954:596 Regression Time Series

Fall2023

Title: "Regression Analysis on Household Victimization" Subtitle: "A Comparative Study between Mortgage Crisis (2008) and COVID Crisis Validation (2018)"

Group 10:

Udayveer Singh Andotra – ua118

Arun Mishra - am3464

Ojas Sharma – os273

Introduction and dataset

Household victimization during crises represents a critical area of study, providing insights into the vulnerabilities and resilience of communities when faced with significant challenges. This project delves into the dynamics of household victimization with a specific focus on two distinct crises – the Mortgage Crisis of 2008 and the COVID Crisis in 2018. By undertaking a comparative analysis, we aim to construct a robust regression model that not only captures the intricacies of household victimization but also evaluates the model's performance across different crisis periods.

The significance of this study lies in its dual objectives: first, to develop a comprehensive understanding of household victimization dynamics during crises, and second, to assess the adaptability and predictive power of the model over time. By examining both crisis and non-crisis periods, our research contributes to the identification of patterns and responses that can inform future policy-making and enhance community resilience.

This project not only sheds light on the specific factors influencing household victimization but also provides a methodological framework for future studies in this domain. The practical implications of our findings extend to policymakers, law enforcement agencies, and community organizations working towards mitigating the impact of crises on household security. Through this project, we aim to contribute valuable insights to the field of criminology, fostering a deeper understanding of household victimization dynamics in the face of crises. The dataset utilized for this project spans from 1992 to 2022 and is sourced from the National Crime Victimization Survey (NCVS): ICPSR 38604. . It is a collection year basis, containing records from interviews conducted in the 12 months of each given year from 1992 to 20222 and has 4,189,318 entries or observations with 226 total columns or variables.

Variable Selection

A pivotal step in the analysis is the careful selection of variables, ensuring that the model incorporates the most influential factors shaping household victimization.

Response Variable: VFLAG (Household Victimization Flag). Based on the outcome we shortlisted 25 variables which could help establish a relation with the response.

The following variables were selected.

Regressor:

V2135: 1ST PERSON'S HH COMPOSITION CODE, V2132: PRINCIPAL PERSON ATTENDING SCHOOL, V2129: MSA STATUS, V2122: FAMILY STRUCTURE CODE, V2121B: INDIAN RESERVATION/LANDS, V2120: PUBLIC HOUSING, V2119: COLLEGE/UNIVERSITY, WGTHHCY: ADJUSTED HOUSEHOLD WEIGHT, V2107: NEIGHBORHOOD WATCH GROUP, V2106: HOUSEHOLD DEVICES AGAINST INTRUDERS, V2006: HOUSEHOLD NUMBER, V2143: URBANICITY, V2025: DIRECT OUTSIDE ACCESS, V2025B: BUILDING WITH RESTRICTED ACCESS, V2026: HOUSEHOLD INCOME, V2031: RACE OF HH HEAD, V2033: PRINCIPAL PERSON AGE, V2034: PRINCIPAL PERSON MARITAL STATUS, V2036: PRINCIPAL PERSON SEX, V2038:

PRINCIPAL PERSON EDUCATIONAL ATTAINMENT, V2043: REFERENCE PERSON MARITAL STATUS, V2046: REFERENCE PERSON NOW IN ARMED FORCES, V2073: NO. CRIME INCIDENT REPORTS, V2074: OPERATE BUSINESS FROM ADDRESS, V2077: NO. TIMES BROKEN IN OR ATTEMPTED, V2078: NUMBER MOTOR VEHICLES OWNED, V2080: NO. TIMES MOTOR VEHICLE THEFT

Response:

VFLAG: HOUSEHOLD VICTIMIZATION FLAG

Data Cleaning and Preparation

Dataset was divided into four distinct subsets, with two capturing periods before and after the 2008 mortgage crash and the other two capturing crisis periods, namely the mortgage crisis and the COVID crisis.

A simple, factor to numeric conversion was done and a response (VFlag) was derived with levels 0 and 1 rather than the original (VFLAG) with levels 1 and 2.

NA responses were dropped to keep the efficacy of model and to avoid biased results.NA observations were replaced with 0 for regressors to avoid errors while model building.

A few variables had a large majority of values as NA namely, (V2120, V2107, V2106, V2025, V2031, & V2046), those were dropped before proceeding with linear regression since these would be insignificant in drawing a linear relation.

Multiple Regression Model

A linear regression model was tried to fit with the objective to examine whether a linear relationship exists between the selected regressors and the response variable (VFlag).The best-fit model is derived through backward stepwise regression, employing a selection criterion of prem=0.05.. The null hypothesis underlying this test posits that the distribution of residuals follows a normal distribution. This alternative testing approach was adopted to assess the normality of the residuals, a crucial assumption for linear regression models and with p value significantly lower than the significance level of 0.05, null hypothesise was rejected implying residuals are not normaly distributed. The normality of residuals was assessed using Q-Q plot, the residuals do not conform to a straight line.

After fitting Linear Regression Model and performing required tests as mentioned above, following observations were made:

- The fitted values were not bound to the range 0 to 1 which is needed for binary response.
- The residuals fail the normality test (Anderson Darling Test) and are notably heteroscedastic. with p-value significantly low then the threshold value of 0.05, null hypothesis (Residuals' distribution is normal) can be hold comfortably.

- The Q-Q plot also points towards non-normality of the residuals.

Consequently, the violations of Ordinary Least Squares (OLS) assumptions were identified, necessitating the adoption of a more appropriate regression technique. Subsequently, classification model was selected.

Logistic Regression Model

Logistic regression model was chosen as the classification model. The created binary variables which were significant in SLR replaced their respective categorical counterparts, while the remaining categorical variables are included in their original form. The likelihood ratio test was conducted to compare the null model and the full logistic regression model. The chi-squared statistic for the test was 52857.96 with 29 degrees of freedom, resulting in a p-value of 0. As the p-value is less than the significance level of 0.05,null hypothesis was rejected, indicating that the full model is statistically more significant than the null model. The summary of the logistic regression model reveals several significant and non-significant coefficients. Notably, the intercept is highly significant (p < 2e-16), suggesting a strong association between the predictors and the binary response variable (VFlag). Among the predictors, variables such as MSA (Metropolitan Statistical Area) and V2073 exhibit significance at conventional levels (p < 0.05), indicating their substantial impact on the response. Furthermore, certain variables like V2119 (a categorical variable related to education), motorGe1, and V2043(2) (indicating the widowed status) are statistically significant with associated p-values less than 0.05. These variables play a crucial role in predicting household victimization. On the other hand, some variables, such as V2135 and V2006, do not exhibit statistical significance. In terms of model performance, the null deviance (=74623) and residual deviance (=21765) were obtained. The AIC (Akaike Information Criterion) is 21825, serving as a measure of the model's goodness of fit, considering the trade-off between the complexity of the model and its ability to explain the data. Additionally, binary variables for the significant variables (V2119 and V2073) and potential variables (V2025B, V2026, V2006, V2034, and V2036) were identified.

Stepwise Logistic and Cross Validation

In the process of constructing a predictive model, binary variables, such as 'noUniF' (indicating no university attendance) and 'crimInc' (signifying the occurrence of a criminal incident), noResAc' (denoting no restricted access to the household) and 'repHous' (indicating the occurrence of a replacement household), 'IncGe35k' (representing a binary variable for household income greater than $35,000), 'widowed' (indicating whether the primary person is widowed or not), and 'female' (denoting the gender of the primary person as female) were systematically created based on logistic significance. In the stepwise logistic regression analysis, a model was developed to predict the response variable (VFlag) based on several predictor variables. The selection process involved backward elimination using the Akaike Information Criterion (AIC) as the criterion for variable inclusion or exclusion. The initial model included 16 variables, and through a series of steps, the algorithm systematically removed less significant variables. Ultimately, seven variables—noUniF, famstr2, IncGe35k, widowed, female, crimInc, and motorGe1—were retained in the final model.

Correlation matrix was computed for the selected independent variables, including noUniF, famstr2, IncGe35k, widowed, female, crimInc, and motorGe1. The results indicated low correlations among these variables, suggesting independence. Subsequently, a logistic regression model was developed using a 5-fold cross-validation approach to predict the response variable VFlag. The model's performance was assessed through various metrics, namely as root mean square error (=0.11), R-squared(=0.76), and mean absolute error (=0.027),the variance inflation factors (VIFs) were also calculated, confirming the absence of multicollinearity among the predictors.

Accuracy

The model exhibited remarkable accuracy across different datasets, with 98.31% on the training set, 97.55% before the crisis, and 99.04% after the crisis. Validation on an additional dataset yielded a high accuracy of 99.23% (see Fig 7.8). These results affirm the logistic regression model's reliability and effectiveness in predicting VFlag outcomes across various time periods.

Comparison of Coefficients

Logistic regression model was fitted to the two individual test datasets using the identical set of variables that had been initially selected and trained upon in the training dataset. By employing the same variables, the goal was to assess the model's performance and its ability to generalize to new data, particularly in predicting household victimization patterns during distinct time periods represented by the test datasets.

For the "before crisis" period, the coefficients indicated that households with certain characteristics, such as being widowed or having a higher income, were more likely to experience victimization. In contrast, during the "crisis" period, attending university appeared to increase the likelihood of household victimization, reflecting a reversal in the impact of this variable.

Conclusion

Examining crisis impact on household victimization, two key factors stand out: "noUniF" (main person attended university) and "IncGe35K" (household income > $35,000). Surprisingly, during crises, university attendance correlates with a higher risk of victimization, contrary to the norm. Outside crises, high-income households face an elevated risk. This insight informs targeted interventions for diverse households in crisis situations.

Secondly, Linear regression proved inadequate for modeling the binary response variable due to the unbounded nature of the fitted response.

The assumptions of OLS for residuals were violated so it was in best interest to use a classification model for a categorical response.

The logistic regression model was robust as it performed well in all the four sets. Accuracy values test0: **97.55%,** train : **98.31%,** test1: **99.04%,** val : **99.23%**

Reference:

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7665755/

https://www.cmu.edu/tepper/faculty-and-research/assets/docs/covid-19-mortgage-meltdown.pdf

https://www.degruyter.com/document/doi/10.1515/erj-2022-0366/html?lang=en

https://www.mdpi.com/1911-8074/15/8/371

https://www.researchgate.net/publication/353416773_A_comparative_analysis_of_COVID-19_and_global_financial_crises_evidence_from_US_economy

https://www.atlanticcouncil.org/blogs/new-atlanticist/can-we-compare-the-covid-19-and-2008-crises/

https://www.urban.org/urban-wire/understanding-differences-between-covid-19-recession-and-great-recession-can-help-policymakers-implement-successful-loss-mitigation