MULTIDISCIPLINARY SCIENTIFIC JOURNAL

NEYART

Vol. 4 No. 4

july - december 2026

VALIDATION OF A PREDICTIVE MODEL USING LINEAR REGRESSION AND ANOVA

Analysis | Data | Innovation | Results

Science that generates knowledge

ISSN: 2992-7161

NEYART Journal
ISSN: 2992-7161

VALIDATION OF A PREDICTIVE MODEL USING LINEAR REGRESSION AND ANOVA

VALIDACIÓN DE UN MODELO PREDICTIVO MEDIANTE REGRESIÓN LINEAL Y ANOVA

López Tarango Daniel Axel
Tecnológico Nacional de México/I. T. De Chihuahua II
https://orcid.org/0009-0005-7829-7275
daniel.lt@chihuahua2.tecnm.mx

Barraza Álvarez Alberto Guerrero
Tecnológico Nacional de México/I. T. De Chihuahua II
https://orcid.org/0000-0002-0197-1859
alberto.ba@chihuahua2.tecnm.mx

García Aguirre Ilse Haydeé
Tecnológico Nacional de México/I. T. De Chihuahua II
https://orcid.org/0009-0000-4012-476X
ilse.ga@chihuahua2.tecmn

Gallegos Orozco Carmen Angelina
Tecnológico Nacional de México/I. T. De Chihuahua II
https://orcid.org/0000-0002-4872-4927
carmen.go@chihuahua2.tecnm.mx

Orduño Aguirre Edgar Gerardo
Tecnológico Nacional de México/I. T. De Chihuahua II
https://orcid.org/0009-0007-8975-6734
edgar.oa@chihuahua2.tecnm.mx

DOI: https://doi.org/10.61273/neyart.v4i4.221 | **Received:** 18/05/2026 | **Accepted:** 24/07/2026 | **Published:** 28/08/2026

Abstract-- This paper verifies a simple linear regression prediction model, whose function is to predict in advance the monthly number of new customers of an institution. This verification uses the institution's full five-year historical monthly customer growth data. The tools used to verify the reliability of the model are the commonly used Excel, and the professional statistical software Minitab 19. The specific verification method is analysis of variance, which is the ANOVA mentioned in studies. After the verification, the model draws a clear conclusion: the correlation between the passage of time and customer growth is very high. The three statistical values supporting this conclusion are the coefficient of determination $R^2=0.8094$, the F-value=246.37, and the p-value≈0. Finally, the authors of the paper propose that in the future, these variables that may affect customer growth can all be incorporated into the model to carry out multiple regression analysis, to further help the institution formulate more reasonable decisions.

Keywords—Forecasting, Mathematical model, Planning, Quantitative methods, Statistical analysis**.**

Resumen— Este artículo verifica un modelo de predicción de regresión lineal simple, cuya función es predecir con anticipación el número mensual de nuevos clientes de una institución. Esta verificación utiliza el conjunto completo de datos históricos de crecimiento mensual de clientes de la institución durante cinco años. Las herramientas utilizadas para verificar la confiabilidad del modelo son el programa Excel, de uso común, y el software estadístico profesional Minitab 19. El método específico de verificación es el análisis de varianza, conocido como ANOVA, mencionado en diversos estudios. Tras la verificación, el modelo llega a una conclusión clara: la correlación entre el paso del tiempo y el crecimiento de la clientela es muy alta. Los tres valores estadísticos que respaldan esta conclusión son el coeficiente de determinación $R^2 = 0,8094$, el valor F = 246,37 y el valor p ≈ 0.. Finalmente, los autores del artículo proponen que, en el futuro, estas variables que pueden afectar el crecimiento de la clientela se incorporen al modelo para realizar un análisis de regresión múltiple, con el fin de ayudar aún más a la institución a tomar decisiones más razonables.

Palabras Clave— Previsión, Modelo matemático, Planificación, Métodos cuantitativos, Análisis estadístico.

INTRODUCTION

At present, institutions and enterprises in all walks of life all face an unavoidable demand—they need reliable statistical tools to help predict the trends of several core operational indicators in advance. For example, they need to have a clear idea ahead of time of figures such as the market demand for their own products, the level of current-period sales that can be achieved, and whether the factory's output can keep up. Currently, tools that can accurately support such predictions still cannot fully meet everyone's needs, and this study is carried out to fill this gap.

To complete the prediction of such operational indicators, the most widely used quantitative analysis method in the industry at present is linear regression. If linear regression is paired with another statistical method, the ANOVA mentioned in the original text, two key functions can be achieved: not only can a complete prediction model be built, but the reliability of this model can also be strictly verified from a statistical perspective—confirming whether it can really be used for prediction, and whether its ability to explain the causes of indicator changes is sufficiently strong.

The following section will first sort out the background of applied research on these methods at both the international and domestic levels, that is, what relevant studies at home and abroad in the past have adopted the combined method of linear regression plus ANOVA; then clarify the practical application, theoretical, and research-methodological reasons for carrying out this study, explaining why this research is conducted from three dimensions; after that, list the overall objectives of this study, that is, what core tasks this research will ultimately achieve; finally, it will also point out various constraints that limit the scope of this study, clarifying what boundaries this research will not exceed, and explicitly defining the coverage of this study.

International Background

Globally, the two methods of linear regression and ANOVA are widely used in a large number of research fields to validate predictive models; that is, they are used to judge whether a model used for prediction is reliable and usable. Research by Jahan (2017) points out that when non-normal errors occur in the process of multiple linear regression análisis to put it simply, the calculated errors do not conform to the conventional situation of normal distribution the stability of ANOVA F-test's will be mainly disturbed by two factors, and the originally stable and reliable test characteristics will encounter problems: one is excess kurtosis, which means the kurtosis of the data is too high, and the other is insufficient sample size, meaning the number of collected samples is too small. These problems will ultimately affect people's interpretation of the model's significance, making it impossible to accurately judge whether the model is actually useful, whether it can pass statistical tests, and clarify its true value.

In the healthcare field, Zou et al. (2024) developed and validated multiple linear regression models to predict the diameter of the acetabular prosthesis used in total hip arthroplasty, demonstrating that this type of model, based on analysis of variance, allows for the generation of clinically useful predictive tools from anthropometric variables.

In the energy sector, scholars Amral, Özveren, and King completed a relevant study in 2007. The research method they adopted was multiple linear regression, which was used to forecast electricity demand over a short period in the future, that is, to calculate the total amount of electricity that would be consumed during that period. During the research, they also conducted statistical significance tests a set of specialized statistical tests designed to verify whether results are reliable, which ultimately confirmed that two types of data can be stably and reliably used for prediction: one is various climate-related variables, and the other is historically recorded historical electricity consumption data.

There is another similar study completed by Kandananond in (2011). At that time, he conducted a set of comparative tests, using two methods, linear regression and artificial neural networks, to calculate the forecast values of Thailand's electricity demand respectively, and then carefully compared the actual performance of the two methods after application. Finally, he reached a clear conclusion: no matter which of the two methods is used, rigorous statistical verification must be carried out to ensure that the final calculated forecast results are sufficiently accurate, and this verification step cannot be skipped to directly use the results.

In relevant research in the field of project management, Ottaviani and De Marco launched a new set of tools in (2022), which is the multiple linear regression model. The core purpose of their development of this model is to improve the accuracy of predicting the final cost of a project this predicted final cost is the estimate at completion mentioned in the original text. After completing the model, they did not put it directly into use. Instead, they first used the method of analysis of variance to test the significance of each regression variable used in the model one by one, to confirm that the influencing factors added to the model can indeed play a role in prediction. Compared with the traditional earned value management method commonly used in the industry in the past, this new method has indeed reduced the fluctuation degree of prediction errors, and the prediction stability is better than the old method.

Within the research scope of renewable energy, Arias Velásquez (2022) carried out a relevant study. This study adopted analysis tools such as regression models, which included the specific regression model Ridge regression, and these regression models were not used alone; instead, they were combined with deep learning technologies to jointly conduct analysis and calculation, with the goal of predicting the energy output of photovoltaic power stations, that is, solar power stations. To verify the reliability of these prediction results, he also used two types of indicators for testing: one type is error measurement indicators used to measure the size of errors, and the other type is statistical significance indicators, and finally completed the verification of the accuracy of these prediction results.

In environmental research, Abdullah et al. (2020) constructed a multiple linear regression model to predict the concentration of inhalable particulate matter PM10 during cross-border haze periods in Malaysia. The cross-border haze mentioned here refers to haze weather that drifts into Malaysia from foreign countries, while PM10 is inhalable particulate matter with a diameter not exceeding 10 micrometers, which directly affects air quality and public health. Predicting its concentration changes in advance is of great importance for responding to such sudden air pollution events. After building the model, they also used two professional statistical methods, analysis of variance and the coefficient of determination, to verify the reliability of the model and confirm that it could produce sufficiently accurate prediction results.

In another study in the same field, Yusuf et al. (2019) tested the multiple linear regression model and the artificial neural network model respectively to explain the fluctuation pattern of PM10 concentration during haze periods in Malaysia that is, why PM10 concentrations in the air can vary in level and change constantly even on all haze days. They conducted a comparative assessment of these two commonly used models, and finally concluded that using ANOVA for statistical verification is the key to selecting the model with the strongest predictive ability, and this step is indispensable.

Latin American Context

In Mexico, linear regression and analysis of variance have been applied in various productive and academic sectors for the construction and validation of forecasting models, demonstrating the relevance of this line of research in the national context.

Morales Castro, Ramírez Reyes, and Rodríguez Albor (2019) forecasted sales for Mexican companies in the food sector—including Bachoco, Bimbo, and Grupo Lala—by comparing multiple linear regression models with artificial neural networks, finding that the linear models outperformed the 90% adjustment level in most of the cases analyzed.

Enríquez Zárate and Rodríguez Lozada (2021) applied and compared various forecasting techniques, including simple linear regression, for inventory planning at a small-to-medium-sized enterprise— —a building materials distributor in Tlaxcala, Mexico, and statistically validated the fit of each model against actual demand behavior.

Baeza-Serrato and Vázquez-López (2014), in a study conducted at a company in the Mexican textile sector, proposed transitioning from a predictive multiple linear regression model to an explanatory nonlinear regression model, using analysis of variance and mean squared error as validation criteria to increase the reliability of the predictions.

Ruiz Mendoza and Hernández González (2021), at the Celaya campus of the National Technological Institute of Mexico, applied quantitative forecasting methods—including linear regression—to determine the future demand for a plastic part manufactured by a company in the Bajío region, comparing the performance of each technique by calculating the forecast error.

Limones López, in his thesis developed at the Mexican Corporation for Materials Research (COMIMSA), used the linear regression model and its corresponding analysis of variance table as part of the methodology for the statistical control of defective production within an industrial process, validating the significance of the relationship between the variables under study.

Morantes-Quintana, Rincón-Polo, and Pérez-Santodomingo (2019), in an article published in the International Journal of Environmental Pollution (published in Mexico), developed a multiple linear regression model to estimate particulate matter concentration, validating the model using the adjusted coefficient of determination and the global significance test.

Scholars Fierro Torres, Castillo Pérez, and Torres Saucedo conducted a comparative analysis in 2022. The two types of models they compared are all demand forecasting tools that can be used in small and medium-sized enterprises in Mexico — simply put, they are tools to help these small and medium-sized businesses predict how much of their own products or services the market will need. One category is traditional models, and the other is modern models of the same type. Throughout the comparative analysis, the three scholars specifically pointed out the limitations of the multiple linear regression model: when encountering scenarios where data does not change linearly following a fixed trend, that is, when a non-linear change occurs, or when encountering unconventional sudden special events, this model cannot function properly and becomes unusable.

Pérez Paredes, Cruz de los Ángeles, Guatemala Villalobos, and Juárez Fonseca (2018) analyzed the importance of forecasting methods—including linear regression—in decision-making at Latin American micro, small, and medium-sized enterprises, emphasizing the need for managers to understand the statistical fundamentals of these techniques before using automated computational tools.

Study Objective

To statistically validate a forecasting model based on linear regression using analysis of variance (ANOVA), in order to determine the significance of the predictor variables and the reliability of the model for estimating the future behavior of the study variable.

Rationale for the Study

From the perspective of practical application, conducting this research is necessary. For any institution, the decisions that need to be made actually fall into two categories: one is operational decisions that manage daily operations, and the other is strategic decisions that set long-term directions. Regardless of the category, forecasting is an indispensable core support. With reliable forecasting, various future conditions linked to demand, production, and sales can be clarified in advance, so that one will not be caught off guard when problems arise, and there will be a solid basis for making decisions. Pérez Paredes et al. (2018) proposed that if decision-makers can master multiple forecasting methods and apply them correctly, they can systematically anticipate the daily operational activities of the institution, reduce uncertainty, allocate the resources at their disposal to the areas that need them most, and achieve more reasonable resource allocation.

From a theoretical perspective, the study is grounded in the statistical foundations of linear regression and analysis of variance, which allow the total variability of a response variable to be broken down into the portion explained by the model and the portion attributable to error, serving as the basis for testing the hypothesis of the model's significance. In this regard, analysis of variance is considered essential for comparing the explained and unexplained variability within a regression model and for determining whether the observed relationship between the variables is statistically significant (FasterCapital, 2024).

From a methodological perspective, the research is justified by the need to employ a rigorous statistical procedure that allows not only for the generation of a predictive model but also for the validation of its assumptions and significance through formal tests. In this regard, Jahan (2017) emphasizes that the ANOVA procedure applied to multiple linear regression models constitutes the appropriate methodological approach for testing the hypothesis of the model's overall significance, although its validity may be compromised when the assumptions of normality and homogeneity of variance are not met, which requires a systematic review of these assumptions before accepting the forecast results.

Study Limitations

This study itself has limitations, and all these issues are directly related to the types of models used. First of all, the linear regression model applied in the study has a core premise for its validity: there must exist a linear corresponding variation relationship between all the variables it analyzes—when one variable changes by a certain amount, another changes by a fixed proportion accordingly. However, if the variable that is the key focus of the actual research does not exhibit such linear variation at all, whether it shows curved fluctuations that rise and fall, periodic cyclic changes that follow a certain pattern, or is affected by a special event that has never appeared in historical records, the predictive ability of this linear regression model will be greatly weakened, and the calculated results will be completely unreliable.

Second is the ANOVA used in this study. For the results derived from it to be valid, three strict premises must be met simultaneously: first, the residuals conform to a normal distribution; second, the residuals have homogeneity of variance; third, the residuals are independent of one another. If any one of these requirements is not met, it becomes impossible to correctly interpret the statistical significance of the model, the final assessment results will be skewed, and no accurate conclusions can be drawn from the analysis.

In addition to these inherent problems of the models themselves, the historical data used in the study has additional limitations. Whether these data can be successfully obtained, whether the quality of the data itself meets standards, and whether the total accumulated records are sufficient—each of these factors will affect the final results of the study. If the observation data used for analysis is too small, the conclusions drawn from the estimation will be untenable, unable to remain stable and reliable, and even less able to be directly applied to other scenarios or other time periods..

DEVELOPMENT

Research Approach

This research was conducted using a quantitative approach, as it is based on the numerical measurement of the study variables—the number of customers served over time—and on the use of statistical analysis to identify behavioral patterns and generate a forecasting model. In this regard, Acosta Faneite (2023) notes that the quantitative approach is characterized by predictions, which constitute interpretations of observed facts, such that their direction is predictable; he also specifies that this approach is based on the measurement of numerical data to answer research questions and test hypotheses through sequential statistical processes.

Research Paradigm

The study was conducted within the positivist paradigm, also known as the quantitative or empirical-analytical paradigm, whose main objective is to explain, predict, and control phenomena by verifying theories and laws through empirical observation and statistical analysis. According to Herrera Castrillo (2024), quantitative research is grounded in this paradigm by employing rigorous methods and techniques to collect and analyze numerical and statistical data, with the aim of obtaining precise and verifiable results, which is consistent with the statistical validation of the forecasting model proposed in this study.

Research Method

The hypothetical-deductive method was employed, which begins with the formulation of an assumption or hypothesis regarding the relationship between the study variables, to subsequently test it empirically through statistical analysis of the collected data. De la Cruz Sullca (2020) notes that the validity of a scientific idea is established by deriving its implications for the real world and verifying whether the resulting prediction is correct or not; thus, the scientific method is, in essence, hypothetical-deductive; this approach is relevant to the present study, in which a regression model was derived from historical data and tested using analysis of variance.

Research Design

The research design was non-experimental, longitudinal, and retrospective, since the study variables were not deliberately manipulated; rather, historical data spanning a five-year period were analyzed, observing the behavior of the variables as they occurred in their real-world context. In this regard, Sienes Bailo et al. (2021), in their retrospective observational study on crystalluria, illustrate how this type of design allows for the analysis of existing data, collected over a defined period of time, to identify relationships between variables without direct intervention by the researcher in the events under study.

Type of Research

The study was descriptive-correlational in nature, as it sought, first, to describe the historical trends of the variable “number of patients seen” and, subsequently, to determine the degree of relationship between this variable and time (months), with the aim of constructing and validating a predictive model. Cairampoma Carrillo (2015) notes that correlational research is a type of study that seeks to measure the degree of relationship between two or more variables without experimentally manipulating any of them, relying instead on observation and statistical analysis of the available data.

Participants

The sample for this pilot study consisted of a total of 15 students enrolled in the first semester of their vocational training at a public university, covering the academic term from January to June 2025. The group of students participating in the study did not use a professional sampling method that could guarantee sample randomness; instead, they adopted non-probability convenience sampling—in simple terms, only selecting individuals that were easily accessible to the researchers. They were chosen because these students were all participating in the school’s new technology literacy course at that time, so the research team could easily gather this group regardless of coordinating schedules or securing a venue, which met the selection requirements for convenience sampling.

A total of 15 participants were included in this study. The gender distribution of their demographic characteristics was as follows: 11 males, accounting for 73.33%, and 4 females, accounting for 26.67%. All participants were first-year university students who had just enrolled, with ages ranging from 18 to 20 years old. The researchers believed that selecting this group for the assessment was perfectly suitable, as it could accurately measure one specific thing—the baseline level of numeracy that students bring with them when they transition from high school to the university environment, which is the baseline numeracy to be measured in the study.

Procedure

The entire process of this study was completed within the semester from January to June 2025. During the six months of the research's advancement, we first retrieved all historical data from the previous five years. These data detailed the total number of clients we had served each month, and we conducted dedicated analysis centering on these data. There were two clear goals to achieve for this data analysis: first, to clarify the exact nature of the correlation between the two variables; second, to build a set of models that can be used for prediction, relying on the statistical analysis method of linear regression.

After the preliminary construction of this prediction model, we did not put it into use directly; instead, we launched a statistical verification process, and the verification method used was analysis of variance, that is, the originally marked ANOVA. The purpose of conducting this verification was also clear: first, to confirm the significance of the model, meaning that it is valid and has research value; second, to test whether its actual prediction ability can meet the requirements.

After completing all verification steps, we finally produced a prediction model with higher reliability for the institution involved in this study..

Ethical Considerations

Throughout this research, the fundamental ethical principles governing scientific practice were upheld, including the confidentiality of the information obtained and the responsible use of historical data provided by the organization, which were used exclusively for academic and statistical analysis purposes. In this regard, Iriarte et al. (2017) note that human rights violations throughout history have led to the development of ethical guidelines in scientific research, with respect for life, autonomy, confidentiality, and the responsible handling of information being some of the fundamental ethical principles that every researcher must uphold during their research work—principles that were observed in the handling of the information used in this study.

Procedure for Data Analysis and Statistical Analysis

When processing the collected data, Microsoft Excel was used for the first step of data organization and statistical analysis. First, it was used to sort out the accumulated historical database, then the built-in data analysis tool of Excel was called to generate a linear regression model. With this model, all coefficients in the regression equation and the corresponding coefficient of determination were calculated.

After the first round of analysis results were obtained, the statistical software Minitab 19 was used to conduct subsequent verification. The verification to be completed included two parts: first, to check whether the results previously calculated with Excel were accurate; second, to confirm statistically that the model was qualified and usable through the method of analysis of variance.

The statement in Gestión de Operaciones, published in (2016), is fully consistent with this set of operation procedures. The book mentions that to build a regression model that can be used for prediction, there are two optional tools: one is the Excel data analysis tool used just now, and the other is the statistical software Minitab. It also clearly states that when these two tools are used to complete the same analysis process, the only difference is the way the results are presented, and the final calculated coefficients are completely identical.

Data Analysis

Model Fit

We conducted a simple linear regression analysis and calculated the multiple correlation coefficient as 0.8997. This coefficient is a value used to measure the strength of the linear correlation between two types of data, with a range from 0 to 1. The closer it is to 1, the stronger the linear correlation between the two variables, and it is a positive correlation—that is, when one value rises, the other will also rise accordingly. The value calculated this time is close to 0.9, very near to 1, which means that between the passage of time counted on a monthly basis and the number of customers served in the same period, there exists a strong positive linear correlation—the longer the time that passes, the larger the cumulative number of customers served, and the trend of synchronous growth of the two is very obvious.

Next, the coefficient of determination of the model we calculated, that is, R^2, has a value of 0.8094. The coefficient of determination helps us judge how much of the variation in another set of data can be explained by the variables included in the model. The situation here is that the number of customers we serve each month is not fixed, it fluctuates up and down every month, and this kind of numerical fluctuation is called "volatility". This R^2 value means that approximately 80.9% of all the observed volatility in the number of customers served can be clearly explained by the time variable incorporated into the model—only the factor of "how many months have passed" can explain the vast majority of the causes of fluctuations in the number of customers.

The remaining 19.1% of the volatility in the number of customers comes from various factors that were not included in the model's calculation, and these fluctuations cannot be explained by the time variable. These factors that were not included in the model are also listed in the original text, including seasonal changes, marketing activities, market environment, and all external events that are not controlled by the model's input variables; all of these are possible sources of that part of the volatility that cannot be explained by time.

The adjusted R^2, which corrects the coefficient of determination based on the number of observations and the number of included predictor variables, was 0.8062. The closeness between R^2 and the adjusted R^2 confirms that the model's explanatory power is not artificially inflated by the sample size, and that the independent variable provides genuinely relevant information for explaining the behavior of the dependent variable.

Meanwhile, the standard error of the estimate was 30.04 customers, representing the expected average deviation between the observed values and the values predicted by the model. Considering that the number of customers served ranged from relatively low values at the beginning of the period to considerably higher values toward the end, this margin of error can be considered reasonable and consistent with the nature of the data analyzed.

Overall Significance of the Model (Analysis of Variance)

We can use Analysis of Variance (ANOVA) to determine whether the regression model we have on hand truly has overall statistical significance, or if it is merely an invalid result derived from accumulated random errors. After conducting analysis with this tool, we calculated two key values: the first is the F-statistic, which is 246.37, and the second is the critical F-value, which is 1.53×10^{-22}. There is a general judgment threshold in the field of statistics, that is, a significance level of 0.05, and the critical F-value we just calculated is far smaller than this threshold. With this result, we can reject the null hypothesis we initially set—the preliminary judgment that there is no linear relationship between variables. In addition to rejecting the null hypothesis, we can also draw two definite conclusions: first, the regression model we proposed has high statistical significance and is indeed a valid model with reference value; second, the independent variables in the model can indeed effectively explain the changes in the dependent variable, and their explanatory power is qualified.

Interpretation of the Regression Coefficients

The estimated regression equation, based on the coefficients obtained, can be expressed as follows:

$$\hat{Y} = 47.503 + 3.515\ (X)$$

The intercept coefficient (47.503) represents the estimated number of customers served at the origin of the period analyzed, that is, the baseline level of service before considering the effect of the passage of months. This value has a t-statistic of 6.049 and an associated probability of 1.13×10^{-7}, confirming that the intercept is statistically different from zero and, therefore, significant within the model.

The Month (X) coefficient of 3.515 indicates an average monthly increase of approximately 3.5 customers served. This positive effect is highly statistically significant ($t=15.696$, $p=1.53 \times 10^{-22}$), supported by a 95% confidence interval of 3.066 to 3.963 that excludes zero.

Discussion

The results obtained in this study, summarized by a coefficient of determination (R^2) of 0.8094 and a highly significant F-test ($F = 246.37$; $p < 0.05$), empirically confirm the theoretical and methodological approaches outlined in the reviewed literature, according to which analysis of variance (ANOVA) is the ideal statistical method for testing the overall significance of a regression model and for decomposing the total variability of the dependent variable into the portion explained by the model and that attributable to error. In this regard, the finding of an F-value with an associated probability of virtually zero (1.53×10^{-2}) aligns with the role that FasterCapital (2024) attributes to ANOVA as a tool for determining whether the observed relationship between variables is statistically significant and not the result of chance.

The magnitude of the R^2 obtained in this study is consistent with the findings reported by Morales Castro, Ramírez Reyes, and Rodríguez Albor (2019) in their study on sales forecasting for Mexican companies in the food sector, where multiple linear regression models achieved levels of fit exceeding 90% in most of the cases analyzed. Although the goodness-of-fit obtained here (80.9%) is slightly lower, both results point in the same direction: linear regression models, when applied to variables with systematic behavior and a defined trend over time, offer satisfactory explanatory performance within the Mexican business context, without necessarily requiring the use of more complex techniques such as artificial neural networks.

Similarly, the pattern of sustained growth identified in the “customers served” variable corresponds to that reported by Amral, Özveren, and King (2007) and by Ottaviani and De Marco (2022), who, in different contexts—short-term electricity demand forecasting and the estimation of final project costs, respectively—confirmed through statistical significance tests that the variables included in their multiple linear regression models were reliable predictors, capable of reducing error variability compared to alternative methods. The present study reinforces this evidence by showing that even a simple regression model, with a single explanatory variable of a temporal nature, can achieve robust levels of statistical significance when the underlying phenomenon exhibits a clear and sustained trend.

However, it is important to contrast these results with the warnings issued by Jahan (2017), who noted that the robustness of the F-test in ANOVA procedures applied to regression models is compromised when the error does not follow a normal distribution, particularly in the presence of excessive kurtosis or small sample sizes. In the case analyzed, the number of observations (sixty monthly data points, equivalent to five years) can be considered adequate to mitigate this risk; however, the interpretation of the model's overall significance must be made with the methodological caveat that, within the framework of this analysis, full compliance with the assumptions of normality and homogeneity of variance of the residuals was not exhaustively verified—an aspect that Jahan (2017) identifies as critical to the validity of inferential conclusions.

Furthermore, it is important to consider the fit obtained within the limitations inherent in linear regression models, as pointed out by Fierro Torres, Castillo Pérez, and Torres Saucedo (2022), who note that this type of model loses predictive power when faced with scenarios involving nonlinear behavior or atypical events not captured in the historical data series. The 19.1% of variability not explained by the model estimated here could, in fact, be associated with factors of this nature—seasonality, sales campaigns, or changing market conditions—that a simple regression model, based exclusively on the passage of time, is unable to capture. Along the same lines, Kandananond (2011) argues that both linear regression models and alternative approaches, such as artificial neural networks, require rigorous statistical validation to ensure the accuracy of their estimates, which confirms the importance of having subjected the proposed model to an ANOVA test before accepting its results as a basis for decision-making.

On the other hand, the individual significance obtained for the coefficient of the Month variable ($t = 15.696$; $p \approx 0$) is consistent with the findings of Enríquez Zárate and Rodríguez Lozada (2021), who, when applying forecasting techniques—including simple linear regression—for inventory planning in a small and medium-sized Mexican company, emphasized the importance of statistically validating the fit of each model against the actual behavior of the studied variable before using it as a planning tool. In the case analyzed, both the narrow confidence interval for the slope (3.066 to 3.963) and the absence of the value zero within that interval support the existence of a positive and stable relationship between time and the number of customers served, which gives the model an adequate level of reliability for use in operational planning, in line with the practical relevance highlighted by Pérez Paredes, Cruz de los Ángeles, Guatemala Villalobos, and Juárez Fonseca (2018) attribute to a correct statistical understanding of forecasting methods by decision-makers in micro, small, and medium-sized enterprises.

Finally, the standard error of the estimate (30.04 customers) allows us to qualify the practical scope of the model: while the statistical significance of the forecast is beyond dispute, the absolute magnitude of this error suggests—in line with the arguments put forward by Zou et al. (2024) regarding predictive models applied in clinical contexts, that the usefulness of a regression model should not be evaluated solely on the basis of its statistical significance, but also on the basis of the magnitude of the expected error in relation to the scale of the variable of interest. Overall, the evidence gathered supports the conclusion that the prognostic model validated in this study is statistically robust and consistent with the reviewed literature, although its practical application must be complemented by continuous monitoring of the model's assumptions and the eventual incorporation of additional variables that capture the variability not explained by the simple passage of time.

Conclusions

The objective set forth in this study—to statistically validate a forecasting model based on linear regression using analysis of variance—was successfully achieved. The results obtained allow us to conclude that the "month number" variable significantly explains the behavior of the number of clients served, with a coefficient of determination of 0.8094, which demonstrates strong explanatory power, and an F-test (246.37; $p < 0.05$) that confirms the overall significance of the model as a whole.

The analysis of the regression coefficients reinforced this conclusion by showing that both the intercept and the slope associated with the time variable were statistically different from zero, with high t-statistics and associated probabilities that were practically zero. The 95% confidence interval obtained for the slope—which is narrow and does not include zero—supports the existence of a positive, stable, and predictable relationship between the passage of time and the growth in the number of customers served, thereby validating the model as a useful tool for planning the operational capacity of the service under analysis.

In general terms, it can be stated that simple linear regression, when subjected to a rigorous statistical validation procedure using ANOVA, is an appropriate technique for building reliable forecasting models in contexts where the variable of interest exhibits a systematic and sustained trend over time. However, the 19.1% of variability not explained by the model, as well as the identified standard error of estimation, indicate that there is room for improvement that could be addressed by incorporating additional predictor variables or through complementary methodological approaches.

As a first line of future research, we propose extending the current model to a multiple linear regression that incorporates additional variables of a seasonal, commercial, or macroeconomic nature—such as marketing campaigns, holidays, local market conditions, or regional economic indicators—with the aim of reducing the proportion of variability not explained by the current model and strengthening its predictive capacity in more complex scenarios. This line of research would also allow for the evaluation of multicollinearity among the new predictors and, , the use of ANOVA to test the individual contribution of each variable to the overall significance of the expanded model.

As a second line of future research, we recommend exploring the application of alternative or complementary models to linear regression—such as time-series models with seasonal components, nonlinear regression, or machine learning techniques—in order to compare their predictive performance against the linear model validated here, particularly in the presence of cyclical behaviors or atypical events that simple linear regression is unable to capture. This comparison would make it possible to determine, based on error criteria and statistical significance, which approach offers the best balance between predictive accuracy and ease of interpretation for the organization under study.

Conflict of Interest

The authors declare that they have no conflicts of interest.

Data Availability

All datasets relevant to the results of this study are available in their entirety in the article.

Source of Funding

This study was not funded by any organization.

Statement on Generative AI

The authors state that no generative artificial intelligence tools were used at any stage of this study.

References

Aguado, D., González, A., Antúnez, M., & de Dios, T. (2017). Assessment of transferable skills among college students. Initial psychometric properties of the Transferable Skills Questionnaire. *REICE.*

Ibero-American Journal on Quality, Effectiveness, and Change in Education, *15*(2), 129–152. https://doi.org/10.15366/reice2017.15.2.007

Asonitou, S. (2022). Measuring the gap between academic hard skills training and modern industry demands in higher education. *Higher Education, Skills and Work-Based Learning*, *12*(3), 455-472. https://doi.org/10.1108/HESWBL-05-2021-0098

Asonitou, S. (2022). The digital transformation era: Redefining technical mastery and soft skills in higher education. *European Journal of Higher Education*, *12*(3), 289-305.

García-Vargas, M., López-Maldonado, J., & Restrepo, A. (2022). Gaps in technical training upon entry into higher education in Latin America. *Ibero-American Journal of Higher Education*, *13*(36), 88-105. https://doi.org/10.22201/iisue.20072872e.2022.36.1192

He, T., Zhang, L., & Chang, Y. (2021). The definitive role of hard skills in early engineering employment outcomes. *Journal of Engineering Education Research*, *29*(2), 114-129. https://doi.org/10.1016/j.jeer.2021.02.004

Jones, P., & Broadbent, M. (2020). OECD metrics and the standardization of technical competencies in global higher education. *Educational Evaluation Policy Analysis*, *42*(4), 567-584. https://doi.org/10.3102/0162373720934521

Martínez-Sánchez, F., & Rodríguez, C. (2023). Technical competencies and professional qualification frameworks in Mexican higher education. *Journal of Higher Education*, *52*(206), 45-63. https://doi.org/10.36857/resu.2023.206.2415

Pérez-Franco, J. (2021). Structural challenges in the assessment of disciplinary skills in the Southern Cone. *Quality in Education*, (54), 112–139. https://doi.org/10.31619/caledu.n54.968

Ramírez, G., & González, E. (2023). Digital literacy and diagnostic quantitative skills among first-year students. *Apertura*, *15*(1), 34–49. https://doi.org/10.32870/Ap.v15n1.2301

Silva, M., & Santos, R. (2022). Gender gaps in the self-perception of hard technological skills among college students. *Latin American Journal of Educational Technology*, *21*(2), 75-91. https://doi.org/10.17398/1695-288X.21.2.75

Smith, J., & Martínez, K. (2023). Quantitative deficiencies in first-year university students and their correlation with retention. *Journal of College Student Retention: Research, Theory & Practice*, *25*(1), 142-161. https://doi.org/10.1177/15210251211011342

Torres-Cáceres, L. (2024). The disconnect between hard academic skills and Latin American SMEs. *Estudios Gerenciales*, *40*(170), 89-102. https://doi.org/10.18046/j.estger.2024.170.6120

Tymon, A. (2021). The 4.0 Industry Paradigm: Redefining Technical Mastery Within Academic Structures. *Studies in Higher Education*, *46*(8), 1611-1624. https://doi.org/10.1080/03075079.2019.1698532

Valencia-Maldonado, G. (2023). Diagnostic competency-based assessment in Andean higher education. *Formación Universitaria*, *16*(3), 55-66. https://doi.org/10.4067/S0718-50062023000300055

Wilson, H., & Taylor, S. (2022). Objective performance testing vs. self-assessment of technical instrument skills in freshmen. *Assessment & Evaluation in Higher Education*, *47*(5), 789-804. https://doi.org/10.1080/02602938.2021.1963452

Zhao, Q., & Wang, Y. (2024). Technological hard skills entry requirements and academic adaptations in STEM fields. *International Journal of Educational Research*, *123*, 102-117. https://doi.org/10.1016/j.ijer.2023.102117

Collaborative Work Table

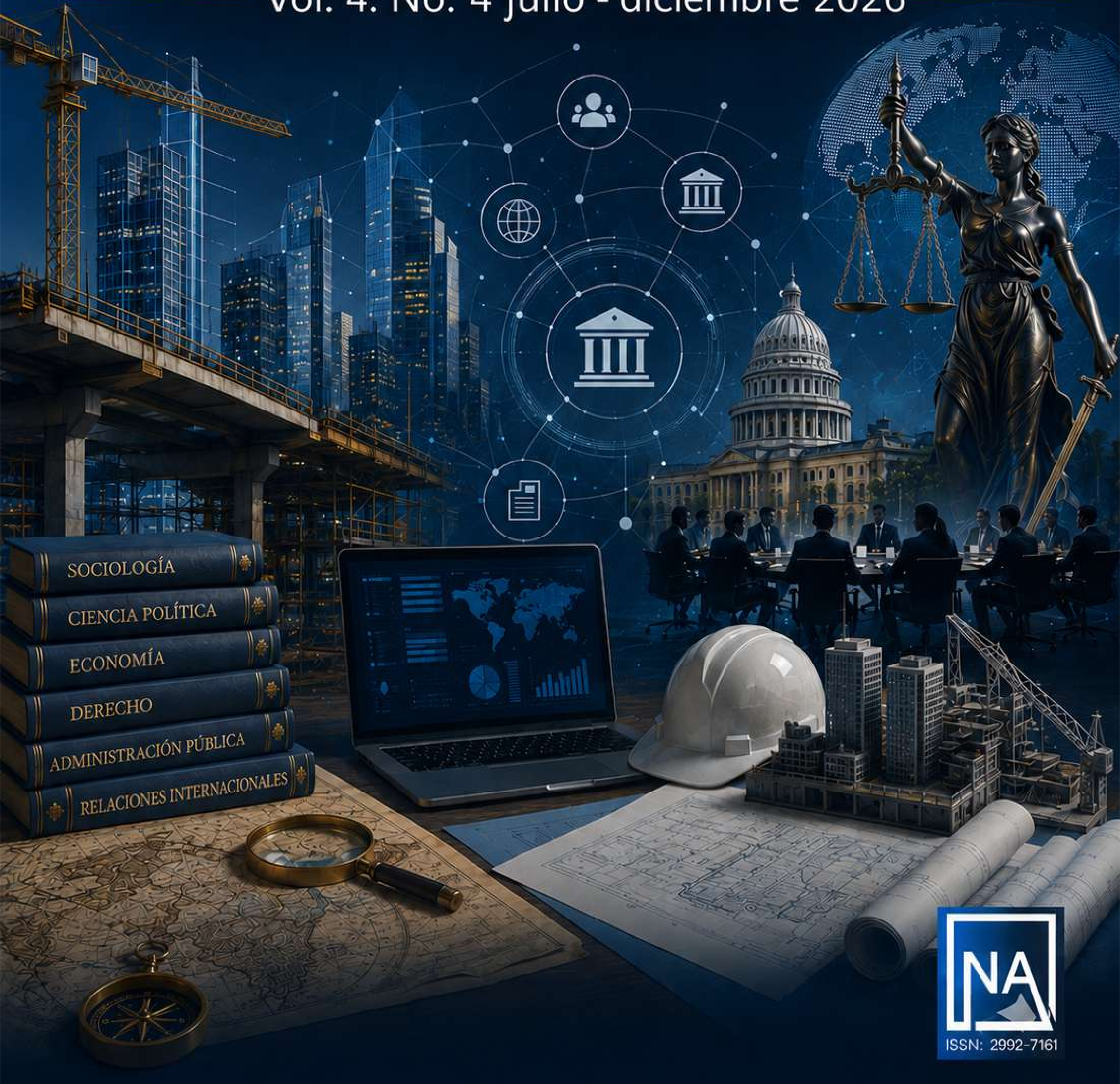
Role	Author(s)
Conceptualization	Daniel Axel López Tarango
Methodology	Alberto Guerrero Barraza Álvarez, José Roberto Holguín Beltrán
Software	Ilse Haydeé García Aguirre, Edgar Gerardo Orduño Aguirre
Validation	Brenda López Martínez
Formal Analysis	Ricardo García Parada, María Eugenia Astrid Macías Sagarminaga
Research	Víctor Villar Laguna, Fátima Yaraset Mendoza Montero
Resources	Brenda López Martínez
Data Curation	Ricardo García Parada, María Eugenia Astrid Macías Sagarminaga
Writing - Preparation of the original draft	Víctor Villar Laguna, Fátima Yaraset Mendoza Montero

Writing - Review and Editing	Brenda López Martínez
Visualization	Ricardo García Parada, María Eugenia Astrid Macías Sagarminaga
Supervision	Víctor Villar Laguna, Fátima Yaraset Mendoza Montero
Project Management	Brenda López Martínez
Fundraising	Ricardo García Parada, María Eugenia Astrid Macías Sagarminaga

REVISTA CIENTÍFICA MULTIDISCIPLINARIA

NEYART

Vol. 4. No. 4 julio - diciembre 2026



ISSN: 2992-7161