Estimation of default rates using the regression model and forward-looking modeling

Stelian STANCU[1], Ion-Florin RĂDUCU[2*]and Andreea PERNICI[3]

To cite this article:

Stancu, S., Răducu, I.-F., Pernici, A. (2024). Estimation of default rates using the regression model and forward-looking modeling. Romanian Journal of Economics, 58(1), pp. 109 - 116

Abstract. Objective: *The overall background of this paper is the economic and financial reality around credit risk and default probability estimates made by banks and factored in as a risk element inside predicted losses. This paper describes how to build a regression model for calculating forward-looking adjustment factors. In practice, determining a link between the default rate curve used to calculate the adjustments for expected losses related to credit risk exposures and macroeconomic factors is what this requirement entails. The default rate curve will be adjusted in accordance with the developed model based on forecasts of macroeconomic factors. Furthermore, forward-looking models include rational expectations, and economic agents understand the correct future values of each variable. The primary objective of the current paper analysis is to recognize statistical relationships between the default rate of a portfolio of physical clients, which is variable dependent, and one or more macroeconomic variables, which are variable independent. This will be accomplished by employing a number of linear rule models and, ultimately, by selecting a single model that will accurately represent the statistical and economic aspects of this relationship. Following that, on the basis of predictions, factors of adjustment k will be calculated over a three-year period, using PD-based historical conditions to calculate provision. The time span for which predictions will be fulfilled is from June 2019 to September 2022. Estimating default rates and discovering relationships between them and various macroeconomic factors can provide an overview of economic reality and reflect the extent to which a person has the necessary resources for the development and capitalization of their heritage. Higher default rates can indicate a lack of individual resources to repay loans, a change in macroeconomic reality that affects population incomes, or, indirectly, an inability to capitalize, and develop heritage.* ***Method:*** *This section will contain the rules for multiple-line regression. It was chosen because it is both robust and interpretable. There are multiple input variables that could have an impact on the outcome, or target variable, according to multiple regression. The classic multiple linear regression must be used here with a modification to make the dependent variable specifics more predictable. Rates by default will be between 0 and 1 (0% and 100%), with the emphasis on the fact that no negative values will be*

[1] The Bucharest University of Economic Studies, Bucharest, Romania; Centre for Industrial and Services Economics, Romanian Academy, Bucharest, Romania; stelian.stancu@csie.ase.ro

[2] The Bucharest University of Economic Studies, Bucharest, Romania; *Corresponding author: florin.raducu@csie.ase.ro

[3] The Bucharest University of Economic Studies, Bucharest, Romania; andreea.pernici@csie.ase.ro

allowed. To address this shortcoming, a logistical transformation will be implemented. The end result will be that all model predictions will be correct and will have values between 0 and 1. ***Results:*** *The model selected for predicting default rates is the one made up of the independent variables consumer price index and gross domestic product, as well as the dependent variable DR, for which a cubic transformation was performed. Lag4 and lag1 transformations were used for the two independent variables. The values derived from the projections are reasonable and appropriately capture the trend. Default rates climbed during the pandemic of 2020 as unemployment also increased, people had trouble making loan payments, some of them lost their jobs, and rates then started to decline. Additionally, using these forecasts, the yearly adjustment factors k were derived as a ratio between the average of the forecasted values for the previous four quarters and the average of the historical default rates during the preceding three years:* $k_1 = 0.59$, $k_2 = 0.70$, $k_3 = 0.39$. ***Originality:*** *The original approach entails establishing a correlation between macroeconomic variables that represent potential recent shocks to credit default rates in order to forecast these rates and, based on them, determine adjustment factors for the probability of default using forward-looking modeling.*
Keywords: *forward-looking, regression, default rates, probability of default, performance measures*
JEL classification*: C58, E44*

1. Introduction

This papers' starting point is the increasing interest in machine learning techniques and methods (such as the regression model in this case), which are becoming widely applied and have a lot of actuality these days across a wide range of fields, as well as economic-financial reality. The paper is divided into sections that each follow a phase in the construction of the forward-looking model. Forward-looking components must be used in accordance with IFRS 9 (International Financial Reporting Standards).

The goal of this papers' analysis is to discover a statistical relationship among the default rate of a portfolio of clients (individuals), as a dependent variable, and any number of macroeconomic indicators, as independent variables. This will be accomplished by employing a number of linear regression models before settling on a single model that statistically and economically reproduces this relationship. Following that, based on the forecasts, correction factors k will be generated and applied to the historical conditional PDs. The prediction will be made between June 2019 and September 2022. The analysis was done in *RStudio*.

The model of linear regression is one of the most widely used and straightforward algorithms in machine learning predictive analysis. This model operates well with tight assumptions such as the number of data points, the linearity of variables, normality of errors, multicollinearity, homoskedasticity, measurement reliability. Furthermore, there has been little consideration for processing and clearing discordant samples in data sets yet. These samples may have a significant impact on the results of multiple linear regression concerning these assumptions and multiple characteristics such as adjusted R-square, intercept slopes, exogenous variables, and reliability of predictions (Rasyidah et al., 2023).

Other researchers (Li et al., 2023) have adopted the forward-looking method as well to assist capital-markets investors.

Estimating default rates and discovering relationships between them and various macroeconomic factors can provide an overview of economic reality and reflect the extent to which a person has the necessary resources for the development and capitalization of their heritage. Higher default rates can indicate a lack of individual resources to repay loans, a change in macroeconomic reality that affects population incomes, or, indirectly, an inability to capitalize and develop heritage.

The article is divided into several sections that describe the steps of carrying out the case study after the introduction to this field of research, the review of the academic literature, the selection of the methodology and data used in the analysis, the results obtained, and the conclusions that can be debated based on these results.

2. Literature review

The literature has always been an instrument to target the evolution of scientific and real-world phenomena as well as a source of inspiration for research papers. Various financial and economic studies have been conducted regarding the estimation of default rates, the use of regression models, and forward-looking modeling.

Baltazar et al. (2020) computed the quarterly default rate of customer loans in the United Kingdom for various simulated stress-scenarios using a macroeconomic model. These stress-scenarios run from the first quarter of 2019 to the fourth quarter of 2023. These researchers simulated default rates and compared them to the default rates experienced during the 2007-2009 financial crisis. A simulated default rate's uplift, level of extremeness, and elasticity were also computed.

Another study in the financial field, prepared by Wattanawongwan et al. (2023), considers the development of a model which predicts the conditional quantiles of the Exposure At Default (EAD), i.e. the card balance at the time of default. For this study they use linear and D-vine copula-based quantile regression by utilizing a large dataset that contains the credit card defaults. The results of EDA (exploratory data analysis) show marginal distributions of EAD and its covariates are non-normal, have a high variance, and also violate the homoscedasticity hypothesis of errors. As a result, interval estimate models, such as quantile regression, that do not rely on parametric distribution assumptions and don't call for constant variance, are usually better suited to modeling such data than point estimate models. On the other hand, quantile regression models have a further benefit of letting variable effects to vary according to the EAD quantile of interest.

The forward-looking technique has been widely used in the literature to develop numerous case studies. For example, Guender (2002) demonstrates that hybrid nominal revenue targeting and strict inflation targeting are that two efficient monetary policy strategies in terms of that they are particular instances of the optimal strategy on the monetary market. These criteria are embraced by all tax policies that focus on the ultimate objective of the variables, namely the final product gap and the rate of inflation. The problem that needs to be solved is determining the optimal rule's weight on real output. This weight is shown to be dependent on the policymaker's preferences as well as the structural parameter that depends on the final product gap in the curve known as the Phillips Curve.

Regression is a well-known and widely used machine learning technique in a variety of research fields. Lin & Chen (2023) developed a logistic regression model to assess the credit risk of individual lenders. Their article develops an indexing system that can determine whether a borrower can repay a loan in full and on time in accordance with the loan contract, accurately assessing individual lenders' credit status, increasing commercial bank loan efficiency, and providing an adequate basis for commercial bank credit decisions.

Zhou et al. (2023) propose an efficient model to forecast credit default risk based on various types of inherent user relationships. The study's contribution is the development of a GAT-based model aimed at capturing latent user relationships and leveraging that lead to improve prediction accuracy. The research paper is well-focused and relevant, emphasizing the use of GNN methods to identify numerous connections between users, thereby addressing the issue of previous studies neglecting relationships. To guarantee meaningful results, their research centers around on user relationship construction and discovers that credit default risk is influenced by user dependencies rather than individual attributes.

Peláez et al. (2024), in a future paper for the future year, propose a nonparametric estimator of the probability of default that considers the existence of a group of cured people who will never experience the default. This estimator is based on López-Cheda's (2018) nonparametric survival estimator for mixture treatment models. The asymptotic bias and variance of the NPCM probability of the default estimator, as well as its asymptotic normality, are demonstrated. Given the low degree of prediction errors, the NPCM estimator outperforms traditional approaches, including semiparametric models, for estimating the probability of default. Beran's estimator, which appears in the comparative study as another nonparametric method, exhibits impressive behavior.

3. Methodology and data

The linear regression model, one of the most well-known models that is relevant to numerous studies in various domains, was used to conduct the analysis.

This section will contain the rules for multiple-line regression. It was chosen because it is both robust and interpretable. There are multiple input variables that could have an impact on the outcome, or target variable, according to multiple regression. The classic multiple linear regression must be used here with a modification to make the dependent variable specifics more predictable. Rates by default will be between 0 and 1 (0% and 100%), with the emphasis on the fact that no negative values will be allowed. To address this shortcoming, a logistical transformation will be implemented. The end result will be that all model predictions will be correct and will have values between 0 and 1. The model looks like this:

$$\ln\left(\frac{Y}{Y-1}\right) = X\beta + \varepsilon \quad (1)$$

where Y is the dependent variable, X represent the independent variables, β are the coefficients of the regression model, and ε is the error.

The analysis involved stationary data series (series with a constant mean, variance, and autocorrelation over time). The assumption of stationarity is used in the majority of statistical forecasting methods. There are few macroeconomic series that exhibit stationarity in reality, but with the help of simple mathematical modifications, they can be brought to an approximate form of stationarity. Given that trend can introduce non-stability into historical default rate series, we used three statistical tests for check stationarity: Augmented Dickey-Fuller (ADF), Kwiatkowski–Phillips–Schmidt–Shin (KPSS), Phillips-Perron (PP).

In terms of level of significance, we utilized a p-value of 10% for ADF and PP, and a p-value of 5% for KPSS. Setting the value at 10% for ADF and PP tests is a "compromise" that tends to include the most statistically significant results while also taking into account the fact that, in practice, the chances of a series of data being stationary are very small. About the KPSS test, this is contrary to ADF and PP tests to test stationarity and is based on linear regression which breaks the time series into deterministic trend, random walk, and stationary error. As is typical in both literature and practice, we have selected a p-value of 5% in this instance.

If the p-value for stationarity tests is less than 10%, or 5% for KPSS, the null hypothesis is rejected; if the p-value of the test is greater than the required significance level, the null hypothesis is accepted.

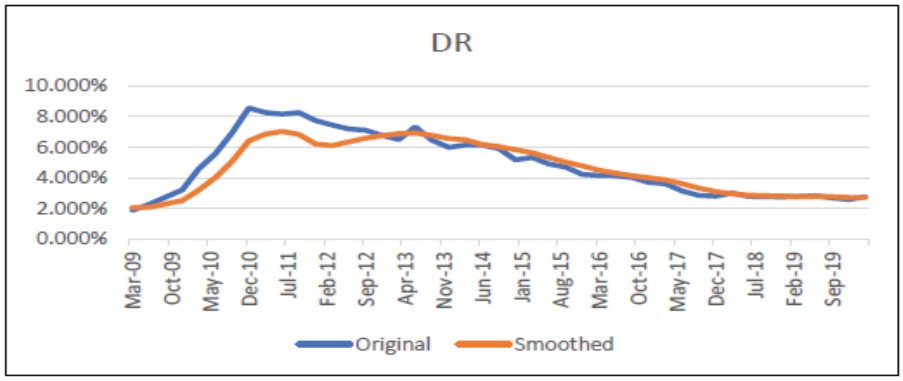
Also, correlation between independent variables can be a serious issue and cause multicollinearity, which violates one of the main assumptions of linear regression. If there are significant correlations (>50%) between the independent variables, there is a danger that the resulting model will be unsuccessful and have a high variance. The correlation matrix was used as a starting point to determine which combinations of elements to avoid. To eliminate multicollinearity, it was determined that only one form of each variable (GDP/UR/HICP) would be employed in a model.

To reduce the fluctuations caused by the methodological change that appeared in the history (modification of the default definition), the series of DRs is smoothed as follows:

- The goal is to remove the extreme values from the series because they contradict the hypothesis of a linear trend;
- Determine the constant and slope of the line corresponding to the initial series (taking into account a linear trend);
- The values in the 90% quantile (higher extreme values) are replaced by the values calculated in the previous point using the two parameters;
- The values in the 10% quantile were not replaced because they are very close to the rest of the series' historical observations;
- The final series is smoothed using a moving average calculated over the previous four quarters.

The graph for the default rate variable before and after smoothing, which shapes the data series so that potential differences can be observed, is shown below.

Figure 1. The original default rate and after smoothing



Source: Owner processing

As can be seen, there were no significant fluctuations in the original data series. In 2010-2011, there was a period of expansion. The series did not change significantly after the smoothing. As a result, the analysis was carried out on the original series.

3.1. General data presentation

The database was compiled by the National Bank of Romania and Eurostat and comprises quarterly data from March 2008 to March 2019. We opted for this time span because it was the full-time span that was available at the time of the analysis, in addition to the fact that we wanted to collect as many observations as possible and perform a relevant analysis.

The following are the database's primary fields: *start_window* (the snapshot when the credit is observed), *DR* (default rate), *PIB* (gross domestic product), *UR* (unemployment rate), *HICP* (consumer price index).

The dependent variable is the 12-month default rate observed between March 31, 2008, and March 31, 2019. The series is released on a quarterly basis. The 12-month default rates were calculated over as long a time period as possible in order to capture at least one economic cycle and to have a time series with as many observations as possible from which to derive the most relevant regressions from both a statistical and a business perspective. The independent variables are a historical sequence of macroeconomic parameters relating to Romania.

3.2. Variable transformations

The following transformations were used on the series of default rates: logit, 1st order differentiation, and cube root.

The following treatments were performed on the series of independent variables, one by one: 1st and 2nd order differentiation, lag 1 and lag 4 (we moved the variable in the past from time t by 1 and 4 quarters, anticipating that the impact of macroeconomic shocks would be integrated into default rates with a delay), growth rate (the variable's growth rate at time t was calculated in comparison to the same period the previous year).

4. Research results and comments

The model selected for predicting default rates is the one made up of the independent variables consumer price index and gross domestic product, as well as the dependent variable DR, for which a cubic transformation was performed. Lag4 and lag1 transformations were used for the two independent variables.

The following is the model equation on which the forecasts will be based:

$$DR\_cube = 0.352 + 0.752 \cdot HICP\_lag4 - 0.669 \cdot PIB\_lag1 \quad (2)$$

The predictions obtained using equation (2) can be seen in Table 1 below.

Table 1. Predicted outcomes for default rates

Period	DR Predictions
30/06/2019	4.46%
30/09/2019	4.62%
31/12/2019	4.61%
31/03/2020	4.40%
30/06/2020	4.92%
30/09/2020	8.93%
31/12/2020	7.18%
31/03/2021	5.69%
30/06/2021	5.08%
30/09/2021	1.91%
31/12/2021	3.52%
31/03/2022	4.70%
30/06/2022	3.68%
30/09/2022	4.30%

Source: Owner processing

The following table shows the performance measures and checklist validation for this model.

Table 2. Performance indicators of the regression model

P-value	Adjusted R^2	AIC	The maximum correlation coefficient (module)
$3.053 \cdot 10^{-13}$ (< 0.05)	73.44%	-194.68	0.34

Source: Owner processing

The model's coefficients are statistically significant, and the model's overall performance is relatively good when compared to the estimated performance metrics.

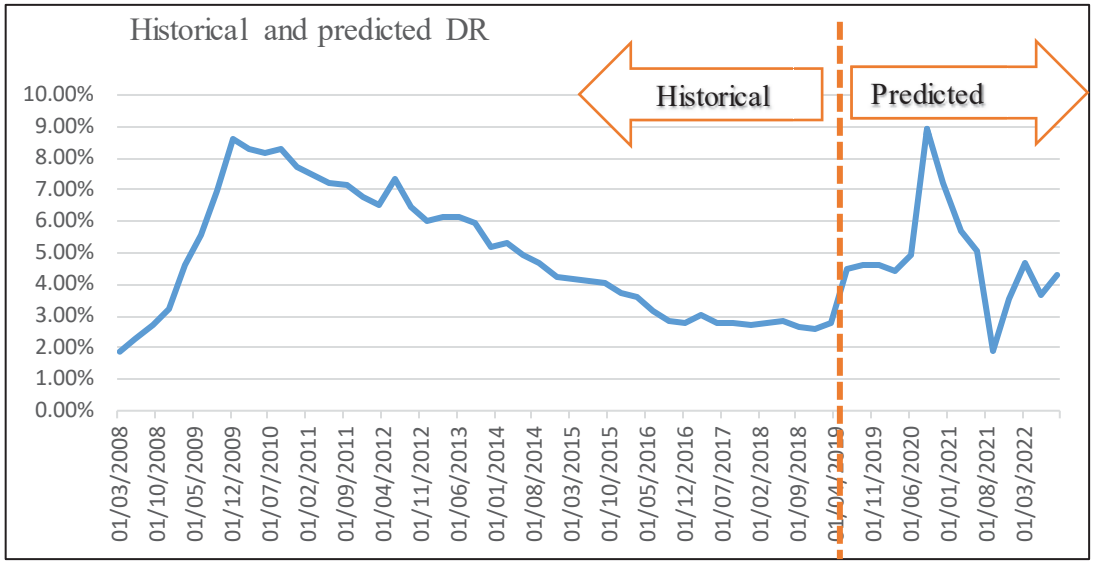
The default rate values will be forecasted using this equation for the time period of June 2019 to September 2022. The rationale behind selecting this time frame is that the regression model's underlying assumption is that the time perimeter for prediction should be constrained rather than very wide. Capturing the impact of the COVID-19 pandemic on default rate values was a second objective.

The predictions obtained using equation (2) are reasonable and appropriately capture the trend. The default rates climbed during the pandemic of 2020 as unemployment also increased, people had trouble making loan payments, some of them lost their jobs, and rates then started to decline. The lowest rate was recorded at the end of September 2021 (1.91%), and the highest rate was recorded in September 2020, at the height of the COVID-19 pandemic (8.93%).

Additionally, using these forecasts, the yearly adjustment factors k were derived as a ratio between the average of the forecasted values for the previous four quarters and the average of the historical default rates during the preceding three years: $k_1 = 0.59$, $k_2 = 0.70$, $k_3 = 0.39$.

The evolution of these rates can also be studied using Figure 2.

Figure 2. Historical and predicted default rates



Source: Owner processing

5. Conclusion

Linear regression is a widely used model that may be used to forecast specific indicators and discover correlations between various variables. It is also regarded as the cornerstone of statistical learning theory.

In this paper, we used multiple linear regression to estimate default rates and discovered a link between these variable and other macroeconomic variables. Also, using the forward-looking technique, we produced some adjustment factors based on the expected default rate values, which will be applied to the probability of loan non-repayment.

The chosen model outperforms the calculated indicators and captures the impact of macroeconomic factors (gross domestic product and consumer price index) in estimating default rates, which characterize the behavior of a banks' customers in relation to their loans. The estimated default rates are in line with expectations, with higher values for the pandemic period.

Future research directions may focus on applying particular models – like ARIMA (Autoregressive Integrated Moving Average), SARIMA (Seasonal Autoregressive Integrated Moving Average), or VAR (Vector Autoregressive) – to time series while accounting for the data set used in the analysis. An assessment of the models' performances and a relevant choice regarding the final model can result from observing any potential differences between the time series models specified previously and the classical regression model.

The data's accessibility could be one of the article's potential limitations. The entire set of data from 2008 to 2019 is analysed. The use of the regression model, one of the fundamental statistical models that is occasionally surpassed by the functionality of more recent and sophisticated models, can also be seen as a limitation. For the purposes of this article's analysis, nevertheless, it works.

Banks' efforts to research these methods and models and use them to compute the likelihood that loans from clients will default (probability of default) can be seen as reflecting the policy implications of this paper.

***Authors' contribution**: Introduction, I.F.R.; Literature review, S.S.; Methodology and data, I.F.R.; Research results and comments, A.P.; Conclusion, A.P.*
***Acknowledgement**: We are grateful to the editors of this journal for inviting us to publish the article. This work was supported by the Bucharest University of Economic Studies/Research Institute.*

References

Baltazar, J., Reis, J., Amorim, M. (2020). Sustainable economies: Using a macro-economic model to predict how the default rate is affected under economic stress scenarios. Sustainable Futures, 2, 100011, ISSN 2666-1888, https://doi.org/10.1016/j.sftr.2020.100011

Guender, A.V. (2002). Optimal and efficient monetary policy rules in a forward-looking model. Journal of Macroeconomics, 24(1), 41-49, ISSN 0164-0704, https://doi.org/10.1016/S0164-0704(02)00014-9

Huang, S. (2023). Linear regression analysis. International Encyclopedia of Education, 4th edition, 548-557

Kamada, K., Muto, I. (2000). Forward-looking Models and Monetary Policy in Japan, Research and Statistics Department, Bank of Japan

Li, C., Yan, Y., Liu, X., Wan, S., Xu, Y., Lin, H. (2023). Forward looking statement, investor sentiment and stock liquidity. Heliyon, 9(4), e15329, https://doi.org/10.1016/j.heliyon.2023.e15329

Lin, M., Chen, J. (2023). Research on Credit Big Data Algorithm Based on Logistic Regression. Procedia Computer Science, 28, 511-518, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2023.11.058

Peláez, R., Van Keilegom, I., Cao, R., Vilar, M.J. (2024). Probability of default estimation in credit risk using mixture cure models. Computational Statistics & Data Analysis, 189, 107853, ISSN 0167-9473, https://doi.org/10.1016/j.csda.2023.107853

Rasyidah, Efendi, R., Nawi, N.M., Deris, M.M., Burney, S.M.A. (2023). Cleansing of inconsistent sample in linear regression model based on rough sets theory. Systems and Soft Computing, 5, 200046, https://doi.org/10.1016/j.sasc.2022.200046

Wattanawongwan, S., Mues, C., Okhrati, R., Choudhry, T., Chi So, M. (2023). Modelling credit card exposure at default using vine copula quantile regression. European Journal of Operational Research, 311(1), 387-399, ISSN 0377-2217, https://doi.org/10.1016/j.ejor.2023.05.016

Zhou, B., Jin, J., Zhou, H., Zhou, X., Shi, L., Ma, J., Zheng, Z. (2023). Forecasting credit default risk with graph attention networks. Electronic Commerce Research and Applications, 62, 101332, ISSN 1567-4223, https://doi.org/10.1016/j.elerap.2023.101332