INTERNATIONAL JOURNAL OF TRENDS IN EMERGING RESEARCH AND DEVELOPMENT

Volume 2; Issue 4; 2024; Page No. 202-206

Received: 17-04-2024
Accepted: 30-06-2024

To study the impact of regression analysis methods for robust business forecasting in marketing management

Dr. Manish Kumar Srivastava

Corresponding Author: Dr. Manish Kumar Srivastava

Abstract

Business prediction findings are vital for assessing a company's future financial success in the context of contemporary business practices. Procedures for planning and prediction are particularly crucial for businesses that operate in an uncertain environment. This study provides an illustration of how to plan and forecast business outcomes in the insurance industry when using linear and nonlinear regression to calculate premium trends. It is essential to obtain sufficient assets to cover the risks because of the uncertainty around the claim's incidence and amount. Predicting future premium movements for individual insurance lines is necessary for asset-liability matching, which is the fundamental idea behind the growth and functioning of insurance businesses. This study examines how regression models can help create strategic financial insights and examines the use of regression analysis in financial forecasting, particularly for small enterprises. Small firms find it difficult to use advanced forecasting techniques because they have limited access to large datasets and financial models. To forecast revenue, expenses, profitability, and other financial metrics that are crucial for small firms, a variety of regression techniques, such as logistic, multivariate, and linear regression models, can be modified. The strengths, drawbacks, and implementation strategies of regression analysis are further examined in this study, along with case examples illustrating its real-world uses.

Keywords: Revenue, expenses, profitability, forecast, environment

1. Introduction

A popular statistical method for analyzing the relationship between variables in business and finance research is regression analysis. Its importance stems from its capacity to shed light on the intricate interactions between variables that affect different occurrences in these fields. The significance of regression analysis in business and finance research is explained in this study, along with its main uses, advantages, and contributions to decision-making.

Quantifying the relationship between variables is one of regression analysis's main purposes. Making informed judgments in business and finance requires knowing the direction and degree of correlations between variables including stock prices and risk factors, interest rates and investment returns, and sales and advertising spend.

By using patterns in previous data, regression analysis helps researchers predict future events. Regression models, for example, can be used in finance to forecast economic indicators, market movements, and stock prices, which helps investors, legislators, and financial institutions make decisions.

Regression analysis is used by companies and financial organizations to evaluate and efficiently manage risks. Regression models aid in the development of risk mitigation techniques and the optimization of portfolio management by identifying variables that impact risk factors like default rates, loan losses, or market volatility.

Regression analysis is essential for assessing how well different financial instruments and business strategies perform. Organizations may improve efficiency and optimize their plans by examining how elements like pricing strategies, marketing campaigns, or investment portfolios affect performance measures like sales revenue, profitability, or returns on investment.

By differentiating between correlation and causation, regression analysis makes it easier to determine the causal links between variables. Researchers can evaluate the causal influence of initiatives, policies, or outside influences on business and financial outcomes by using methods like instrumental variable regression or difference-in-differences analysis.

2. Scope of research work

The practice of oversimplifying regression models through excessive parameter definition or needless introduction of mathematical terms in regression analysis is the root cause of my research. The inability to standardize regression models leads to users or serious researchers being perplexed about which regression models are best suited for their needs.

3. Problem on hand

The challenge at hand is creating an extended version of linear regression that can handle a number of popular regression models as special instances. Showing that different regression models may be formed or obtained from the general model as specific cases is one of the goals of accomplishing this. The primary benefit of doing so is that many of the general model's optimality features will inevitably apply to its special situations. Simplifying the computational process for fitting the regression model for a particular dataset is the second goal of creating a general regression model. Lastly, it will be feasible to compare several regression model types by combining them into a single generic linear model. When every model is a specific case of the general model, such comparisons are simple. For this reason, it is suggested that all models be represented using a generalized linear form.

4. Objectives of the study

1. To create and suggest a general model for Business Forecasting in Marketing Management
2. To determine which general linear regression model is best suited or most important for use in the future.

5. Results and Data analysis

1. Linear regression: The most basic regression model is simple linear regression. Interestingly, the genesis of simple linear regression helps explain why the Gaussian distribution, or normal distribution, is so crucial to the linear regression model. Let's say that W and Y are two random variables that have a bivariate normal distribution. Their respective means are w and dw, their respective standard deviations are w and dw, and their correlation coefficient is rho. Consequently, W's marginal distribution is N (mean = w, standard deviation = w). The conditional distribution of Y | W = w is a normal distribution with variance = [σ2 (1 – ρ 2)] and mean = [μy – ρ * σy/ σw (w – μw)]. This demonstrates that, given W = w, the conditional expectation of Y is a linear function of the constant w. When W and Y have a bivariate normal distribution, linear regression is linear as regression of Y on W is defined as the conditional expectation of Y | W. Additionally, the residual is independent of W and Y and has a normal distribution with zero mean and variance. When the dependent variable Y and the P independent variables W 1, W 2, …, W P jointly follow the multivariate normal distribution, multiple linear regression is also the obvious choice. If we see that the conditional distribution of Y given W 1 = w1, W 2 = w2, …, W P = wP is a linear function of the variables w1, w2, …, wP, then the regression may be easily derived.

2. Polynomial Regression: The scatterplot of W and Y occasionally suggests that there may not be a linear relationship between W and Y. In these situations, polynomial regression is frequently used. Higher order powers of the predictor variable W are used in polynomial regression models, which have the following structure. In this case, Y = y0 + y1 W + y2 W2 + y3W3 + … + yk Wk + ε.

It represents the polynomial regression of the kth degree. Interestingly, the predictor variable and not the regression coefficients are non-linear in polynomial regression. The polynomial regression model will be identical to the multiple linear regression model if we define W1 = W, W2 = W2, W3 = W3, … W k = Wk. This is the cause of the paucity of multinomial regression literature. Terms like W 1 \ W2, W2 \ W3, W2 \ W2, and so on that include products of predictor variables or their powers are occasionally found in the polynomial regression model. Even so, each of these is given a new predictor variable, making the resulting model a multiple linear regression model. Because the more complicated the polynomial expression, the better the model fits the data, it can be quite tempting to fit highly intricate polynomial regression models to the data. It's crucial to remember that adding too many terms to a polynomial regression expression can result in an overfitting issue.

3. Logistic Regression: The logistic regression model's response variable is binary, meaning it can have only two potential values: 0 and 1, which are sometimes referred to as failure and success, respectively. Since the binary answer variable is bounded and a straight line is unbounded, it is evident that the linear regression model does not appear to be suitable. The challenge of predicting the value of the answer (the goal variable) is reformulated as the problem of predicting the probability of success (p), or the likelihood that the response variable will have a value of 1. Even though the probability p is confined between 0 and 1, a linear function cannot be used to predict it. There is no upper constraint on the odds ratio p / (1-p), which is non-negative. The response variable in the logistic regression model is log (p / (1-p)). As p goes along the unit interval from 0 to 1, it can be observed that log (p / (1-p)) spans the whole real line. A linear function of the regressors can be used to forecast this function because it is real-valued. Thus, the logistic regression expression is provided as follows. y0 + y1 W 1 + y2 W 2 + … + yp W P + ε = log (p/(1-p)).

It's crucial to remember that no predicted variable has a linear impact on the initial response in a logistic regression. Because of this, it is difficult to interpret logistic regression coefficients. Unlike linear regression, the effects of predictor variables are multiplicative rather than additive. Furthermore, the logistic regression model deviates from the homoscedasticity condition. Furthermore, the response variable deviates from the normal distribution due to its binary nature. For the same reason, residuals also deviate from the normal distribution.

4. Quantile regression: The following is how the quantile regression model is different from the linear regression model. Whereas the quantile regression model looks for the specified quantile of the conditional distribution of responses that correspond to given values of the explanatory variable(s), the linear regression model looks for the conditional expectation of the response variable corresponding to given values of the explanatory variables. Stated differently, the quantile regression model transforms the idea of a quantile into a conditional quantile. If the distribution function of a random variable Y is F(y) = P(Y < y), then the inverse distribution function Q(q) = inf {y: F(y) > q} = F-1(q) for 0 < q < 1 defines the qth quantile (Q (q)) of Y. The median, for instance, is Q (1/2). The sample median (Md), minMd $\sum n$ | yi – Md |, is known to minimize the total absolute deviation of sample values surrounding a random sample y1, y2, …, yn. Since quantiles lack the sample mean's ability to have the smallest squared deviations when subtracted from the mean, it should be clear that the quantile regression model is unable to apply the squared error loss function. Rather, in quantile regression, the objective function that is minimized is provided by minz in R {Sum wq(yi – z), where I() is the indicator function and wq(w) = w (q – I(w < 0)). In order to determine the conditional mean as the best way to solve the problem of minimizing squared error loss function, the linear regression model applies the sample mean's property of minimizing the total square of the deviations from the sample observations. Similarly, the quantile regression model generates the conditional quantile function for any given quantile q, 0 < q < 1, and extends the property of the sample quantile that minimizes the total weighted deviations from sample values when the weights are given by the function wq() established above. When the dependent variable's distribution is skewed or the data is heteroscedastic, quantile regression is the recommended method. Additionally, quantile regression is resistant to outliers. Keep in mind that quantile regression's regression coefficients varies significantly from linear regression's. The use of quantile regression is not warranted if this does not occur.

5. Ridge Regression: We will briefly discuss the idea of regularization in regression before defining ridge regression. Overfitting occurs when a model fits training data well but performs poorly on test data. Regularization is a technique to address this issue. Regularization uses the penalty function to manage the objective function. Regularization is helpful when there is multicollinearity between the independent variables and when the sample size is too small in relation to the number of independent variables. The L1 and L2 norms are used as the punishment functions in the two most popular regularization techniques. By requiring that the absolute values of the regression coefficients total up to unity, the L1 norm—also referred to as the absolute norm—restricts the regression coefficients. By requiring that the squares of the regression coefficients total up to unity, the L2 norm—also referred to as the quadratic norm—restricts the regression coefficients. The L2 norm is the regularization technique used in ridge regression. The ridge regression's objective function is to minimize ∑(yi – y0 – y1 w1i – y2 w2i - … - ypwpi)2 + λ ∑ y2. The solution to the normal equations for ridge regression is (W‘‘ W + λ I)-1 W' y Ridge regression was first put out as a solution to the multicollinearity issue. As a result of regularization, assuming that the error terms are regularly distributed loses its significance.

6. LASSO regression: By substituting the L1 norm for the L2 norm used in ridge regression, Lasso regression was put up as an alternative to ridge regression. The abbreviation for Least Absolute Shrinkage and Selection Operator is LASSO. The goal function ∑(yi – y1 w1i – y w2i - … - y wpi) 2 + λ ∑ |y | is minimized. Because all variables are normalized prior to model fitting, LASSO does not regularize the intercept. Consequently, the regression has no intercept since it goes through the origin. Since LASSO regression lacks an explicit mathematical solution, the regression coefficients are determined iteratively with the aid of statistical software. As the name implies, LASSO handles the multicollinearity issue and automatically chooses variables to include in the model. LASSO is superior to ridge regression in this regard. Ridge does, however, have the benefit of being more computationally efficient. A direct comparison between Ridge and LASSO is not possible. Both ought to be fitted to training data, and their performance on test data should be the basis for selection. The model with the best performance on test data ought to be chosen.

7. Elastic Net Regression: When multicollinearity is present and it is unclear whether ridge regression or lasso regression is superior, elastic net regression was also suggested. Because it makes use of both L1 and L2 norms, elastic net regression is a hybrid of ridge and lasso regression. There is no intercept term in the regression model since all the variables are normalized before it is fitted. Elastic net regression's objective function is displayed below.

The formula is ∑(yi – y0 – y1 W 1 – y2 W 2 - … - yp W p)2 + λ1 ∑ y2 +λ2 ∑ |p|. Then, it becomes clear that the assumption that mistakes are normally distributed is not made by elastic net regression.

5.1 Diagnostics of a regression model

Prior to concluding the model, it is necessary to verify whether the assumptions taken into consideration during model fitting are adhered to once the desired regression model has been fitted to the provided data. We can confirm the assumptions with the use of residual plots and residual analysis. Typically, outliers or significant data are found using scaled residuals.

The normality of the error terms is examined using normality plots.

Plotting the cumulative probabilities against the ascending sequence of probabilities is known as the normal probability plot. The points should ideally be on a straight line.

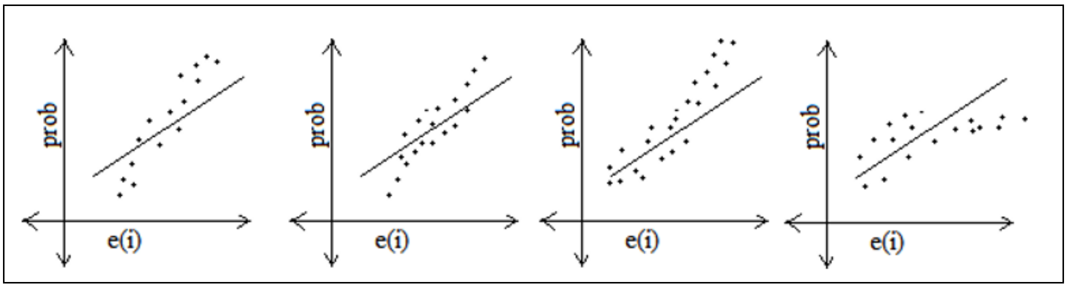

Fig 1: Diagnostics of a regression model

The normal distribution is said to have thicker tails if the points at the ends deviate from the line.
The normal curve has narrower tails at the ends if the points exhibit flat trends at the endpoints. The distribution is favourably skewed if the plot displays an upward trend in the upper portion of the plot.

The distribution is negatively skewed if the points in the upper portion of the plot exhibit a decreasing tendency from the straight line.
The homoscedasticity of the error terms can be checked with the use of the plot of the error term versus the expected values.

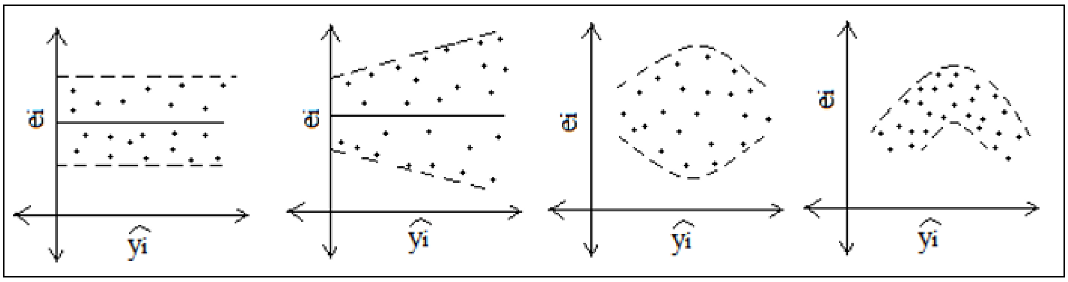

Fig 2: The normal distribution is said to have thicker tails if the points at the ends deviate from the line

The model's suitability is demonstrated by the horizontal plot. Near the expected values, errors are dispersed at random. The premise of homoscedasticity is said to be broken if the plot displays an outward opening funnel-type shape, indicating that error variances are neither totally random nor constant. The linearity assumption is said to be broken if the graphs display a curve.

Other Residual Plots

When used in R for linear regression analysis, these are also known as "Diagnostic Plots." The model's fit to the data is assessed using residual plots. The distribution of residuals is normal. As demonstrated below, a variety of residual plots are used as a straightforward method to verify the regression model's acceptance and normality assumption.

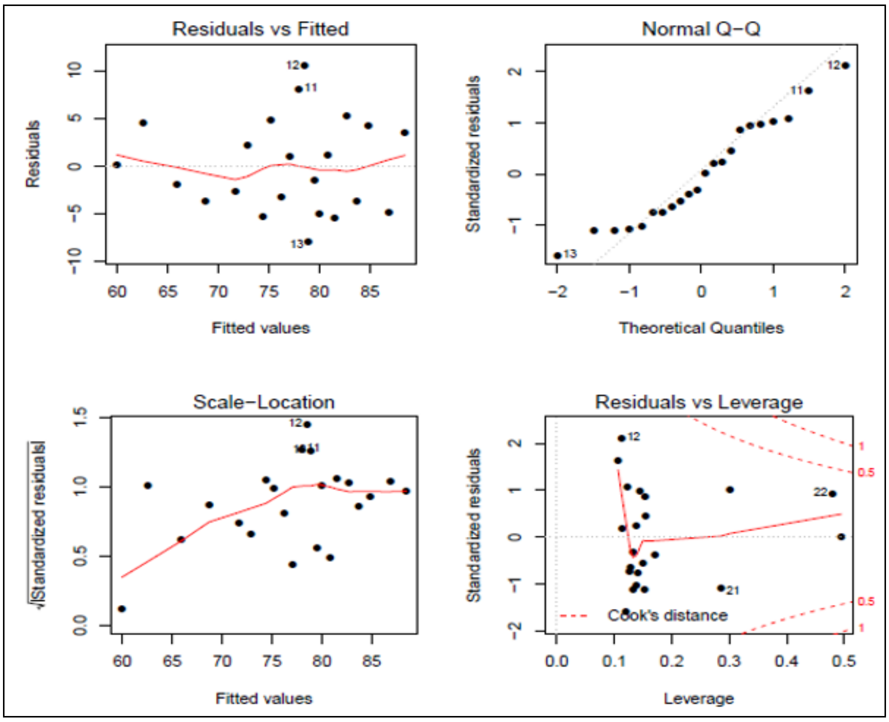

Fig 3: The model's suitability is demonstrated by the horizontal plot

1. We can discuss the assumption of homogeneity of error variances in a normal Q.Q plot if residuals are shown against fitted values. It ought to appear arbitrary.
2. The residuals are shown against cumulative normal probabilities in an ascending manner in a normal probability plot (Normal Q.Q). In an ideal normal probability plot, the points should be on a straight line.
3. Location Scale The plot should appear haphazard and devoid of patterns. It displays the distribution of points over the anticipated value range. Uniform variance throughout the data range is indicated by a red, horizontal line.
4. Cook's distance plot identifies the leverage points-points that have the biggest impact on the regression. Large residuals are linked to these sites. Outside of Cook's distance lines (the red dashed line), in the top or lower right corner of this map, are any significant observations. Excluding certain observations will change the regression findings.

6. Conclusion

In business and finance research, regression analysis plays a key role since it provides a flexible toolkit for relationship analysis, outcome prediction, risk management, and decision support. Forecasting, risk management, performance assessment, causal inference, decision assistance, and strategic planning are just a few of the many fields in which it finds use. Researchers and practitioners can improve performance, promote sustainable growth, and stimulate innovation by using regression analysis to provide important insights into the intricate dynamics of business and finance.

One of the most effective statistical tools in business is regression analysis, which is used to create mathematical models that forecast the value of one variable depending on another. It can be roughly divided into two categories: multiple linear regression, which uses many independent variables to make predictions, and simple linear regression, which uses a single independent variable to predict a dependent variable. Effective decision-making in a variety of commercial scenarios, including sales forecasting, marketing strategy evaluation, and consumer behavior analysis, depends on an understanding of the intricate interactions among variables, which this analysis helps to provide.

Regression analysis, however, is predicated on a number of assumptions, including the accuracy of the model and the caliber of the data, both of which are frequently jeopardized in practical situations. Because the correlations between variables might be complex and influenced by a number of factors, analysts must carefully interpret the results. Furthermore, other regression approaches can be used to meet particular objectives. For example, multivariate regression can be used to examine multiple dependent variables at once, while time series analysis can be used to foresee patterns over time. All things considered, regression analysis offers insightful information that can guide strategic choices and guarantee that companies successfully adjust to changing market conditions.

7. References

1. Angrist JD, Pischke JS. Mostly Harmless Econometrics: An Empiricist's Companion. Princeton: Princeton University Press; c2009.
2. Blanchard O. Macroeconomics. Pearson; c2017.
3. Gujarati DN, Porter DC. Basic Econometrics. New York: McGraw-Hill Education; c2009.
4. Hair JF, Black WC, Babin BJ, Anderson RE. Multivariate Data Analysis. Boston: Cengage Learning; c2019.
5. Hill RC, Griffiths WE, Lim GC. Principles of Econometrics. Hoboken: Wiley; c2018.
6. Kennedy P. A Guide to Econometrics. Hoboken: John Wiley & Sons; c2008.
7. Montgomery DC, Peck EA, Vining GG. Introduction to Linear Regression Analysis. Hoboken: John Wiley & Sons; c2012.
8. Pindyck RS, Rubinfeld DL. Econometric Models and Economic Forecasts. New York: McGraw-Hill Education; c2017.
9. Wheelen TL, Hunger JD, Hoffman AN, Bamford CE. Strategic Management and Business Policy: Globalization, Innovation, and Sustainability. Pearson; c2017.
10. Wooldridge JM. Introductory Econometrics: A Modern Approach. Toronto: Nelson Education; c2015.
11. Hsieh Y, Chen H. Customer data mining for personalized marketing: A predictive analytics approach. J Bus Res. 2020;112:75–84.
12. IDC. Predictive Analytics Market Outlook 2023–2026. International Data Corporation; c2023.
13. Islam NMR, Rahaman NMM, Bhuiyan NMMR, Aziz NMM. Machine learning with health information technology: Transforming data-driven healthcare systems. J Med Health Stud. 2023;4(1):89–96. doi:10.32996/jmhs.2023.4.1.11
14. McAfee A, Brynjolfsson E. Big data: The management revolution. Harv Bus Rev. 2012;90(10):60–68.
15. Rahaman M, *et al*. Machine learning in business analytics: Advancing statistical methods for data-driven innovation. J Comput Sci Technol Stud. 2023;5(3):104-111. doi:10.32996/jcsts.2023.5.3.8.
16. Zhang Z, Zhou X. Predictive analytics in healthcare: Progress and applications. Healthc Inform Res. 2019;25(3):157–163.