Regression analysis

1. Under what conditions would you use correlation and/or regression analysis Include comments on the type of data needed, the types of questions that these tools can answer, and a work-related suggestion for their use. Consider both linear and multiple correlation and regression. Answer.

The multiple regression was used in the context of comparing and estimating the effects of two variables of which one is independent and the other dependent on first one. eg; The increase of population and the increase in demand of land and housing. In the prescribed case the increase of population is taken as independent and the demand which depends on it is taken as a dependent variable. The change in the dependent value for a variation in the independent value is estimated in the regression analysis. Multiple regression takes into consideration all the assumptions of correlation. It takes place when the independent variable is dichotomy. In the above prescribed case, if the increase of men and women were to be considered separately; Multiple regression is used. In the case of linear correlation no power terms are found as it will not reflect curvilinear changes in independent variables. In the context of multiple regression the powers to the variables were found to represent the curvilinear variations in independent and dependent variables. Correlation is the percent of variance in the dependent explained by the given independent when all other independents are allowed to vary. In the final result the magnitude of r2 reflects not only the unique covariance it shares with the dependent, but uncontrolled effects on the dependent attributable to covariance the given independent shares with other independents in the model. For example in the above case the increase of male female population can be taken as

covariance.

2. During the years 1790 to 1820, the correlation between the number of churches built in New England and the barrels of Rum imported into the region was a perfect 1. 0. What does this tell you - that church building causes rum drinking, that rum drinking causes church building, or something else If something else, what

Answer:

During the years 1790 and 1820 the correlation between the church building and rum drinking was found as one. This does not mean that church building causes rum drinking. If statistically considered, the church building and rum drinking were both independent, i. e. each of them is not dependent on the other which is considered in the context of calculating correlation. The condition that the rum drinking must be dependent on the church building (in this case) was not followed. So there is a common case which causes both. If that cause is abundant availability of money, then the church building and rum drinking can be considered as dependents to abundant availability of money and two correlations can be calculated. That work and the result can give the realistic results regarding the statistical calculations.

3. Political science question (from " How to Think About Statistics" 5th edition by John L. Phillips, Jr. W. H. Freeman and Company, New York 1996): Researchers have frequently asked whether there is any relationship between the amount of domestic conflict within a given country (X) and the amount of foreign conflict which that country initiates (Y). Assume you have constructed a conflict scale and collected data for 50 countries on the values of both X and Y. What statistic will answer your question

Answer:

If the correlation obtained is 1 for calculation between the domestic conflict in a particular country and the foreign conflict it initiates, then it can be told that the domestic conflict in a country is a cause for foreign conflict. If this is < . 5 then its effect in causing foreign is negligible. The statistics of the fifty countries obtained on the basis of domestic conflicts initiating foreign conflicts can be taken on a scale of 1 to 10, in which, 1 means the lowest and the 10 means the highest in causing the conflict. A country placed at 1 on the domestic conflict scale is initiating a foreign conflict of 10, then these two variables have no effect on one another. If the level of foreign conflict caused by a country on the scale of conflict is same as the level of domestic conflict in that country, then the correlation between them will be nearer to 1 and it will be established that domestic conflict in the country is the cause for initiating of foreign conflict by it.

4. In the mid 60's, the Department of Education had a study performed on educational achievements of students. The researcher entered the variables into the equation according to a time-based theory of the impact of variables. This means he entered the variables in the order a person would run across them in real time. So, the first two variables entered were race and sex. The study concluded that the educational system discriminated among students on the basis of race. As a student of statistics, what comments might you make about the study's entering of variables

Answer:

In contrast to the above case another study of discrimination caused on the basis of race and sex can be taken. Let a researcher had entered the variables; race and sex of the students to study the achievements of them, in the form a time based theory of impact variables. Let us say this study

revealed that, the students of a particular and sex were found of achieving more and other races and sex were found achieving less. This can be regarded as discrimination on that particular race and sex. A student of statistics can justify the entering of variables of race and sex in the study regarding admission in to the institution but not on the results of it. If the variables were entered in the study of results, then the result obtained is not the discrimination of the education system but other factors of that race and sex were dependent upon. So an independent variable which decides the course of variance in race and sex was neglected. So while entering the race and sex as the variables for a study of achievement in education, they must be considered as dependents and other factors which can be considered as independent variables that affect the various races and sexes also must be taken into consideration and should be entered as independent variables.

5. One issue that many companies are now facing is in determining what the best production practices are for their products. This often involves examining not only the quality and specs of incoming raw materials but also the process variables such as how long to heat something, what temperature to use, etc. If you were charged with maximizing the effectiveness of a manufacturing process, how might you go about the task (Assume you have all of the needed measurements on the different variables involved in the process.)

Answer:

When the dependent variable is a dichotomy like in the case of variables of manufacturing process, assumptions of multiple regression cannot be met; discriminant analysis or logistic regression is used instead. Partial least squares regression is sometimes used to predict one set of response

variables from a set of independent variables, which here mean the variables involved in the manufacturing process. Logit regression uses log-linear techniques to predict one or more categorical dependent variables. Poisson regression is a form of log-linear analysis common in event history analysis and other research involving rare events where assumptions of a normally distributed dependent do not apply. Categorical regression is a variant which can handle nominal independent variables.

6. What is the primary purpose of residual analysis

Answer:

While considering only variables some times the results were not accurate. Residuals are the difference between the observed values and those predicted by the regression equation. (subtract mean, divide by standard deviation). The purpose here is to show how residual analysis can not only perform its usual statistical purpose (i. e. check model adequacy) but also serve as a final check for glitches. Besides, the authors share a graphical display of residual that retains in a single display, the information in plots of residuals versus all of the following: predicted response, actual response, and level of the independent variables, and effects, which helps in visualizing how residuals fit into experimental results. There was concern that despite extreme care in executing and measuring the process they could account for only 82% of the variation. A graph with residuals plotted as straight line segments between the observed response and the response predicted by a model shows large unexplained variation in some points. A rerun of the analysis, removing the data points results in an R^2 of . 998, the residuals then will show a random pattern. The residual pattern makes clear; the negative residuals occurred at the middle values for feed where the planar

regression surface lay above the curved surface. The positive residuals occurred where the regression plane lay below the curved surface at the extreme values for feed. Rerunning the analysis with the quadratic terms includes will show a strong quadratic relation between finish and feed. R^2 increased to 0. 999, and the new residual pattern seems to be random. Thus use of residuals helps to illustrate the importance of keeping accurate records of the process: The residual plot shows a very large residual for observation where a misfed or a maximizing effect occurred.

References:

helios. bto. ed. ac. uk/bto/statistics/tress11. html

abyss. uoregon. edu/js/glossary/correlation. html