Review Article

A Brief Introduction to Regression Analysis and Its Types

ABSTRACT

This article may be used for research, teaching, and private study purposes. In this study, I focused on regression analysis and with a special focus on types of regression analysis and some types of regression. The term regression is used to indicate the estimation or prediction of the average value of one variable for a specified value of another variable. And Regression Analysis is a statistical tool used to estimate the relationship between a dependent variable and an independent variable. For example, if a Manger of a firm wants to the exact relationship between advertisement expenditure and sales for future planning then the regression technique will be most suitable for him.

Key Words: *Regression, Types of Regression*

INTRODUCTION

Regression analysis is a statistical tool for the investigation of relationships between variables. Usually, the investigator seeks to ascertain the causal effect of one variable upon another—the effect of a price increase upon demand, for example, or the effect of changes in the money supply upon the inflation rate (Sykes, 1993). To explore such issues, the investigator assembles data on the underlying variables of interest and employs regression to estimate the quantitative effect of the causal variables upon the variable that they influence. The investigator also typically assesses the "statistical significance" of the estimated relationships, that is, the degree of confidence that the true relationship is close to the estimated relationship (Cook et al, 1999).

WHAT IS REGRESSION?

For purposes of illustration, suppose that we wish to identify and quantify the factors that determine earnings in the labor market. A moment's reflection suggests a myriad of factors that are associated with variations in earnings across individuals—occupation, age, experience, educational attainment, motivation, and innate ability come to mind, perhaps along with factors such as race and gender that can be of particular concern to lawyers. For the time being, let us restrict attention to a single factor—call it education. Regression analysis with a single explanatory variable is termed "simple regression."(Sykes, 1993).

RESULT AND DISCUSSION

There are different types of regression analysis, let's talk about it in more details:

1. LINEAR REGRESSION

Linear regression is a type of model where the relationship between an independent variable and a dependent variable is assumed to be linear. The estimate of variable "y" is obtained from an equation, y'-y_bar = byx(x-x_bar)…… (1) and estimate of variable "x" is obtained through the equation x'-x_bar = bxy(y-y_bar)….. (2). the graphical representation of linear equations on (1) & (2) is known as Regression lines. These lines are obtained through the Method of Least Squares (Draper et al, 1998).

UNDER PEER REVIEW

There are two kinds of Linear Regression Model:

A. Simple Regression

In reality, any effort to quantify the effects of education upon earnings without careful attention to the other factors that affect earnings could create serious statistical difficulties (termed "omitted variables bias"), which I will discuss later. But for now, let us assume away this problem. We also assume, again quite unrealistically, that "education" can be measured by a single attribute—years of schooling. We thus suppress the fact that a given number of years in school may represent widely varying academic programs.

At the outset of any regression study, one formulates some hypothesis about the relationship between the variables of interest, here, education and earnings. Common experience suggests that better-educated people tend to make more money. It further suggests that the causal relation likely runs from education to earnings rather than the other way around. Thus, the tentative hypothesis is that higher levels of education cause higher levels of earnings, other things being equal.

To investigate this hypothesis, imagine that we gather data on education and earnings for various individuals. Let E denote education in years of schooling for each individual, and let I denote that individual's earnings in dollars per year. We can plot this information for all of the individuals in the sample using a two-dimensional diagram, conventionally termed a "scatter" diagram. Each point in the diagram represents an individual in the sample (Fox, J., 1997).

Then, the hypothesized relationship between education and earnings may be written

$$I = a + bE + e$$

Where

a = a constant amount (what one earns with zero education);

b = the effect in dollars of an additional year of schooling on income,

Hypothesized to be positive; and

e = the "noise" term reflecting other factors that influence earnings.

B. Multiple Regression

Earnings are affected by a variety of factors in addition to years of schooling, factors that were aggregated into the noise term in the simple regression model above. "Multiple regression" is a technique that allows additional factors to enter the analysis separately so that the effect of each can be estimated. It is valuable for quantifying the impact of various simultaneous influences upon a single dependent variable. Further, because of omitted variables bias with simple regression, multiple regression is often essential even when the investigator is only interested in the effects of one of the independent variables.

For purposes of illustration, consider the introduction into the earnings analysis of a second independent variable called "experience." Holding constant the level of education, we would expect someone who has been working for a longer time to earn more. Let X denote years of experience in the labor force and, as in the case of education, we will assume that it has a linear effect upon earnings that is stable across individuals (Hamilton, L. C. 1992). The modified model may be written:

$$I = a + bE + \gamma X + e$$

Where γ is expected to be positive.

UNDER PEER REVIEW

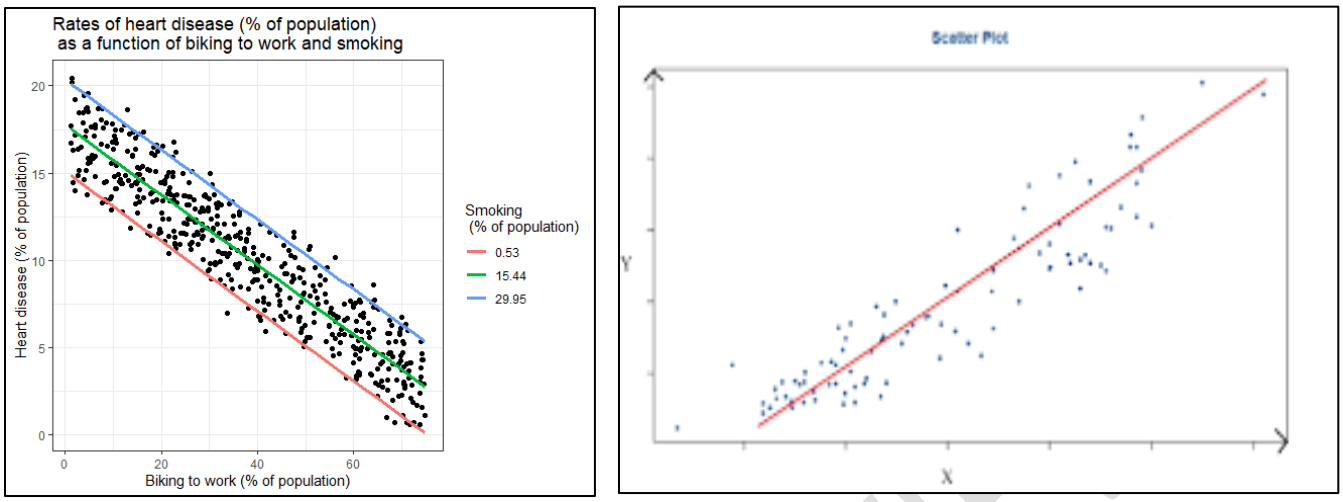

Figure 1. Graphical Representation of Linear Regression, A. multiple regression, and B. Simple regression.

2. POLYNOMIAL REGRESSION

It is a type of Regression analysis that models the relationship of values of the Dependent variable "x" and Independent variables "y" as non-linear. It is a special case of Multiple Linear Regression even though it fits a non-linear model to data. It is because data may be correlated but the relationship between two variables might not look linear (Montgomery, D. C., 2001). Generally takes the form y = m x + b where m is the slope, and b is the y-intercept. It could just as easily be written f(x) = c0 + c1 x with c1 being the slope and c0 the y-intercept

3. LOGISTIC REGRESSION

Logistic Regression is a method that was used first in the field of Biology in the 20th century. It is used to estimate the probability of certain events that are mutually exclusive, for example, happy/sad, normal/abnormal, or pass/fail. The value of probability strictly ranges between 0 and 1 (Ahmadzai, W. K., & Aryobi, H. G. 2021). This is the equation used in Logistic Regression. Here (p/1-p) is the odd ratio.

4. QUANTILE REGRESSION

Quantile Regression is an econometric technique that is used when the necessary conditions to use Linear Regression are not duly met. It is an extension of Linear Regression analysis i.e., we can use it when outliers are present in data as its estimates strong against outliers as compared to linear regression (Weisberg, 1985),

5. RIDGE REGRESSION

To understand Ridge Regression we first need to get through the concept of Regularization. Regularization: There are two types of Regularization, L1 regularization & L2 regularization. L1 regularization adds an L1 penalty equal to the value of coefficients to restrict the size of coefficients, which leads to the removal of some coefficients. On the other hand, L2 regularization adds a penalty L2 which is equal to the square of coefficients. Using the above method Regularization solves the problem of a scenario where the model performs well on training data but underperforms on validation data (Walker, E., 2002).

6. LASSO REGRESSION

LASSO (Least Absolute Shrinkage and Selection Operator) is a regression technique that was introduced first in geophysics. The term "Lasso" was coined by Professor Robert Tibshirani. Just like Ridge

UNDER PEER REVIEW

Regression, it uses regularization to estimate the results. Plus it also uses variable selection to make the model more efficient (Davis, C. S. 2002).

7. ELASTIC NET REGRESSION

Elastic net regression is favored over ridge and lasso regression when one has to deal with exceedingly correlated independent variables (Neter, J. et al, 1996).

8. PRINCIPLE COMPONENTS REGRESSION (PCR)

Principle components regression technique is broadly used when one has various independent variables. The technique is used for assuming the unknown regression coefficient in a standard linear regression model. The technique is divided into two steps,

1. Obtaining the principal components

2. Go through the regression Analysis on Principle components.

9. PARTIAL LEAST REGRESSION (PCR)

It is a substitute technique of principal components regression when one has a widely correlated independent variable. The technique is helpful when one has many independent variables. Partial least regression is widely used in the chemical, drug, food, and plastic industry (Draper, N. R., & Smith, H. 1998).

10. SUPPORT VECTOR REGRESSION

Support vector regression can be used to solve both linear and nonlinear models. Support vector regression has been determined to be productive to be an effective real-value function estimation.

11. ORDINAL REGRESSION

Ordinal regression is used to foreshow ranked values. The technique is useful when the dependent variable is ordinal. Two examples of Ordinal regression are Ordered Logit and ordered probity (Fahrmeir, L., et al, 2007).

12. POISSON REGRESSION

Poisson Regression is used to foreshow the number of calls related to a particular product on customer care. Poisson regression is used when the dependent variable has a calculation. Poisson regression is also known as the log-linear model when it is used to model contingency tablets. Its dependent variable y has Poisson distribution (Ryan, T. P., 2008). Thus, the module for the Poisson regression model: log (λi) =β0+β1xi l o g (λ i) = β 0 + β 1 x i where the observed values Yi∼ Y i ∼ Poisson with λ=λi λ = λ i for a given xi.

13. NEGATIVE BINOMIAL REGRESSION

Similar to Poisson regression, negative Binomial regression also accord with count data, the only difference is that the Negative Binomial regression does not predict the distribution of count that has variance equal to its mean (Weisberg, S. 2005).

14. QUASI POISSON REGRESSION

Quasi Poisson Regression is a substitute for negative Binomial regression. The technique can be used for over dispersed count data (Wright, R. E. 1995).

UNDER PEER REVIEW

15. COX REGRESSION

Cox Regression is useful for obtaining time-to-event data. It shows the effect of variables on time for a specific period. Cox Regression is also known as proportional Hazards Regression (Seber, G. A., et al, 2012).

16. TOBIT REGRESSION

Tobit Regression is used to evaluate linear relationships between variables when censoring (observing independent variable for all observation) exists in the dependent variable. The value of the dependent is reported as a single value.

CONCLUSION

The types of regression analysis are listed above but choosing a correct regression model is a tough grind. It requires vast knowledge about statistical tools and their application. The correct method was chosen based on the nature of the variable, data, and the model itself. Overall the different types of Regression Analysis have calculated discrete and distinct data very easily in recent, not only in the field of mathematics/statistics but it has many applications in the real world as well. Hence, Regression analysis is a boon for mankind.

REFERENCES

1. Sykes, A. O. (1993). An introduction to regression analysis.
2. Cook, R. D., and Weisberg, S. (1999), Applied Regression Including Computing and Graphics, New York: Wiley.
3. Draper, N. R., and Smith, H. (1998), Applied Regression Analysis (3rd ed.), New York: Wiley.
4. Fox, J. (1997), Applied Regression Analysis, Linear Models, and Related Methods, Thousand Oaks, CA: Sage Publications.
5. Hamilton, L. C. (1992), Regression with Graphics: A Second Course in Applied Statistics, Belmont, CA: Wadsworth.
6. Montgomery, D. C., Peck, E. A., and Vining, G. (2001), Introduction to Linear Regression Analysis (3rd ed.), Hoboken, NJ: Wiley.
7. Neter, J., Kutner, M. H., Nachtsheim, C. J., and Wasserman,W. (1996), Applied Linear Statistical Models (4th ed.), Boston: McGraw-Hill.
8. Weisberg, S. (1985), Applied Linear Regression (2nd ed.), New York: Wiley.
9. Walker, E., and Wright, S. P. (2002), “Comparing Curves Using Additive Models,” Journal of Quality Technology, 34, 118–129.
10. Davis, C. S. (2002), Statistical Methods for the Analysis of Repeated Measurements, New York: Springer-Verlag.
11. Ahmadzai, W. K., & Aryobi, H. G. (2021). Natural and Socio-economics Factors Affecting the Household Food Security in Rural Area of Paktia Province, Afghanistan. *Asian Journal of Agricultural Extension, Economics & Sociology*, 1-11.
12. Draper, N. R., & Smith, H. (1998). *Applied regression analysis* (Vol. 326). John Wiley & Sons
13. Fahrmeir, L., Kneib, T., Lang, S., & Marx, B. (2007). *Regression*. Springer-Verlag Berlin Heidelberg.
14. Ryan, T. P. (2008). *Modern regression methods* (Vol. 655). John Wiley & Sons.
15. Weisberg, S. (2005). *Applied linear regression* (Vol. 528). John Wiley & Sons.
16. Wright, R. E. (1995). Logistic regression.
17. Seber, G. A., & Lee, A. J. (2012). *Linear regression analysis* (Vol. 329). John Wiley & Sons.

UNDER PEER REVIEW