Examining Predictive Relationship Between Intrinsic Motivation and Total Packages Contributes

Prashasti Tripathi

Teachers College, Columbia University

HUDM 50026001 Programming for Data Science

Professor Dr.Bryan Keller

December 16, 2025

Abstract

This exploratory data analysis investigated the linear relationship between a researcher's Intrinsic Motivation score (mintrinsic) and their Total Number of Packages Contributed (npkgs) to the Comprehensive R Archive Network (CRAN). A Simple Linear Regression (SLR) was performed on a sample of 841complete cases. The sample was carefully derived from an intermediate dataset of 1,087 cases following a necessary step of listwise deletion applied to the relevant variables. The analysis indicated a statistically significant positive predictive relationship ($\beta = 0.34$, $p = 0.030$) between intrinsic motivation and number of package contributions made. However, the model's explanatory power was minimal, accounting for less than 1% of the variance in the outcome ($R^2 = 0.006$). Limitations and Implications for future research have been stated.

Introduction

The sustained development and success of open-source software (OSS) ecosystems, particularly those supporting specialized tools like the R programming language and its package repository, CRAN, rely heavily on the voluntary labor of researchers and developers. Understanding the factors that motivate these contributions is a critical area of inquiry within software engineering and organizational research. The core focus of this project is to explore the role of internal psychological drive in predicting quantifiable productivity in a specialized OSS environment.

Literature Review

Research on motivation in OSS is generally grounded in Self-Determination Theory (SDT) (Deci and Ryan,2000), which posits that motivation exists on a continuum from fully extrinsic (driven by rewards or avoidance of punishment) to fully intrinsic (driven by inherent interest, challenge, and enjoyment).

Previous studies examining motivation in OSS often highlight the importance of intrinsic factors. For instance, findings frequently report that developers are motivated by intellectual curiosity, the joy of programming, and the desire to improve skills (Lakhani & Wolf, 2005). These papers typically find that intrinsic factors are the primary drivers for initial participation and sustained engagement in general open-source projects. However, the exact predictive strength of pure intrinsic motivation on measured output (like total packages) remains variable across platforms.

In contrast, other research has shown that professional rewards often become dominant predictors of high-volume contribution. For platforms like CRAN, which intersect heavily with academic and professional careers, contributions often translate into extrinsic benefits including academic citations, tenure requirements, job market visibility, and peer reputation (Hars & Ou, 2002). Papers analyzing high-volume output often find that extrinsic motives such as the necessity of publishing code for grant funding or academic papers are stronger correlates of prolific production than simple intrinsic enjoyment. This suggests a potential ceiling effect for intrinsic motivation where, beyond a certain point, professional pressures take over as the key determinant of volume. This project seeks to contribute to this literature by quantifying the actual predictive variance accounted for by intrinsic motivation in the context of high-stakes CRAN contributions.

Research Questions

This exploratory analysis is guided by the following research questions:

1. What are the descriptive properties (e.g., mean, distribution) of the Total Packages Contributed (npkgs) and Intrinsic Motivation (mintrinsic) variables within this sample of CRAN contributors?
2. Does a developer's enjoyment of the work positively predict the total number of R packages they contribute? This is the Primary Research Question.
3. Is this simple prediction model good enough, or are its results flawed? This is the Secondary Research Question.

Rationale

The rationale behind conducting this secondary research includes the following reasons:

1) Personal Drive for Community Insight: As an R user, I'm personally interested in discovering the true drivers behind the developers building tools for my community. This project is my chance to gain direct, empirical insight into the crucial link between enjoyment and contribution within the R ecosystem.

2) Testing the Foundation of Existing Knowledge: The original study's complex findings are built on the simple idea that enjoyment matters. This analysis is essential to explore and establish the precise strength of this core theoretical link using the simplest model, thereby validating the foundation of the larger research.

3) Exploring Model Sufficiency for Simplicity: Since the original authors found a complex model necessary, the core purpose of this study is to analyse if a basic, single-variable model could be sufficient. By examining the resulting R^2 and model assumptions, I will diagnostically determine if a simple approach is adequate for this real-world complexity, or if the complicated model is truly essential.

Method

Participants and Data Source

The data were sourced from a large-scale survey previously conducted among R developers and contributors. The original data underwent extensive initial processing by the researchers, including rigorous psychometric modelling (e.g., Item Response Theory) to calculate

reliable motivational scale scores. This preparatory work yielded an intermediate dataset containing 1,087 cases (The data were sourced from the RMotivation.tab dataset). The sample consisted of researchers and developers who contribute to the R ecosystem, representing a highly specialized and skilled group.

Data Cleaning and Sample Finalization

However this dataset consisted of missing values for the two variables relevant for the current research - npkgs and mintrinsic. To ensure that the analysis was conducted on a dataset with complete cases, final round of data processing was conducted. This involved performing Listwise Deletion for npkgs and mintrinsic. This step reduced the sample from 1,087 to the definitive 841 complete cases, ensuring the analysis is reproducible and based on a consistent subset of the original data. Also, this method was chosen as the most basic form of missing data handling, consistent with the beginner-level scope of this project.

Research Variables

The two variables used in this research included 2 of the original dataset's 27 variables - the continuous score for Intrinsic Motivation (mintrinsic) and the count variable for total packages contributed (npkgs). While mintrinsic served as the independent (predictor) measure, npkgs served as the dependent (outcome) measure. A description of these variables is given in Table 1 below.

Table 1
Description of variables included in the analysis

Variable name	Meaning	Range of Values
mintrinsic	Continuous Score for Intrinsic Motivation (Predictor)	[−1.72, 0.73]
npkgs	Count of total R packages contributed (Outcome)	[0, 33]

Note. The analysis is performed on the N = 841complete cases. Brackets indicate the observed numeric range.

Statistical Analysis Measures

To determine if a relationship exists between a developer's Intrinsic Motivation score and their Total Packages Contributed, the analysis started with a visual inspection of the scatterplot. This preliminary step was taken to visually assess the direction, form, and strength of the

relationship and to check for obvious distributional issues like skew or outliers. Following the initial inspection, the Primary Research Question whether a statistically significant predictive relationship exists was addressed using a Simple Linear Regression (SLR). The SLR yielded the unstandardized beta coefficient (β) and p-value. To address the Secondary Research Question concerning the model's overall utility, two additional steps were taken: the model's practical utility was determined by examining the Multiple R^2 value, and its technical integrity was rigorously assessed by visually inspecting a series of four diagnostic plots including Residuals vs. Fitted (checking linearity and homoscedasticity), Normal Q-Q (checking normality of residuals), Scale-Location (checking homoscedasticity), and Residuals vs. Leverage (identifying influential outliers) to confirm if the critical statistical assumptions were met.

Findings

Descriptive Properties of the Dataset

Descriptive Analysis of mintrinsic and npkgs for the complete cases resulted in the descriptive properties of the same as shown in Table 2.

Table 2
Descriptive Statistics for Motivation and Contribution Variables (N=841)

Variable	Type	Mean	SD	Range
Intrinsic Motivation (mintrinsic)	Continuous Score	-0.07	0.73	[−1.72, 0.73]
Total Packages Contributed (npkgs)	Count	2.82	3.35	[0, 33]

Note. SD = Standard Deviation. The mean, SD, and range are the appropriate descriptive statistics for these numeric variables.

Simple Linear Regression Analysis

The initial visual inspection of the scatterplot revealed a minimal, highly dispersed relationship between Intrinsic Motivation and Total Packages Contributed. The data points formed a wide cloud, clustered heavily near the origin, with numerous significant outliers, suggesting low predictive power and a severely skewed distribution (Figure 1). The subsequent Simple Linear Regression confirmed the visual finding - the model was statistically significant, indicating a positive prediction between intrinsic motivation and contribution ($\beta = 0.34$, $t(839) = 2.175$, $p = .030$). However, the model demonstrated negligible practical utility, as the Multiple R^2 was only 0.0056, meaning less than one percent of the variance in package contributions is

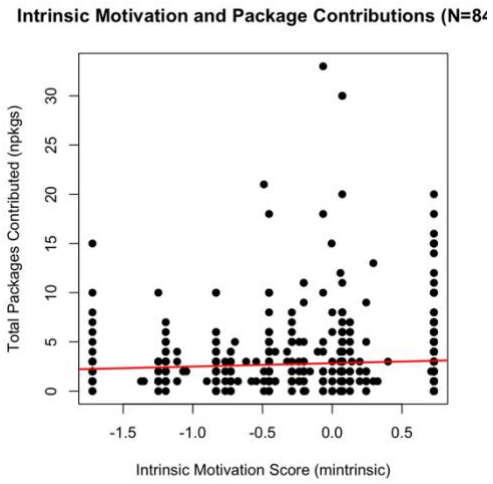
explained by intrinsic motivation (Table 3). Furthermore, as shown in Figure 2, 3 and 4, the analysis of the diagnostic plots confirmed severe technical flaws, including a critical violation of the normality of residuals and the presence of influential outliers, rendering the statistically significant p-value unreliable. These are discussed in detail in Discussion section.

Table 3
Summary of Simple Linear Regression Predicting Total Packages Contributed (N = 841)

Predictor	B	SE	t	p
Intercept	2.84	0.12	24.49	<0 .001
Intrinsic Motivation (mintrinsic)	0.34	0.16	2.18	0.030
R^2	0.006			

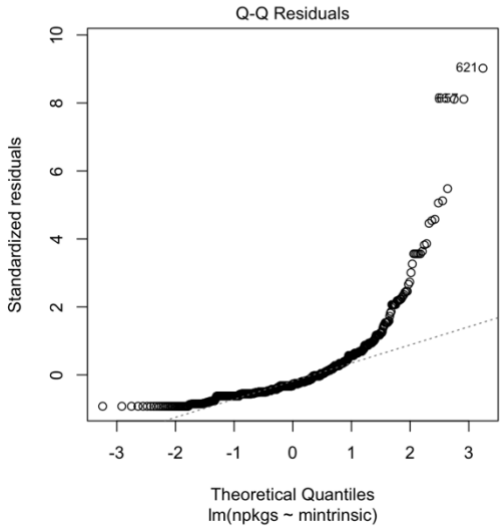
Note. N = 841. The model was statistically significant ($F(1, 839) = 4.73$, $p = 0$ .030). The beta coefficient for Intrinsic Motivation indicates a significant positive prediction, but the R^2 demonstrates extremely low predictive power.

Figure 1. Scatterplot of Intrinsic Motivation (mintrinsic) and Total Packages Contributed (npkgs)



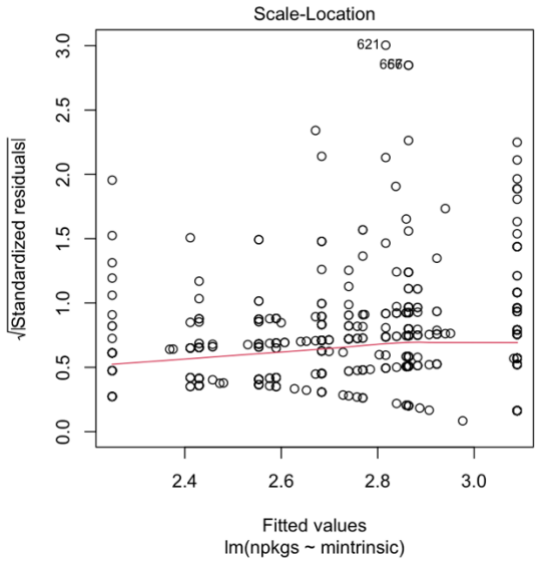
Note. This plot visually assesses the linearity and direction of the relationship. The high dispersion of points confirms the negligible practical utility (R^2 approx 0.0056), and the cluster near the origin indicates high positive skew.

Figure 2. Normal Q-Q Plot for Simple Linear Regression Model Residuals



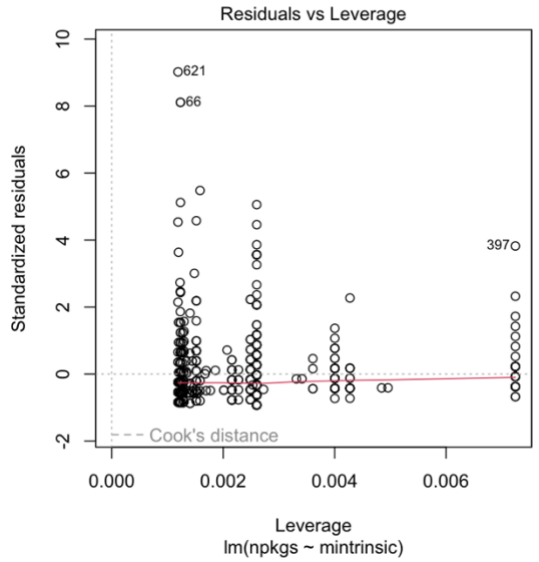
Note. This plot assesses the assumption of normally distributed residuals. The severe upward curve away from the diagonal line, especially at the high end, indicates a critical violation of the normality assumption.

Figure 3. Scale-Location Plot



Note. This plot checks the homoscedasticity assumption (constant variance). A non-horizontal red line (suggesting an increasing trend) indicates that the variance of the residuals is not constant, thus confirming heteroscedasticity.

Figure 4. Residuals vs. Leverage Plot



Note. This plot identifies influential observations. The points labeled 621, 66, and 397 fall outside the dashed Cook's Distance contours, confirming their significant influence and distortion of the regression line.

Discussion

The current study investigated the relationship between Intrinsic Motivation and Total Packages Contributed using Simple Linear Regression (SLR) to address two key questions namely the existence of a predictive relationship (Primary RQ) and the technical integrity and practical utility of the model (Secondary RQ). The findings, while revealing a statistically significant relationship, overwhelmingly suggested that the model is both practically useless and technically unsound, thereby failing to provide a reliable predictive tool.

Interpretation of Statistical Significance and Practical Utility

The SLR analysis yielded a statistically significant positive unstandardized beta coefficient ($\beta = 0.34$, $p = 0.030$), which superficially addresses the Primary Research Question by indicating that higher intrinsic motivation predicts a higher number of packages contributed. This positive, though weak, finding aligns generally with motivational theories suggesting that enjoyment derived from an activity (intrinsic motivation) can translate into greater output or persistence.

However, this statistical significance must be immediately tempered by the analysis of the model's practical utility. The Multiple R^2 value was extremely low at 0.0056, meaning that only 0.56% of the variance in package contribution can be explained by intrinsic motivation alone. Given the massive amount of unexplained variance (over 99.4%), the finding, while technically significant, has virtually no practical explanatory power for this population. This low R^2 was visually foreshadowed by the initial scatterplot, which showed data points forming a highly dispersed, wide cloud around a nearly flat regression line, indicating that knowing a developer's motivation score provides almost no reliable insight into their contribution count.

Evaluation of Model

The analysis of the diagnostic plots provided definitive evidence regarding the Secondary Research Question concerning the model's technical integrity, revealing multiple critical violations of SLR assumptions.

Violation of Normality

The Normal Q-Q Plot showed a severe departure of the standardized residuals from the theoretical diagonal line, particularly at the extreme positive end, confirming a major violation of the assumption of Normality of Residuals. This skew is inherent to the nature of the dependent variable (npkgs), which is a positively skewed count variable (packages contributed), with a concentration of zero and low values and a long tail of high-contributing outliers. The violation of this assumption invalidates the calculated p-values, meaning the conclusion that the relationship is “statistically significant” ($p = .030$) cannot be trusted.

Heteroscedasticity and Outliers

Further flaws were evident in the data distribution. The scatterplot visually demonstrated Heteroscedasticity (non-constant variance), where the spread of the residuals is much wider for higher predicted values than for lower ones, a pattern that would also be confirmed by the Scale-Location plot. This means the model's error rate is inconsistent across the entire range of package contributions.

Moreover, the Residuals vs. Leverage Plot confirmed the presence of highly influential outliers (e.g., cases 621, 66, and 397). These individuals, who contributed a disproportionately large number of packages, exert an excessive pull on the regression line,

artificially inflating the β coefficient and leading to a false sense of a statistically meaningful result. The low R^2 and the severe influence of these few data points confirm that the model's output is driven more by the presence of extreme contributors than by a stable, generalizable linear relationship across the population.

Limitations and Future Research

The primary limitation of this study is the inadequacy of the SLR model for analyzing this specific data structure. A simple linear model is inappropriate for highly skewed count data with non-normal residuals and clear heteroscedasticity.

Future research should focus on employing more appropriate statistical techniques. The highly skewed count nature of the dependent variable (npkgs) strongly suggests that a Generalized Linear Model (GLM), such as a Poisson regression or a Negative Binomial regression (to account for overdispersion), would be technically necessary. Alternatively, transforming the dependent variable (e.g., using a logarithmic transformation) could normalize the distribution and stabilize the variance, although interpretation of the resulting beta coefficients can become less intuitive. Finally, future studies should explore the addition of other predictors (e.g., skill level, tenure, company support) in a Multiple Regression framework to account for the 99.4% of unexplained variance.

Conclusion

In conclusion, while the Simple Linear Regression suggested a statistically significant positive link between intrinsic motivation and package contribution, the model's extremely low practical utility ($R^2 = 0.0056$) and its numerous, critical technical flaws demonstrated by the diagnostic plots render this finding unreliable and the model unsuitable for prediction or explanation. Future analysis must utilize models appropriate for count data to draw valid conclusions about the relationship between intrinsic motivation and developer output.

References

Deci, E. L., & Ryan, R. M. (2000). The "what" and "why" of goal pursuits: Human needs and the self-determination of behavior. *Psychological Inquiry*, *11*(4), 227–268.

Hars, A., & Ou, S. (2002). Working for free? Motivations for participating in open-source projects. *International Journal of Electronic Commerce*, *6*(3), 25–39.

Lakhani, K. R., & Wolf, R. G. (2005). Why hackers do what they do: Understanding motivation and effort in free/open source software projects. In J. Feller, B. Fitzgerald, S. Hissam, & K. R. Lakhani (Eds.), *Perspectives on free and open source software* (pp. 3–21). MIT Press.

Mair, P., Hofmann, E., Gruber, K., Hatzinger, R., Zeileis, A., & Hornik, K. (2015). Motivation, values, and work design as drivers of participation in the R open source project for statistical computing. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(48), 14788–14792.