GLM Models in Actuarial Science: Theory, Validation and Business Applications

Andrés González Ortega

8 August 2025

Topics: Risk Model, General Reggression Models

Abstract

This comprehensive study guide presents Generalized Linear Models (GLMs) as a fundamental framework for actuarial modeling, bridging mathematical rigor with practical business applications in insurance and risk management. The work provides a complete treatment of GLM theory, from exponential family distributions and maximum likelihood estimation to advanced validation techniques including confusion matrices, ROC analysis, and actuarial-specific metrics such as the Gini coefficient and lift charts. Key contributions include detailed interpretability frameworks for business decision-making, comprehensive coverage of model validation strategies tailored to actuarial data structures, and practical implementation guidelines addressing regulatory compliance and production considerations. The guide emphasizes the critical role of performance metrics,precision, recall, AUC-ROC, and deviance-based measures,in assessing model quality for claim frequency, severity, and binary classification problems in insurance contexts.

Introduction to GLM Models

Generalized Linear Models (GLMs) extend ordinary linear regression to handle non-normal response distributions and non-linear relationships through link functions. In actuarial science, GLMs are fundamental for modeling insurance claims, pricing, and risk assessment.

Key Components:

- **Random Component**: Response variable Y with probability distribution from exponential family
- **Systematic Component**: Linear predictor $\eta = X\beta$
- **Link Function**: $g(\mu) = \eta$, connecting expected value μ to linear predictor

Why GLMs in Actuarial Science?

- **Flexibility**: Handle various data types (counts, proportions, continuous positive values)
- **Interpretability**: Coefficients have clear business meaning for rate-making
- **Regulatory Compliance**: Transparent methodology satisfies regulatory requirements
- **Risk Quantification**: Natural framework for modeling insurance phenomena with proper uncertainty quantification

Mathematical Foundation

Exponential Family Distributions

GLMs assume the response variable follows an exponential family distribution:

$$f(y|\theta,\phi) = \exp\left[\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right] \tag{1}$$

Where:

- θ: Natural parameter (canonical parameter linking to mean)
- ϕ: Dispersion parameter (controls variance relative to mean)
- $b(\theta)$: Cumulant function (ensures proper normalization)
- $a(\phi)$: Scale function (typically ϕ/w where w is known weight)
- $c(y,\phi)$: Normalization constant (makes integral equal to 1)

Common Actuarial Distributions:

Distribution	Domain	Actuarial Use
Poisson	Non-negative	Claim counts
Gamma	Positive	Claim amounts
Binomial	$\{0,1\}$	Binary outcomes
Tweedie	Mixed	Claim severity with zeros

Link Functions and Linear Predictors

The link function $g(\cdot)$ connects the expected response to the linear predictor:

$$g(E[Y|X]) = \eta = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p \tag{2}$$

Common Link Functions:

- **Identity**: $g(\mu) = \mu$ (Normal distribution - direct linear relationship)
- **Log**: $g(\mu) = \log(\mu)$ (Poisson, Gamma - ensures positive predictions)
- **Logit**: $g(\mu) = \log\left(\frac{\mu}{1-\mu}\right)$ (Binomial - maps probabilities to real line)
- **Probit**: $g(\mu) = \Phi^{-1}(\mu)$ (Binomial alternative using normal CDF inverse)

Maximum Likelihood Estimation

Parameters are estimated by maximizing the log-likelihood:

$$\ell(\beta) = \sum_{i=1}^{n}\left[\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i,\phi)\right] \tag{3}$$

Actuarial Applications

Claim Frequency Modeling

Model: Poisson GLM with log link

$$\log(E[N|X]) = \beta_0 + \beta_1 \cdot \text{Age} + \beta_2 \cdot \text{Gender} + \beta_3 \cdot \text{VehicleType} + \ldots \tag{4}$$

Business Interpretation: Coefficients represent multiplicative effects on claim frequency.

Claim Severity Modeling

Model: Gamma GLM with log link

$$\log(E[S|X]) = \alpha_0 + \alpha_1 \cdot \text{ClaimType} + \alpha_2 \cdot \text{Region} + \alpha_3 \cdot \text{PolicyLimit} + \ldots \tag{5}$$

Key Considerations:

- Right-skewed claim amounts require appropriate distribution choice
- Heavy tails requiring robust estimation techniques
- Inflation adjustments for historical data (trending factors)

Binary Classification Models

Model: Binomial GLM with logit link

$$\text{logit}(P[Y = 1|X]) = \gamma_0 + \gamma_1 \cdot \text{CreditScore} + \dots \quad (6)$$

Applications:

- **Lapse modeling**: Predicting policy termination probability
- **Fraud detection**: Identifying suspicious claims
- **Underwriting decisions**: Accept/reject classification

Tweedie Models for Pure Premium

Combined frequency and severity modeling using compound Poisson-Gamma distribution:

$$E[Y|X] = \mu(X) = \exp(X\beta) \quad (7)$$
$$\text{Var}[Y|X] = \phi\mu(X)^p \quad (8)$$

Where $p \in (1, 2)$ for compound Poisson-Gamma distribution. This approach models total claim cost directly without separate frequency/severity models.

Model Validation Framework

Pearson Chi-Square (alternative goodness-of-fit measure):

$$X^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)} \quad (9)$$

Tests for Model Adequacy:

- **Deviance goodness-of-fit test**: Tests if model adequately fits data
- **Pearson chi-square test**: Alternative fit assessment using standardized residuals
- **Hosmer-Lemeshow test**: Specifically for binomial models, tests calibration across risk deciles

Residual Analysis

Deviance Residuals (preferred for GLMs):

$$r_{D,i} = \text{sign}(y_i - \hat{\mu}_i)\sqrt{d_i} \quad (10)$$

Pearson Residuals (standardized by variance):

$$r_{P,i} = \frac{y_i - \hat{\mu}_i}{\sqrt{V(\hat{\mu}_i)}} \quad (11)$$

Standardized Residuals (adjusted for leverage):

$$r_{S,i} = \frac{r_{P,i}}{\sqrt{1 - h_{ii}}} \quad (12)$$

Diagnostic Plots:

- **Residuals vs. fitted values**: Check for heteroscedasticity and non-linearity
- **Q-Q plots**: Assess distributional assumptions
- **Cook's distance**: Identify influential observations that disproportionately affect model fit
- **Leverage vs. residuals**: Detect high-leverage points with unusual predictor values

Cross-Validation Strategies

K-Fold Cross-Validation:

1. Split data into K folds
2. Train on $K - 1$ folds, validate on remaining fold
3. Repeat K times
4. Average validation metrics

Time Series Cross-Validation:

- Essential for actuarial data with temporal structure and trends
- Use expanding or rolling windows to respect time dependencies
- Always respect chronological order to avoid data leakage

Performance Metrics & Confusion Matrix

Confusion Matrix Framework

For binary classification problems (e.g., claim/no-claim):

		Actual	
		Claim	No Claim
Predicted	Claim	TP	FP
	No Claim	FN	TN

Definitions:

- **TP**: True Positives (correctly predicted claims)
- **TN**: True Negatives (correctly predicted no claims)
- **FP**: False Positives (Type I error - predicted claim but no actual claim)
- **FN**: False Negatives (Type II error - missed actual claims)

Key Performance Metrics

Accuracy (overall correctness):

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{13}$$

Precision (Positive Predictive Value):

$$\text{Precision} = \frac{TP}{TP + FP} \tag{14}$$

Actuarial Interpretation: Of all predicted claims, what proportion are actual claims?

Recall (Sensitivity, True Positive Rate):

$$\text{Recall} = \frac{TP}{TP + FN} \tag{15}$$

Actuarial Interpretation: Of all actual claims, what proportion did we correctly identify?

Specificity (True Negative Rate):

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{16}$$

F1-Score (harmonic mean of precision and recall):

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{17}$$

Area Under ROC Curve (AUC-ROC):

- Plots True Positive Rate vs. False Positive Rate across all thresholds
- AUC = 0.5: Random classifier performance
- AUC = 1.0: Perfect classifier
- AUC ¿ 0.7: Generally acceptable for actuarial applications

Actuarial-Specific Metrics

Gini Coefficient (related to AUC, measures discriminatory power):

$$\text{Gini} = 2 \times \text{AUC} - 1 \tag{18}$$

Range: [-1, 1], with higher values indicating better discrimination between risk classes.

Lift Charts: Measure how much better the model performs compared to random selection in top percentiles (e.g., top 10% of highest-risk policies).

Concordance Index (C-Index): Proportion of pairs where higher-risk observation has higher predicted probability. Equivalent to AUC for binary outcomes.

Kolmogorov-Smirnov Statistic: Maximum difference between cumulative distributions of scores for positive and negative classes, measuring separation quality.

Regression Performance Metrics

Mean Absolute Error (MAE):

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i| \tag{19}$$

Root Mean Square Error (RMSE) (penalizes large errors more):

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \tag{20}$$

Mean Absolute Percentage Error (MAPE) (scale-independent):

$$\text{MAPE} = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{y_i} \right| \times 100\% \tag{21}$$

Deviance Explained (GLM equivalent of R-squared):

$$D^2 = 1 - \frac{\text{Deviance}_{\text{model}}}{\text{Deviance}_{\text{null}}} \tag{22}$$

Business Interpretability

Log Link (Multiplicative Effects):

- β coefficient: log(rate ratio)
- $\exp(\beta)$: multiplicative factor on the response
- $(\exp(\beta) - 1) \times 100\%$: percentage change

Example: If $\beta_{\text{age}} = 0.05$ in a Poisson claim frequency model:

- Each additional year increases claim frequency by $\exp(0.05) = 1.051$ times
- This represents a 5.1% increase per year

Logit Link (Odds Ratios):

- β coefficient: log(odds ratio)
- $\exp(\beta)$: odds ratio (change in odds for unit increase in predictor)
- For probability: $p = \frac{\exp(\eta)}{1+\exp(\eta)}$

Variable Importance

Statistical Significance:

- **Wald tests**: $z = \frac{\hat{\beta}}{\text{SE}(\hat{\beta})}$ (tests individual coefficient significance)
- **Likelihood ratio tests**: $2[\ell(\text{full}) - \ell(\text{reduced})]$ (compares nested models)

Economic Significance:

- **Elasticity measures**: Percentage change in response per percentage change in predictor
- **Dollar impact calculations**: Translation of coefficient effects to monetary terms
- **Business relevance assessment**: Whether statistical significance translates to practical importance

Model Transparency

SHAP Values (Shapley Additive Explanations):

- Attribute prediction to individual features fairly using game theory
- Provide both local (individual prediction) and global (overall model) interpretability
- Satisfy efficiency, symmetry, dummy feature, and additivity properties from cooperative game theory

Feature Attribution:

$$f(x) = E[f(X)] + \sum_{i=1}^{p} \phi_i(x) \tag{23}$$

Where $\phi_i(x)$ is the SHAP value for feature i, representing its marginal contribution.

Risk Factor Analysis

Relativity Tables show multiplicative factors for each level of categorical variables:

Age Group	Frequency Relativity	Severity Relativity
18-25	1.45	1.20
26-35	1.00 (base)	1.00 (base)
36-50	0.85	1.10
51-65	0.70	1.25
65+	0.60	1.35

Interaction Effects model multiplicative interactions between risk factors:

$$\log(\mu) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 \tag{24}$$

Implementation Guidelines

Data Preparation

Exploratory Data Analysis:

- **Distribution analysis**: Examine response variable distribution to select appropriate GLM family
- **Missing value assessment**: Identify patterns and decide on imputation strategies
- **Outlier detection**: Use statistical methods (IQR, z-scores) and business knowledge
- **Correlation analysis**: Check for multicollinearity among predictors

Feature Engineering:

- **Binning continuous variables**: Create interpretable risk bands for pricing
- **Creating interaction terms**: Model synergistic effects between variables
- **Temporal features**: Capture seasonality, trends, and calendar effects
- **External data integration**: Incorporate economic indicators, weather data, etc.

Variable Selection:

- **Forward/backward selection**: Stepwise procedures based on statistical criteria
- **Regularization**: Ridge, Lasso, Elastic Net to handle high-dimensional data
- **Information criteria**: AIC, BIC for model comparison and selection
- **Business knowledge integration**: Ensure model includes actuarially relevant variables

Model Development Process

Baseline Model:

- Start with simple main effects only
- Establish benchmark performance metrics
- Validate basic distributional assumptions

Model Enhancement:

- Add interaction terms between significant predictors
- Include polynomial terms for non-linear relationships
- Test alternative link functions for better fit
- Consider offset variables for exposure adjustment

Regularization (penalized likelihood):

$$\text{Penalized Log-Likelihood} = \ell(\beta) - \lambda_1 \sum_{j=1}^{p} |\beta_j| - \lambda_2 \sum_{j=1}^{p} \beta_j^2 \tag{25}$$

Model Selection criteria:

- Information criteria comparison (prefer lower AIC/BIC)
- Cross-validation performance on holdout data

- Business interpretability and regulatory acceptance
- Computational efficiency for production use

Production Considerations

- **Performance degradation detection**: Track key metrics over time
- **Data drift monitoring**: Detect changes in predictor distributions
- **Champion-challenger testing**: A/B test new model versions
- **Regular retraining schedules**: Update models with fresh data

Documentation Requirements:

- Model development documentation for audit trails
- Technical specifications for IT implementation
- Business use case description for stakeholders
- Comprehensive validation results for regulators

Case Studies

Auto Insurance Claim Frequency

Objective: Develop GLM for annual claim frequency prediction.

Data Structure:

- Response: Number of claims per policy year
- Exposures: Policy exposure time (to handle partial years)
- Predictors: Age, gender, vehicle type, region, driving record, etc.

Model Specification:

$$\log\left(\frac{E[\text{Claims}]}{\text{Exposure}}\right) = \beta_0 + \beta_1 \cdot \text{Age} + \beta_2 \cdot \text{Gender} + \ldots \quad (26)$$

Key Results:

- Young drivers (18-25): 45% higher claim frequency than base
- Sports cars: 30% higher frequency than sedans
- Urban areas: 20% higher frequency than rural areas

Business Impact:

- Improved risk-based pricing leading to better risk selection
- 5% reduction in loss ratio through better rate adequacy

Life Insurance Lapse Prediction

Objective: Predict policy lapse probability using logistic regression.

Model Specification:

$$\text{logit}(P[\text{Lapse}]) = \gamma_0 + \gamma_1 \cdot \text{Duration} + \gamma_2 \cdot \text{PremiumRatio} + \gamma_3 \cdot \text{CashValue} + \quad (27)$$

Performance Metrics:

- AUC-ROC: 0.78 (good discriminatory power)
- Precision: 0.62 (62% of predicted lapses are actual lapses)
- Recall: 0.71 (captures 71% of actual lapses)
- F1-Score: 0.66 (balanced precision-recall performance)

Business Applications:

- **Targeted retention campaigns**: Focus on high-lapse-probability policies
- **Dynamic pricing strategies**: Adjust premiums based on lapse risk
- **Portfolio risk management**: Assess persistency risk in business planning
- **Regulatory capital allocation**: Better estimate of surrender risk

Workers' Compensation Severity

Objective: Model claim severity using Gamma GLM.

Model Features:

- Injury type classification (medical vs. indemnity components)
- Industry sector effects (construction vs. office work)
- Geographic adjustments for cost variations
- Inflation trends and medical cost escalation

Validation Results:

- Deviance explained: 34% (reasonable for severity modeling)
- MAPE: 28% (acceptable prediction accuracy)
- Residual analysis shows good fit with no systematic patterns
- No significant autocorrelation in diagnostic plots

Implementation:

- Integrated into automated pricing platform
- Enables real-time quote generation

- Supports regulatory filing requirements
- Provides competitive positioning insights

Conclusion

GLMs provide a robust framework for actuarial modeling, offering the flexibility to handle diverse data types while maintaining interpretability crucial for business applications. The key to successful implementation lies in:

1. **Rigorous validation** using appropriate statistical tests and business-relevant metrics
2. **Clear interpretation** of model coefficients in actionable business context
3. **Comprehensive documentation** for regulatory compliance and audit purposes
4. **Ongoing monitoring** to ensure model performance stability in changing environments

The combination of statistical rigor, business interpretability, and regulatory transparency makes GLMs an essential tool in modern actuarial practice. Success requires balancing mathematical sophistication with practical business needs while maintaining model transparency for managment confidence.

Refferences and details

- McCullagh, P. & Nelder, J.A. (1989). *Generalized Linear Models*
- De Jong, P. & Heller, G.Z. (2008). *Generalized Linear Models for Insurance Data*
- Ohlsson, E. & Johansson, B. (2010). *Non-Life Insurance Pricing with Generalized Linear Models*
- Casualty Actuarial Society E-Forum on GLM Applications