UNIT-3

Multiple Linear Regression & Factor Analysis

Multiple Linear Regression Analysis - Inferences from the estimated regression function - Validation of the model - Approaches to factor analysis - interpretation of results.

Multiple Regression Analysis:

Multiple regression analysis is a statistical technique that can be used to analyze the relationship between single dependent variable and several independent variables.

To apply multiple regression analysis:

(1) The data must be metric or appropriately transformed, and

(2) before deriving the regression equation the researcher must decide which variable is to be dependent and which remaining variables will be independent.

Example:

Study of eight families regarding their credit card usage.

Dependent Variable (Y)

- Number of credit cards used.

Independent variables (V_1, V_2 and V_3)

- Family size, family income and number of automobiles owned.

→ Impact of Multicollinearity:

The ability of an additional independent variable to improve the prediction of the dependent variable is related not only on its correlation to the dependent variable, but also to the correlation(s) of the additional independent variable to the independent variable(s) already in regression equation.

- Collinearity is the association, measured as the correlation between two independent variables.

- Multicollinearity refers to the correlation among three or more independent variables.

* The impact of multicollinearity is to reduce any single independent variable's predictive power by the extent to which it is associated with the other independent variables.

* To maximize the prediction from a given number of independent variables, the researcher should look for independent variables that have low multicollinearity with the other independent variables but also have high correlations with the dependent variable.

→ Multiple Regression Equation:

To improve further our prediction of credit card holdings, let us use additional data obtained from our eight families.

Simple Regression

$$Y = b_0 + b_1 V_1$$

The second independent variable to include in the regression model is family income (V_2) which has next highest correlation with the dependent variable.

We simply expand our simple regression model to include two independent variables as follows.

Predicted number of credit cards used

$$= b_0 + b_1 V_1 + b_2 V_2 + e$$

where

b_0 - constant number of credit cards independent of family size and income.

b_1 - change in credit card usage associated with unit change in family size.

b_2 - change in credit card usage associated with unit change in family income.

V_1 - family size

V_2 - family income

e - prediction error

Decision process for Multiple Regression Analysis:

Stage 1: Objectives of Multiple Regression

The researcher must consider three primary issues:

1. The appropriateness of the research problem.
2. Specification of statistical relationship
3. Selection of the dependent & independent variables.

Stage 2: Research Design of a multiple Regression Analysis

The researcher incorporates three feature

(1) Sample size

Multiple regression maintains the necessary levels of statistical power and practical significance across a broad range of sample sizes.

(2) Unique elements of the dependence relationshi

Even though independent variables are assumed to be metric and have a linear relationship with the dependent variable, both assumptions can be relaxed by creating additional variables to represent these special aspects of the relationship.

(3) Nature of the independent variables:

Multiple regression accommodates metric independent variables that are assumed to be fixed in nature as well as those with rando component

Stage 3: Assumptions in multiple Regression Analysis:

The assumption to be examined are in four areas:

1. Linearity of the phenomenon measured
2. Constant variance of the error terms.
3. Independence of the error terms
4. Normality of the error term distribution

→ Assessing Individual Variable Vs the Variate

In multiple regression, once the variate is derived, it acts collectively in predicting the dependent variable, which necessitates assessing the assumptions not only for individual variable but also for the variate itself.

→ Methods of Diagnosis:

- The principal measure of prediction error for the variate is the residual- the difference between the observed and predicted values for the dependent variable.
- The most widely used is the studentized residual, whose values corresponds to t values.
- plotting the residuals versus the independent or predicted variables is a basic method of identifying assumption violations for the overall relationship.

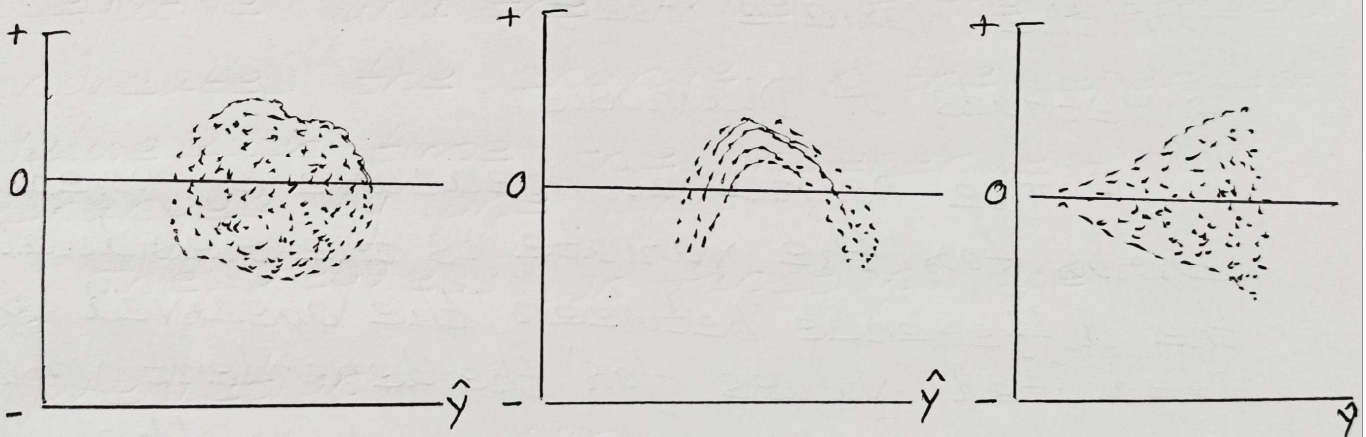
→ Linearity of the phenomenon

The linearity between dependent and independent variables represents the degree to which the change in the dependent variable is associated with the independent variable.

→ Constant variance of the error term:

The prescence of unequal variances (heteroscedasticity) is one of the most common assumption violations.

- plotting the residuals (studentized) against the predicted dependent values and comparing them to the null plot shows a consistent pattern if the variance is not constant.



(a) Null plot (b) Nonlinearity (c) Heteroscedasticity

→ Normality of the Error Term Distribution:

The simplest diagnostic for the set of independent variables in the equation is a histogram of residuals, with a visual check for a distribution approximating the normal distribution.

Stage 4: Estimating the regression model and Assessing overall model fit

→ Selecting an Estimation Technique:

The researcher use the estimation technique to pick and choose among the set of independent variables with either sequential search methods or combinatorial process.

Three approaches to specifying the regression model are

1. confirmatory specification:

The approach for specifying the regression model is to employ a confirmatory perspective wherein the researcher specifies the exact set of independent variables to be included.

2. sequential search methods:

This method provides an objective method for selecting variables that maximizes the prediction while employing the smallest number of variables.

Two types are

- Stepwise estimation
- Forward addition & backward elimination.

3. Combinatorial Approach:

It is a generalized search process across all possible combination of independent variables.

- The best known procedure is all-poss. Subsets regression.

↳ All possible combinations of the independent variables are examined, an the best-fitting set of variables is identif.

Stage 5: Interpreting the regression variate:

→ Using Regression coefficients:

The estimated regression coefficients, termed b coefficients represent both the type of relationship between independent and dependent variables in the regression variate.

- The sign of the coefficients denotes whether the relationship is positive or negative, and the value of the coefficients indicates the change in the dependent value each time the independent variable changes by one unit.

The regression functions are

1. Prediction
2. Estimation
3. Forecasting

Factor Analysis:

Factor analysis is a sophisticated statistical method aimed at reducing a large number of variables into a smaller set of factors.

This technique is valuable for extracting the maximum common variance from all variables, transformi them into single score for further analysis.

Approaches

Factor analysis is a part of Gener Linear Model (GLM) and this method also assumes several assumptions: there is linear relationship, there is no multicolinearity, it includes relevant variables into analysis and there is true correlation between variables and factors.

Approaches to factor Analysis:

1. principal component Analysis (PCA)

- PCA starts extracting the maximum variance and puts them into the first factor.

- After that it removes the variance explained by the first factors and other then starts extracting maximum variance for the second factor.
- This process goes to the last factor.

2. Common Factor Analysis:

- It extracts the common variance and puts them into factors.
- This method does not include the unique variance of all variables.
- This method is used in structural Equation Modelling (SEM)

3. Image factoring:

- This method is based on correlation matrix.
- Image factoring uses ordinary least squares regression to predict factors, making it distinct in its approach to factor extraction.

4. Maximum Likelihood Method:

This technique utilizes the maximum likelihood estimation approach to factor analysis, working from the correlation matrix to derive factors.

other methods of factor analysis:

→ Factor loading

Basically it is the correlation coefficient for the factors & variables. Also, it explains the variable on a particular factor shown by variance.

→ Eigen values

It explains the variance shown by that particular factor out of the total variance.

→ Factor score:

It is the score of all rows and columns that we can use as an index for all variables and for further analysis. moreover we can standardize it by multiplying it with a common term.

→ Rotation method

The five rotation methods are

- No rotation method
- Varimax rotation method
- Quartimax Rotation method
- Direct oblimin rotation method
- Promax rotation method

Interpretation of Results:

1. Determine the number of factors:

The various methods to determine the number of factors are

(a) % Var

- Use the percentage variance to determine the amount of variance that the factors explain.
- The acceptable level depends on the type of application

(b) Variance (Eigenvalues)

If we use principal components to extract factors, the variance equals the eigenvalue.

- The size of the eigen value is used to determine the number of factors

(c) Screen plot:

The screen plot orders the eigen values from largest to smallest.

- The ideal pattern is a steep curve, followed by a bend, and then a straight line.

2. Interpret the factors:

- After determine the number of factors, we can repeat the analysis using the maximum likelihood method.
- Then examine the loading pattern to determine the factor that has the most influence on each variable.
- Loadings close to -1 or 1 indicate that the factor has a ~~weak~~ strong influence on the variable.
- Loading close to 0 indicate that the factor has a weak influence on the variable.
- unrotated factor loadings are often difficult to interpret.
- Factor rotation simplifies the loading structure, allowing you to more easily interpret the factor loadings.
- However, one method of rotation may not work best in all cases. You want to try different rotations and use the one that produces the most interpretable results.

3. check your data from problems

- If the first two factors account for most of the variance in the data, we can use score plot to assess the

data structure and detect clusters, outliers and trends.

- Groupings of data on the plot may indicate two or more seperate distributions in the data.
- If the data follow a normal distribution and no outliers are present, the points are randomly distributed about the value of 0.

References:

1. Joseph F Hair, Rolph E Anderson, Ronald L. Tatham & William C. Black, Multivariate Data Analysis, pearson Education, New Delhi, 2005.

30/9/24

Course Coordinator

(Vidhya S)

3/10/24

HOD

(P.M. SIVA RAJA)