Proceedings of the 4th International Conference on Applied Physics and Mathematical Modeling
DOI: 10.54254/2753-8818/2026.37015

Customer Purchase Frequency Modeling Using Transactional Data and Regression

Rucong Liu

Faculty of Science and Engineering, University of Nottingham Ningbo China, Ningbo, China
Smyrl10@nottingham.edu.cn

Abstract. This review examines regression models for customer purchase frequency using transaction data. It explains how transaction records are aggregated into count outcomes and behavioral predictors, including recency, frequency, monetary value, promotion exposure, and inter-purchase intervals. The review compares ordinary least squares, Poisson regression, negative binomial regression, zero-inflated models, hurdle models, and the Conway-Maxwell-Poisson distribution. It shows that overdispersion, excess zeros, customer heterogeneity, observation-window design, and unequal exposure periods strongly affect model choice and interpretation. Representative studies indicate that negative binomial and two-part count models often fit purchase data better than simple Poisson or linear models, although results depend on data construction and validation procedures. The review also summarizes common evaluation measures, including mean absolute error, root mean squared error, deviance, and information criteria. Current limitations include inconsistent definitions, censoring, temporal leakage, non-stationarity, privacy concerns, and limited causal interpretation. Future research should develop dynamic and seasonal count models, Bayesian uncertainty estimation, causal analysis, and interpretable machine-learning approaches.

Keywords: purchase frequency, transaction data, count regression, zero-inflated models, customer analytics

1. Introduction

Purchase frequency, defined as the number of transactions made by a customer during a specified observation period, is a valuable behavioral signal available to retailers, e-commerce firms, subscription services, and digital platforms. It is associated with customer demand, engagement, and customer lifetime value, and it provides evidence for inventory planning, customer management, and marketing decisions [1, 2]. Modern point-of-sale systems and online shops routinely collect transaction data with a customer identifier, transaction timestamp, product category, purchase amount, and promotion indicators. From these records, analysts can derive customer-level measures such as purchase counts, inter-purchase intervals, and recency, frequency, and monetary value (RFM), which summarize habitual purchasing behavior [3, 4]. Purchase frequency is a non-negative integer count that is often right-skewed. When the analytical cohort includes registered or otherwise eligible customers who make no purchase during the observation window, the outcome may also

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2753-8818/2026.37015

contain many zeros. These properties can make ordinary linear models inappropriate and motivate count-data regression approaches that account for the mean-variance relationship, overdispersion, and excess zeros [5, 6].

Research on modeling purchase frequency from transaction data has evolved considerably over the past decades. Early studies described repeat purchasing mainly through RFM summaries and simple distributional assumptions, and RFM remains a common way to transform transaction records into customer-level profiles [3, 4]. As transactional datasets became more widely available, Poisson regression became a common baseline for count outcomes. However, its equidispersion assumption is often violated in purchasing data, which are typically overdispersed. The negative binomial model addresses overdispersion by introducing a heterogeneity parameter. More flexible count models and count-based specifications can outperform simpler baselines in specific retail and online settings [7, 8]. When many customers make no purchases during the observation window, zero-inflated Poisson and zero-inflated negative binomial models treat zeros as arising from two latent processes, whereas hurdle models separate the decision to purchase from the number of purchases conditional on a positive count [5, 6]. More recently, flexible count distributions such as the Conway-Maxwell-Poisson distribution have been used to model both underdispersion and overdispersion in customer transaction counts [9]. Existing studies also differ in their definitions of the observation window, the purchase event, active customers, and zero-purchase samples. These choices affect the mean-variance relationship, the prevalence of excess zeros, the representation of customer heterogeneity, and comparisons of model assumptions, parameter interpretation, predictive power, and data requirements [1, 10].

Against this background, the present paper examines regression modeling of customer purchase frequency based on transaction data. Its aim is to clarify how raw transaction data are converted into a count outcome and a set of explanatory variables, and how these variables are used to examine the main factors associated with repeat purchasing. The rest of the paper is organized as follows. Section 2 outlines basic concepts of transaction data and purchase frequency. Section 3 reviews linear and count regression methods and common evaluation measures. Section 4 surveys representative empirical studies and discusses their data, models, and findings. Section 5 discusses limitations of current studies and identifies promising future directions, including dynamic and seasonal count models, Bayesian estimation, and machine-learning-based prediction [11, 12]. Section 6 concludes the paper. Overall, this review explains how probability distributions, parameter estimation, and regression analysis support interpretable, data-driven research on customer purchase frequency [1, 5].

2. Basic concepts of transactional data and purchase frequency

Transactional data are event-level records generated by customers' purchases. A representative record contains the customer's identifier, the time of the transaction, product category, order amount and discount or promotion information. To transform the raw records into a regression-ready dataset, an observation window for the period of interest should first be set, and events should then be aggregated at the customer level. Within this window, the purchase frequency is the number of separate purchases made by a customer, and the time between consecutive purchases is called the inter-purchase interval. The above indicators are closely related to the traditional RFM model, which considers the recency of a customer's last purchase, how often they purchase, and how much money they have spent [3]. Recently, some extensions have added behavioral attributes to better reflect the regularity of shopping and price sensitivity, such as basket depth, campaign share, and the standard deviation of intervals between orders [4]. Customer attributes, product-category mix, order amount,

DOI: 10.54254/2753-8818/2026.37015

discount usage, promotional exposure, prior-purchase features, and these behavioral attributes can be used as candidate explanatory variables in a purchase-frequency regression [1, 4].

Purchase frequency is considered a count variable and is not modeled as a continuous outcome. Counts are discrete, bounded below by zero, and often strongly skewed; therefore, methods that assume continuous data may produce inaccurate predictions or inferences unless modified for counts [5]. A sample should be constructed accordingly. The length of the observation window will affect both the maximum observable frequency and the proportion of zero counts. A smaller window will result in more zeros and change the degree of overdispersion. The level of aggregation (order lines, orders or shopping baskets), the handling of duplicate and returned orders, and the treatment of missing records will also affect the resulting distribution. Different studies have defined an "active customer" in various ways, such as requiring at least one repeat purchase or explicitly modeling dropout; thus, both the sample composition and the interpretation of purchase frequency vary [2, 10].

The data construction choices should be explicitly reported to enable meaningful comparisons of regression results in other studies.

3. Regression methods for customer purchase frequency modeling

A natural starting point is ordinary least squares linear regression, which is easy to estimate and interpret. However, applying ordinary least squares to count outcomes can be problematic because predictions may be negative, the error variance is often non-constant, and standard linear-model inference may be unreliable when the outcome is highly skewed and zero-heavy [5]. Poisson regression is the standard generalized linear model for counts, with the log of the expected purchase count related to a linear predictor of customer and transaction features. Poisson regression assumes equidispersion; that is, the conditional mean equals the conditional variance. In purchasing data, unobserved customer heterogeneity often causes overdispersion. The negative binomial model addresses this problem by adding a heterogeneity parameter that allows the variance of purchase counts to be larger than the mean [7]. If there are too many zeros, a two-part count model may be employed. Zero-inflated models combine a structural-zero process with a count distribution, and hurdle models separately model the probability of any purchase and the positive purchase count using a zero-truncated count distribution [5, 6]. For transaction counts that show either overdispersion or underdispersion, the Conway-Maxwell-Poisson distribution can also be used [9].

Distributional selection is not the end of the problem. Feature selection is required because the transaction data may produce many candidate covariates. Multicollinearity among RFM-type variables increases standard errors; therefore, when many predictors are included, regularization, such as ridge or lasso penalties, can stabilize the estimation. Differences in the time of customer observation, such as different registration dates, should be handled by exposure offsets, and frequency is modeled relative to the time at risk. Reasonable indicators should be selected for the count model. Mean absolute error and root mean squared error are measures of predictive accuracy in the original units, and root mean squared error gives more weight to large errors. Poisson deviance indicates how well the model fits the data compared with a saturated model, and the Akaike information criterion and Bayesian information criterion balance goodness of fit with model complexity [5, 12]. Together, the above measures offer a system for comparing predictive error, distributional fit and parsimony.

DOI: 10.54254/2753-8818/2026.37015

4. Recent studies of transaction-based purchase frequency modeling

Empirical studies of purchase frequency differ widely in data sources, customer and product populations, observation windows, target definitions, feature construction, distributional assumptions, validation designs, and evaluation metrics. One line of research compares stochastic count models of repeat buying using customer transaction databases. Ma and Büschken proposed an "always a share" specification and tested its purchase-frequency predictions against the classical negative binomial distribution, Pareto/negative binomial distribution, and beta-geometric/negative binomial distribution models, showing that conclusions about repeat buying depend on the assumed heterogeneity structure [10]. Simon conducted a similar comparison on simulated and empirical transaction data, testing Pareto/negative binomial distribution purchase forecasts obtained from Markov chain Monte Carlo, maximum likelihood, and heuristic estimation, and found that forecast accuracy depends on the estimation method and prediction task. Studies focusing directly on purchase-frequency distributions reach similar conclusions [1]. Trinh et al. showed that a Poisson log-normal model fitted category-level purchase counts better than the classical negative binomial distribution, demonstrating how assumptions about customer heterogeneity can change substantive conclusions [7].

A second stream studies e-commerce and platform settings, where transaction logs are richer but zeros and dispersion problems can be more difficult to handle. Sheikh and Kaur benchmarked linear, Poisson, and random-forest regressions for customer purchase frequency, showing that count-based specifications fit purchase-frequency data better than linear models and that behavioral features carry more predictive power than demographic ones [8]. Mzoughia, Borle, and Limam estimated a Conway-Maxwell-Poisson model for customer transaction counts using Markov chain Monte Carlo, explicitly accommodating both overdispersion and underdispersion that simpler Poisson and negative binomial specifications cannot represent [9]. Yang, Huo, Li, Xi, and Liu fitted a zero-inflated negative binomial regression to live-streaming commerce data, using the two-part structure to represent overdispersion and excess zeros among customers who do not buy [6]. Glady, Lemmens, and Croux jointly modeled transaction timing, transaction spending, and customer dropout behavior across four customer transaction databases, showing that purchase frequency cannot be cleanly separated from inter-purchase timing and latent dropout behavior [2].

Together, these studies point to a general analytical pipeline: clean customer-level transaction records, define an observation window, construct customer features and the purchase-count outcome, and compare candidate models using multiple fit and predictive metrics [1, 8]. The comparisons also reveal why conclusions vary across studies. The length of the observation window alters the proportion of zeros. Definitions of a purchase event and an active customer alter the effective sample. Distributional assumptions determine the mean-variance relationship being modeled, while validation design determines whether in-sample fit or holdout forecasting drives model selection [7, 10]. Across applications, the methodological core consists of count distributions, count regression, parameter estimation, checks for overdispersion and zero inflation, and multi-metric assessment [8, 9].

5. Limitations and prospects

Despite important advances, research on purchase-frequency modeling has several limitations. First, core definitions such as purchase event, active customer, and observation window differ across retail, e-commerce, and subscription settings, so results are often not directly comparable across studies [2, 10]. Second, observation-window design creates unequal exposure and censoring or

DOI: 10.54254/2753-8818/2026.37015

truncation issues. Customers who enter late or leave early contribute different observation times, and poorly designed features - for example, using information from the full period to predict behavior within that same period - can introduce temporal leakage and inflate apparent predictive accuracy. Third, transactional data are often sparse and dominated by zeros. Although zero-inflated and hurdle models address this feature, their two-part structure depends on assumptions about why zeros occur that can be difficult to verify empirically [5]. Fourth, purchase patterns are non-stationary because seasonality, price changes, and promotional campaigns can shift buying rates over time. Static count regressions may therefore become misspecified when purchasing rates change [11]. Fifth, distributional assumptions may fail in ways that standard diagnostics do not capture. The Poisson mean-variance restriction is only the most obvious example [7]. Sixth, customers differ in ways that are only partly observable, and unobserved heterogeneity can confound estimated effects of marketing variables [2]. Seventh, publicly available datasets may differ systematically from real business environments, while granular transaction data raise privacy and fairness concerns that remain insufficiently examined.

These limitations lead to several promising research directions. Dynamic and seasonal count models can relax the stationarity assumption and allow purchase propensities to change with seasons, prices, and promotions. Related work on customer purchase timing has used transformer-based models to represent long-term temporal dependencies and time-series patterns, including trends and seasonality [11]. Bayesian approaches, can quantify parameter uncertainty and propagate it to customer-level predictions, while joint models of timing, spending, and dropout provide a principled way to represent partially observed heterogeneity. Machine-learning predictors such as random forests have also been benchmarked against probabilistic models including extended Pareto/negative binomial distribution models [12]. These comparisons show that additional complexity should be justified by reproducible out-of-sample gains and accompanied by appropriate interpretation. Causal-inference frameworks can move beyond correlational regression by estimating the effects of promotions, discounts, and recommendations on purchase frequency. Uncertainty quantification, transfer across customer contexts, and interpretable prediction methods can further improve the reliability and applicability of findings across retail, e-commerce, and subscription settings. Progress in these directions can make purchase-frequency modeling a more reliable basis for decision-making [1, 11].

6. Conclusion

This paper has reviewed regression modeling of customer purchase frequency using transaction data. The analysis shows that purchase frequency should generally be treated as a count outcome whose statistical properties depend on how the observation window, customer cohort, and purchase event are defined. Model selection should therefore follow the structure of the outcome rather than rely on a single default regression method. Ordinary least squares can provide a simple descriptive baseline, but it may produce negative predictions and unreliable inference for strongly skewed, zero-heavy counts. Poisson regression is appropriate when conditional equidispersion is plausible, whereas negative binomial regression is better suited to overdispersed counts. Zero-inflated and hurdle models are useful when excess zeros have a defensible behavioral interpretation, and the Conway-Maxwell-Poisson distribution provides additional flexibility when dispersion differs from the Poisson assumption. The review also shows that data construction is as important as model choice. Observation-window length, unequal customer exposure, treatment of returns and duplicate orders, definitions of active customers, and temporal separation between predictors and outcomes can materially change the estimated purchase-frequency distribution. Evaluation should therefore

DOI: 10.54254/2753-8818/2026.37015

combine predictive error, distributional fit, parsimony, and out-of-sample validation rather than rely on one metric. The main remaining limitations are inconsistent definitions across studies, censoring and unequal exposure, temporal leakage, non-stationary purchasing patterns, unobserved customer heterogeneity, privacy concerns, and the limited causal interpretation of many predictive regressions. Future research can address these issues through dynamic and seasonal count models, hierarchical or Bayesian uncertainty modeling, carefully designed causal analyses, and interpretable machine-learning methods. Overall, rigorous purchase-frequency modeling requires consistent data definitions, an explicit match between distributional assumptions and observed count behavior, and validation procedures that reflect the intended forecasting or decision task.

References

[1] Simon, L. (2025). A generalised comparison of Pareto/NBD based forecasts using MCMC, maximum likelihood, and heuristics. *Journal of Business Economics, 95*(8), 1079–1105. https://doi.org/10.1007/s11573-025-01237-8

[2] Glady, N., Lemmens, A., & Croux, C. (2015). Unveiling the relationship between the transaction timing, spending and dropout behavior of customers. *International Journal of Research in Marketing, 32*(1), 78–93. https://doi.org/10.1016/j.ijresmar.2014.09.005

[3] Alves Gomes, M., & Meisen, T. (2023). A review on customer segmentation methods for personalized customer targeting in e-commerce use cases. *Information Systems and e-Business Management, 21*(3), 527–570. https://doi.org/10.1007/s10257-023-00640-4

[4] Ozcan, T. (2026). Customer segmentation using an extended RFM model and clustering algorithms in e-commerce. *Journal of Theoretical and Applied Electronic Commerce Research, 21*(5), Article 142. https://doi.org/10.3390/jtaer21050142

[5] Almulhim, F. A., Alharbi, R., Gemeay, A. M., Hussam, E., & Hammad, A. T. (2026). Modeling count data with zero inflation, overdispersion, underdispersion, and multicollinearity using a zero-inflated Conway–Maxwell Poisson Liu-type regression. *Journal of Statistical Computation and Simulation*. Advance online publication. https://doi.org/10.1080/00949655.2026.2676655

[6] Yang, Q., Huo, J., Li, H., Xi, Y., & Liu, Y. (2023). Can social interaction-oriented content trigger viewers' purchasing and gift-giving behaviors? Evidence from live-streaming commerce. *Internet Research, 33*(7), 46–71. https://doi.org/10.1108/INTR-11-2021-0861

[7] Trinh, G., Rungie, C., Wright, M., Driesener, C., & Dawes, J. (2014). Predicting future purchases with the Poisson log-normal model. *Marketing Letters, 25*(2), 219–234. https://doi.org/10.1007/s11002-013-9254-1

[8] Sheikh, N. M., & Kaur, T. (2026). Behavioral vs. demographic predictors of customer purchase frequency: A multi-model analytical approach. In *2026 International Conference on Artificial Intelligence, Systems, and Emerging Technologies (ICAISET)*. IEEE. https://doi.org/10.1109/ICAISET66439.2026.11541713

[9] Mzoughia, M. B., Borle, S., & Limam, M. (2018). A MCMC approach for modeling customer lifetime behavior using the COM-Poisson distribution. *Applied Stochastic Models in Business and Industry, 34*(2), 113–127. https://doi.org/10.1002/asmb.2276

[10] Ma, S., & Büschken, J. (2011). Counting your customers from an "always a share" perspective. *Marketing Letters, 22*(3), 243–257. https://doi.org/10.1007/s11002-010-9123-0

[11] Grigoraș, A., & Leon, F. (2023). Transformer-based model for predicting customers' next purchase day in e-commerce. *Computation, 11*(11), Article 210. https://doi.org/10.3390/computation11110210

[12] Yan, Y., & Resnick, N. (2024). A high-performance turnkey system for customer lifetime value prediction in retail brands. *Quantitative Marketing and Economics, 22*(2), 169–192. https://doi.org/10.1007/s11129-023-09272-x