X or Y or both.

LAGGED VARIABLES

Meaning: In economics generally a cause produces its effect only after a lapse of time, called a lag. These lags are of importance in decision making by the planners to know the fastness of impact of tax on consumers and incentives on production.

For eg. Consumption function $C_t = f(Y_t, Y_{t-1}, Y_{t-2}, \ldots..C_{t-1}, C_{t-2}\ldots.)$

Investment function $I_t = f(X_t, X_{t-1}, X_{t-2}\ldots..)$

Types

The endogenous variable depends upon current and past values or lagged values of the endogenous variables.

1. **Distributed Lag Model**

It is a model including only lagged values of the explanatory variables

$$Y_t = \beta_0 X_t + \beta_1 X_{t-1} + \beta_2 X_{t-2} + \ldots.. + \beta_8 X_{t-8} + u_t$$

The number of lags in the model may be either finite or infinite.

Short run or impact multiplier = β_0

Long run or distributed lag multiplier = $\sum_{i=0}^{n} \beta i$

Delayed or interim multiplier are $\beta_{1,}\ \beta_{2,}\ \ldots\beta_8$

2. **Auto regressive Model**

It is a model including lagged values of the endogenous variable in the form as, $Y_t = \alpha + \beta X_t + rY_{t-1} + u_t$

Sources for Lag

1. Technical Reasons: Production requires time and supply of a commodity depending upon the production process, which also depends upon lagged prices of inputs.
2. Institutional Reasons: Contractual obligations and certain rules like fixity of deposits will take time to respond to better money market conditions.
3. Psychological reasons: Behaviour is often based on taste and habit. The change in consumption habit is a slow process depending upon whether the income change is permanent or transitory.

Consequences

1. One period lag model reduces the sample size by one.
2. If the number of lag is large and n is small and it may not be possible to estimate the parameters for estimating MLRM.
3. Multi-collinearity may arise due to strong correlation between the successive values of the same variable.
4. The OLS estimators may become biased and inefficient.

Solutions

There are two approaches to solve the problem due to lagged variable.

1. Kyock Approach to Distributed lag model

It is based on the assumption that the variables in the distant past have smaller impact. So the β values are declining continously (Geometric lag)

2. **Almon Approach to the Distributed Lag models**

It is a far more flexible method in terms of the form of lag scheme and it does not lead to violation of the basic assumptions of the error term.

DUMMY VARIABLE

Meaning: Dummy variables are those of qualitative characters such as profession, religion, sex, region, season, literacy, and war period. Since such characteristics can not be measured, we assign a value of 1 to the presence and zero to the absence of the attribute. These are also called binary variables, indicator variables or categorical variables.

Types of Dummy Variables

There are two types of model with dummy variables.

1. Analysis of Variance (AOV) Models

Models that contain only dummy variables for explanatory variables are called Analysis of variance (AOV) models, which are common in sociology and education. In econometric research

consider $Y_i=\alpha+\beta D_i+u_i$

Where Yi is salary of i^{th} worker and D_i represents literacy as a dummy variable.

Di = 1 if the worker has high school level education and above

Di = 0 if otherwise.

Assuming that ui satisfies all the assumptions under OLS

We get, $\hat{y} = \hat{\alpha}$ for Di = 0 and $\hat{y} = \hat{\alpha} + \hat{\beta}$ for D_i =1.

So $\hat{\beta}$ measures the difference in the worker's mean salary due to higher education.

2. Analysis of Co-Variance (ACOV) Models

In most econometric research, models contain admixture of qualitative and quantitative variables, which are called Analysis of Co-Variance (ACOV) models.

Consider the Consumption function

$$C_i = \beta_0 + \beta_1 Y_1 + \beta_2 D_{1i} + \beta_3 D_{2i} + u_i$$

Where D_{1i} = 1 if the household has children

D_{1i} = 0 if the household has no children

D_{2i} = 1 if the household owns a house

D_{2i} = 0 if the household does not owns a house

$\hat{\beta}2$ measures the difference in consumption due to children

$\hat{\beta}3$ measures the difference in consumption due to wealth.

Rules of Dummy Variables

1. By a priori consideration we should assign 0 and 1 to two categories. The category for which zero is assigned is referred as base or control category.
2. If the qualitative variable has n categories, then introduce only (n-1) dummy variables to avoid dummy variable trap. Otherwise estimation by OLS is not possible due to perfect multi-collinearity.
3. The coefficient attached to the dummy variable is called the differential coefficient due to that variable.

Uses of Dummy variables

1. To measure the shift of a function over time

During war times controls restrict the availability of consumer goods which may shift the consumption function downwards.

Instead of usual function $C_t = \alpha + \beta_1 Y_t + u_t$

We consider $C_t = \alpha + \beta_1 Y_t + \beta_2 D_t + \beta_3 Y_t D_t + u_t$

Where

D_t = 1 for war years and zero for normal years.

$\hat{\beta}_2$ gives differential effect on subsistence level of consumption and

$\hat{\beta}_3$ gives differential effect on MPC due to war.

$\hat{C}_t = (\hat{\alpha} + \hat{\beta}_2) + (\hat{\beta}_1 + \hat{\beta}_3)Y_t$ War period equation

$\hat{C}_t = \hat{\alpha} + \hat{\beta}_1 Y_t$ Normal period equation

We expect $\hat{\beta}_2$ and $\hat{\beta}_3$ are negative.

2. To isolate seasonal component from time series

In case of quarterly data for economic time series deseasonalisation is done by identifying presence of seasonal pattern in the intercept and slope of the SLRM $Y_t = \alpha + \beta X_t + u_t$ as

$$Y_t = \alpha + \beta X_t + r_1 D_1 + r_2 D_2 + r_3 D_3 + e_1 D_1 X_t + e_2 D_2 X_t + e_3 D_3 X_t + u_t$$

With $D_1 = 1$ if it is Q_2 and 0 otherwise

$D_2 = 1$ if it is Q_3 and 0 otherwise

$D_3 = 1$ if it is Q_4 and 0 otherwise

The significance of r values and e values will reflect the presence of seasonal pattern respectively in the intercepts or slope values

3. To estimate discriminating functions

Consider a qualitative dependent variable $D_i = \alpha + \beta X_i + u^i$ to explain purchase of car (D_1) on family income (X_i)

$D_i = 1$ if the family purchases a new car and 0 otherwise. In this case u_i's will not follow normal distribution and they are heteroscedastic. So OLS estimates although unbiased are not efficient.