[Academic Script]

Multiple Regression Model

Subject:	Business Economics
Course:	B.A., 3rd Semester, Undergraduate
Paper No. & Title:	Paper – 304 Basic Econometrics
Unit No. & Title:	Unit - 3 Multiple Regression Model
Lecture No. & Title:	Lecture – 1 Multiple Regression Model

Academic Script

1. Introduction

Regression analysis is a technique of predicting a dependent variable using one or more independent variables.

In **Simple regression** the model has one dependent and one independent variable. Its equation is given as

$$Y = b_0 + b_1 x + e$$

where Y is the response or dependent variable and X is the independent or explanatory or predictor variable.

Here b_0 is the intercept of the line, $b_1$is the regression coefficient and e is the residual or error term.

Eg1.

- Independent: First year mileage for a certain car model;
  Dependent: Resulting maintenance cost.

Eg2.

- Independent: Height of fathers;
  Dependent: Height of sons.

Multiple regressions is a logical extension of simple regression. Instead of using just one independent variable as in simple regression, several independent variables are used. With multiple regressions, we form a 'linear combination' of multiple variables to best predict an outcome, and then we assess the contribution that each independent variable makes to the equation.

Its equation is given as

$$Y = b_0 + b_1 X_1 + b_2 X_2 + b_3 X_3 + \ldots\ldots\ldots + b_p X_p + e$$

where Y is the response or dependent variable and X_i is the independent or explanatory or predictor variable.

Here b_0 is the intercept of the line, b_i is the regression coefficient of X_i variable and e is the residual or error term.

Eg1.

• Independent: square feet of home, locality of home, quality of construction;

• Dependent: Current market value of home.

Eg2.

• Independent: advertising budget, number of times played on radio station, number of times shown on TV;

• Dependent: Sales of a music CD.

Let us restrict our study to three independent variables only. So the equation will be of the form:

$$Y = b_0 + b_1 X_1 + b_2 X_2 + b_3 X_3 + e$$

2. Derivation of regression coefficients

Using method of least square we can obtain four normal equations. These normal equations are

$$\sum Y = Nb_0 + b_1 \sum X_1 + b_2 \sum X_2 + b_3 \sum X_3$$

$$\sum X_1 Y = b_0 \sum X_1 + b_1 \sum X_1^2 + b_2 \sum X_1 X_2 + b_3 \sum X_1 X_3$$

$$\sum X_2 Y = b_0 \sum X_2 + b_1 \sum X_1 X_2 + b_2 \sum X_2^2 + b_3 \sum X_3 X_2$$

$$\sum X_3 Y = b_0 \sum X_3 + b_1 \sum X_1 X_3 + b_2 \sum X_2 X_3 + b_3 \sum X_3^2$$

Solving these four equations we can get the values of regression coefficients. And substituting these regression coefficients, we get the best fitted regression equation.

3. Assumptions for multiple regression analysis

The basic issue is whether, in the course of calculating the regression coefficients and the predicting the dependent variable, the assumptions of regression analysis are met. We need to check these assumptions because violation of any of these assumptions may lead to undesirable results.

The assumptions to be examined are in four areas:

- Linearity
- Constant variance of error term
- Independence of error terms
- Normality of the error term distribution

1. **Linearity**: There is a linear relationship between dependent and independent variables. It is important as it represents the degree to which the change in dependent variable is associated with the independent variable.
2. **Constant variance of error term**: At each level of the independent variables, the variance of the residual terms should be constant (homoscedasticity). The presence of unequal variances (heteroscedasticity) is one of the most common assumption violations.
3. **Independence of error terms**: For any two observations the residual terms should be uncorrelated or independent.

4. **Normality of the error term distribution**: The residuals in the model should be random and normally distributed variables with mean 0.

Fitting of the regression model

The fitting of multiple regression model can be assessed by the *Coefficient of Multiple determination*, which is a fraction that represents the proportion of total variation of *y* that is explained by the regression plane.

Sum of squares due to error

$SSE = \sum(y_i - \hat{y}_i)^2$

Sum of squares due to regression

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

Total sum of squares

$SST = \sum(y_i - \bar{y})^2$

Obviously,

$SST = SSR + SSE$

Coefficient of multiple determination (R^2) = $\frac{SSR}{SST}$. It represents the proportion of the total variation in *y* explained by the regression model.

4. Adjusted R -square

The adjusted R-squared is a modified version of R-squared that has been adjusted for the number of independent (p) in the model. The adjusted R-squared increases only if the new term

improves the model more than it would be expected by chance. It is always lower than the R-square.

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

$$= 1 - \frac{SSE/(n-p-1)}{SST/(n-1)}$$

5. Interpreting regression coefficient

In regression with a single independent variable, the coefficient tells how much the dependent variable is expected to increase (if the coefficient is positive) or decrease (if the coefficient is negative) when that independent variable increases by one unit. In regression with multiple independent variables, the coefficient tells us how much the dependent variable is expected to increase when that independent variable increases by one, holding all the other independent variables constant. While interpreting regression coefficients, type of variable should be considered i.e. whether it is metric or non-metric variable.

6. Confidence Interval of regression coefficient

We can define 100(1 - α) % confidence interval for regression coefficient b_j as

$b_j \pm t_{n-p-1,\frac{\alpha}{2}} * SE(b_j)$

7. Statistical inferences for the model

The overall goodness of fit of the regression model (*i.e.* whether the regression model is at all helpful in predicting the values

of *y*) can be evaluated, using *F*-test in the format of analysis of variance.

Under the null hypothesis: H_o: $b_1 = b_2 = \ldots = b_p = 0$, the statistic

$$\frac{SSR/p}{SSE/(n-p-1)} = \frac{MSR}{MSE}$$

has an *F*-distribution with *p* and *n*-1 degrees of freedom

ANOVA Table for Multiple Regression

Source of Variation	*Sum of Squares*	*Degrees of freedom*	*Mean Squares*	*F ratio*
Regression	*SSR*	*p*	*MSR*	*MSR/MSE*
Error	*SSE*	(*n*-*p*-1)	*MSE*	
Total	*SST*	(*n*-1)		

Whether a particular variable contributes significantly to the regression equation can be tested as follows:

For any specific variable x_i, we can test the null hypothesis

H_o: $b_i = 0$, by computing the statistic

$$t = \frac{b_i - 0}{SE(b_i)}$$

and performing a one or two tailed *t*-test with *n*-*p*-1 degrees of freedom.

8. Selection methods

The foundation of a multiple linear regression is to evaluate whether one dependent variable can be predicted from a set of independent (or predictor) variables. Certain regression selection approaches are helpful in testing predictors, thereby increasing the efficiency of analysis.

Three selection procedures are used to yield the most appropriate regression equation: forward selection, backward elimination, and stepwise selection.

- Forward selection begins with an empty equation. Predictors are added one at a time beginning with the predictor with the highest correlation with the dependent variable.
- Backward elimination (or backward deletion) is the reverse process. All the independent variables are entered into the equation first and each one is deleted one at a time if they do not contribute to the regression equation.
- Stepwise selection is considered a variation of previous two methods. Stepwise selection involves analysis at each step to determine the contribution of the predictor variable entered previously in the equation. In this way it is possible to understand the contribution of the previous variables now that another variable has been added. Variables can be retained or deleted based on their statistical contribution.

Essentially, the multiple regression selection process enables the researcher to obtain a reduced set of variables from a larger set of predictors, eliminating unnecessary predictors, simplifying data, and enhancing predictive accuracy.

9. Examples with Solution

Example: In survey of 20 flats of a particular locality, information of three variables: flat size, assessed value and selling price were collected. Fit a regression model.

Flat size	Assessed value	Selling Price
1531	5730000	7480000
1520	6380000	7400000
1625	6540000	7290000
1433	5700000	7000000
1457	6380000	7490000
1733	6320000	7600000
1448	6020000	7200000
1491	5770000	7350000
1525	5640000	7450000
1389	5560000	7350000
1518	6260000	7150000
1444	6340000	7100000
1487	6020000	7890000
1863	6720000	8650000
1520	5710000	6800000
2576	8960000	10200000
1905	6860000	8400000
1537	6010000	6900000
1806	6630000	8800000
1635	6580000	7600000

Solution:

This data was analyzed using SPSS software and following results were obtained.

Table1: Variables Entered/ Removed

Model	Variables Entered	Variables Removed	Method

1	Assessed value, Flat size[b]	.	Enter

Table 1 shows how many variables are included in the model (equation) and it also shows which selection procedure is used for fitting the model. Here ***Enter*** method is used for fitting.

Table 2: Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.913	.834	.815	347253.887

Table 2 gives the value of R square which is 0.834, indicating that the data exhibit strong regression relationship. The value of adjusted R Square is 0.815.

Table 3: ANOVA

Model		Sum of Squares	Df	Mean Square	F	Sig.
1	Regression	10328750550521.172	2	5164375275260.586	42.828	.000
	Residual	2049949449478.829	17	120585261734.049		
	Total	12378700000000.000	19			

Table 3 gives the ANOVA analysis which provides the statistical tests for the overall model fit in terms of the F ratio. Here F ratio is 42.828 and significance level is 0.000 which indicates that the model is a good fit.

Table 4: Coefficients

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	3096656.634	788220.844		3.929	.001
	Flat size	2634.400	785.599	.875	3.353	.004
	Assessed value	.045	.285	.041	.158	.876

Using Table 4 we get the fitted equation:

Selling Price =

3096656.634+2634.4(**Flat size**) +0.045(**Assessed value**)

Here 2634.4 is the regression coefficient of flat size which indicates that with every increase of square feet in the flat size, the selling price increases by 2634.40 Rs.

0.045 is the regression coefficient of Assessed value which indicates that with every 1 Rupee increase in assessed value, the selling price increases by 0.045 Rs.

Table 4 also gives the t values of t test along with the significance value. The significance value in the last column suggests that the variable assessed value is not contributing significantly to the model so it can be dropped.

95 % Confidence interval for regression coefficient of flat size is given by

$$b_j \pm t_{n-p-1,\frac{\alpha}{2}} * SE(b_j)$$

$= 2634.4 \pm t_{17} * 785.60$

$= 2634.4 \pm 2.11 * 785.60$

$= 2634.4 \pm 1657.62$

$= (976.78, 4292.02)$

Similarly confidence interval for other independent variable assessed value can also be obtained.

10. Summary

- Multiple regression is a 'linear combination' of multiple variables to best predict an outcome, and then we assess the contribution that each predictor variable makes to the equation.
- The assumptions to be examined are in four areas:
  - Linearity

- Constant variance of error term
- Independence of error terms
- Normality of the error term distribution

- The **coefficient of multiple determinations is the r**atio of *SSR* and *SST* represents the proportion of the total variation in *y* explained by the regression model. This ratio is denoted by R^2.

- The **adjusted R-squared** is a modified version of R-squared that has been adjusted for the number of predictors in the model.
- 100(1 - α) **confidence interval** for regression coefficient b_j is given as follows:

$$b_j \pm t_{n-p-1,\frac{\alpha}{2}} * SE(b_j)$$

- Three **selection procedures** are used to yield the most appropriate regression equation: forward selection, backward elimination, and stepwise selection.