GLM and Computational Statistics

Tarek SAADE

January 2026

1 Modeling Micro-Credit Uptake in Rural Communities Using GLM

1.1 Introduction and Scientific Question

Micro-credit programs aim to stimulate economic development by improving access to financial resources in rural communities. However, the effectiveness of such programs depends on more than the number of informational meetings organized. Differences in human capital, structural accessibility, and economic capacity may strongly influence whether information is converted into concrete action.
This study analyzes data from 200 rural villages that participated in a micro-credit outreach program. The primary scientific question is:

> *To what extent do **literacy** rates and local **infrastructure** influence the conversion of informational **meetings** into formal loan **applications**?*

To answer this question, we model the number of loan applications submitted per village using a generalized linear modeling framework that explicitly accounts for differences in informational exposure.

1.2 Modeling Framework and Procedure

Response, Exposure, and Link Function

The response variable is the number of loan applications, a non-negative count. Villages differ in the number of informational meetings held, which represent opportunities for application. To model conversion rather than raw volume, the logarithm of the number of meetings is included as an offset.
A Poisson GLM with a log link is specified:

$$\log\left(E(Y_i)\right) = \log(\text{meetings}_i) + \beta_0 + \sum_k \beta_k X_{ik}$$

All coefficients are therefore interpreted as multiplicative effects on the application rate per meeting.

Nested Model Strategy

Rather than fitting a single model, we adopt a nested modeling strategy, progressively adding covariates in a conceptually motivated order:

$$\text{Meetings} \rightarrow \text{Literacy} \rightarrow \text{Infrastructure} \rightarrow \text{Income}$$

This ordering reflects a movement from proximal mechanisms to background controls. Each transition between models corresponds to a specific scientific question:

Transition	Question Answered
M0 $\rightarrow$ M1	Does literacy influence the conversion of meetings into applications?
M1 $\rightarrow$ M2	Does infrastructure quality matter beyond literacy?
M2 $\rightarrow$ M3	Are literacy and infrastructure effects robust after controlling for income?

This approach ensures that likelihood ratio tests answer interpretable, mechanism-oriented questions rather than purely statistical ones.

1.3 Evaluation Criteria

Model performance and validity are assessed using complementary metrics:

Likelihood-Based Metrics

- Residual Deviance: Remaining lack of fit.
- Likelihood Ratio (LR) Tests: Differences in deviance between nested models test whether added predictors significantly improve fit.
- AIC: Balances fit and complexity.

Goodness-of-Fit Diagnostics

- Deviance / DF: Global adequacy of the Poisson mean–variance assumption.
- Pearson χ^2 / DF: Diagnostic for over- or under-dispersion.

Coefficient-Based Metrics

- Estimates: Direction and magnitude of effects.
- Standard Errors: Estimation uncertainty.
- Wald z-tests and p-values: Statistical significance of individual predictors.
- Coefficient stability across models: Large changes may indicate confounding or misspecification.

Together, these metrics allow evaluation of both global model adequacy and individual predictor behavior.

1.4 Results

Global Model Comparison

Model	AIC	Residual Deviance	Residual DF	Deviance / DF	ΔDeviance	LR p-value	Pearson χ^2/DF
M0	1445	387	199	1.94	–	–	1.91
M1	1297	237	198	1.20	149.0	2.35×10^{-34}	1.20
M2	1236	174	197	0.88	63.7	1.43×10^{-15}	0.88
M3	1178	114	196	0.58	59.5	1.19×10^{-14}	0.58

Table 1: Global model comparison metrics for nested Poisson GLMs. Each model is color-coded for easier visual distinction.

Coefficient Estimates Across Models

Model	Predictor	Estimate	Std. Error	p-value
M0	Intercept	-0.6265	0.0125	$< 2 \times 10^{-16}$
M1	Intercept	-1.0530	0.0382	$< 2 \times 10^{-16}$
M2	Intercept	-1.1265	0.0394	$< 2 \times 10^{-16}$
M3	Intercept	-1.6072	0.0744	$< 2 \times 10^{-16}$
M1	Literacy rate	0.7393	0.0609	$< 2 \times 10^{-16}$
M2	Literacy rate	0.7273	0.0609	$< 2 \times 10^{-16}$
M3	Literacy rate	0.6732	0.0613	$< 2 \times 10^{-16}$
M2	Infrastructure	0.2029	0.0253	9.6×10^{-16}
M3	Infrastructure	0.2076	0.0253	$< 2 \times 10^{-16}$
M3	Avg. income	0.00101	0.00013	1.37×10^{-14}

Table 2: Coefficient estimates across nested Poisson GLMs

1.5 Interpretation

Overall Assessment of Nested Models

The nested modeling strategy reveals a clear and consistent improvement in model performance as additional explanatory variables are introduced.

From M0 to M1 (Meetings → Literacy) Adding literacy rate leads to a very large reduction in residual deviance ($\Delta D = 149$, $p < 10^{-33}$) and a substantial drop in AIC. This indicates that differences in literacy levels explain a major portion of the variability in the conversion of meetings into loan applications. The Pearson chi-square per degree of freedom drops from ~ 1.9 to ~ 1.2, showing a marked improvement in goodness of fit.

From M1 to M2 (Literacy → Infrastructure) Introducing infrastructure further reduces deviance ($\Delta D = 63.7$, $p < 10^{-14}$) and improves AIC. This shows that, even after accounting for literacy, local infrastructure provides additional explanatory power. The Pearson chi-square ratio falls below 1, suggesting the model now captures the variability in the data well.

From M2 to M3 (Infrastructure → Income) Adding average income yields a statistically significant but smaller improvement ($\Delta D = 59.5$, $p < 10^{-13}$). The continued decrease in AIC indicates that income contributes meaningful information, though its role is more incremental compared to literacy and infrastructure. The Pearson chi-square ratio further decreases, suggesting no evidence of overdispersion.

Overall, each successive model fits the data better than the previous one, and none of the improvements appear to be driven by overfitting. The monotonic decrease in AIC and deviance supports the final model as the most adequate description of the data.

Interpretation of Key Predictors

Literacy Rate

Literacy rate has a strong and stable effect across all models. The estimated rate ratios (≈ 1.96–2.09) indicate that villages with higher literacy levels have nearly twice the expected rate of loan applications per meeting, holding other factors constant. The slight reduction in the coefficient when additional variables are added suggests partial overlap with infrastructure and income, but the effect remains dominant.

Infrastructure

Infrastructure consistently increases the expected application rate by approximately 23% ($RR \approx 1.23$). The coefficient is remarkably stable across models, indicating that infrastructure has an independent effect that is not explained by literacy or income.

Average Income

Average income shows a positive but modest per-unit effect ($RR \approx 1.001$). While small at the unit level, the effect becomes meaningful over realistic income differences. Its statistical significance aligns with the observed improvement in model fit, confirming that income plays a secondary but non-negligible role.

1.6 Conclusion

By combining a conceptually ordered nested modeling strategy with likelihood-based comparisons and coefficient diagnostics, this analysis demonstrates that literacy rate and infrastructure quality are central determinants of micro-credit uptake, while income provides additional but non-substituting explanatory power.

From a modeling perspective, the consistency between global fit improvements and stable, significant coefficients supports the validity of the chosen GLM specification and highlights the importance of aligning model-building procedures with scientific questions rather than purely predictive criteria.

2 Poisson GLM via Manual IRLS Implementation

In this task, we implemented a custom R function to estimate the coefficients of a Poisson Generalized Linear Model using the Iteratively Reweighted Least Squares (IRLS) algorithm. The function takes a design matrix and a response vector as input, iteratively updates the coefficients through weighted least squares, and outputs both coefficient estimates and their standard errors.

During the implementation, we encountered numerical difficulties due to the small size and collinearity of the Galapagos dataset, which caused instability in the matrix inversion step.
These were solved by using QR decomposition for the weighted least squares step and by capping extreme values of the linear predictor to prevent overflow in the exponential function.
After these adjustments, the function ran smoothly, and the coefficient estimates and standard errors obtained were very close to those produced by R's built-in `glm()` function, confirming the correctness of our implementation.

Variable	Manual_Poisson	GLM_Poisson	SE_Manual_Poisson	SE_GLM_Poisson	Manual_Quasi	GLM_Quasi	SE_Manual_Quasi	SE_GLM_Quasi
(Intercept)	2.82841e+00	2.82841e+00	5.95821e-02	5.95821e-02	2.82841e+00	2.82841e+00	0.214818902	0.214818929
Endemics	3.38809e-02	3.38809e-02	1.74116e-03	1.74116e-03	3.38809e-02	3.38809e-02	0.006277613	0.006277614
Area	-1.06730e-04	-1.06730e-04	3.74079e-05	3.74079e-05	-1.06730e-04	-1.06730e-04	0.000134871	0.000134871
Elevation	2.63788e-04	2.63789e-04	1.93425e-04	1.93425e-04	2.63788e-04	2.63789e-04	0.000697382	0.000697382
Nearest	1.04760e-02	1.04760e-02	1.61111e-03	1.61111e-03	1.04760e-02	1.04760e-02	0.005808724	0.005808725
Scruz	-6.83525e-04	-6.83526e-04	5.80215e-04	5.80215e-04	-6.83525e-04	-6.83526e-04	0.002091925	0.002091925
Adjacent	4.53862e-05	4.53862e-05	4.79987e-05	4.79987e-05	4.53862e-05	4.53862e-05	0.000173056	0.000173056

Table 3: Comparison of manual IRLS Poisson estimates with built-in GLM for both Poisson and quasi-Poisson models, including standard errors.

2.1 IRLS and Variance Dependence

The IRLS algorithm naturally accounts for the fact that in a GLM, the variance of the response depends on the mean. At each iteration, IRLS computes a weight matrix where each weight is equal to the expected value of the response under the current parameter estimates (for Poisson, $w_i = \mu_i$). By solving a weighted least squares problem using these weights, the algorithm automatically adjusts the updates of the coefficients to reflect the heteroscedasticity implied by the mean-variance relationship.

2.2 Overdispersion and Quasi-Poisson Adjustment

The Galapagos species dataset is overdispersed, meaning the observed variance exceeds the mean, which violates the strict assumptions of the Poisson model. Our standard IRLS implementation does not automatically adjust the standard errors for overdispersion. To account for this, we can implement a quasi-Poisson model by multiplying the variance-covariance matrix of the coefficients by an estimated dispersion parameter (ϕ), calculated as the Pearson chi-squared statistic divided by the residual degrees of freedom. A quasi-Poisson model maintains the Poisson mean structure but allows the variance to be

$$\mathrm{Var}(Y_i) = \phi\, \mu_i,$$

thus providing more accurate standard errors in the presence of overdispersion.

3 Bootstrap versus Asymptotic Inference

3.1 Methodology

We compared **asymptotic** and **non-parametric bootstrap inference** for a Poisson Generalized Linear Model fitted to the Galápagos species data, with particular focus on the coefficient of *Area*. Since the data are known to exhibit overdispersion, bootstrap resampling provides a useful robustness check against reliance on large-sample approximations.

A non-parametric bootstrap with $B = 1000$ replications was performed. In each iteration, the dataset was resampled with replacement, the Poisson GLM was refitted, and the coefficient associated with Area was stored. A **95% percentile bootstrap confidence interval** was then constructed from the empirical distribution of these estimates. For comparison, an asymptotic confidence interval based on the profile likelihood was obtained using the `confint()` function in R.

3.2 Results

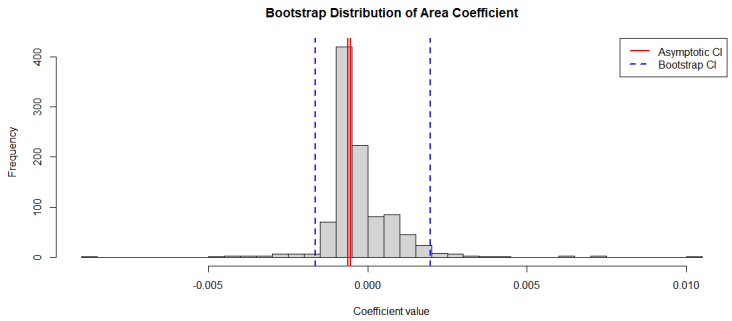

Figure 1: Bootstrap distribution of Area coefficient.

Estimated coeff.:
$\hat{\beta}_{\text{Area}} = -0.00058$

95% CI:

- BootS.: $[-0.00164, 0.00198]$
- Asym.: $[-0.00063, -0.00053]$

Bootstrap summary:

- Mean: -0.00025
- Median: -0.00052
- Mode: -0.00057

3.3 Interpretation

The asymptotic confidence interval is narrow and entirely negative, suggesting a statistically significant negative effect of Area under standard GLM assumptions. In contrast, the bootstrap confidence interval is substantially wider and includes zero, indicating greater uncertainty in the estimated effect.

The bootstrap distribution is centered close to the original GLM estimate, as reflected by the proximity of the median and mode to $\hat{\beta}_{\text{Area}}$. However, the less negative mean reveals right skewness in the sampling distribution, leading to a longer positive tail and a confidence interval that crosses zero. This behavior is consistent with overdispersion and finite-sample effects, which are not fully captured by asymptotic inference.

Overall, the bootstrap analysis provides a more conservative assessment of uncertainty and highlights the value of resampling-based methods when classical GLM assumptions may be violated.