Introduction to Predictive Modeling Using GLMs

A Practitioner’s Viewpoint

Dan Tevet, FCAS, MAAA
Anand Khare, FCAS, MAAA, CPCU
Outline

• Overview of predictive modeling
• Predictive modeling in the actuarial world
• Simple linear models vs generalized linear models (GLMs)
• Specification of GLMs
• Interpretation of GLM output
• Frequency/severity vs pure premium modeling
• Model validation
What is Predictive Modeling

• Model – an abstraction of reality, generally with a random or probabilistic component
  – Simplification of a real world phenomenon
• Model types include:
  – Linear models – predict target variable using linear combination of predictor variables
  – Trees – split dataset, one variable at a time, into subgroups that behave similarly
  – Neural networks – “self-learning” algorithms that adapt to best predict a quantity of interest
How Do Actuaries Use Modeling?

• Rating plans – model insurance loss data to build plans that charge actuarially fair rates

• Underwriting plans – knowing relative riskiness of policyholder can inform underwriting decisions

• Enterprise risk management – model correlations between lines of business or probability of ruin

• Customer retention – model probability of customer renewing each year
Predictive Modeling Process

- Collect Data
- Exploratory Data Analysis
  - Examine univariate distributions
  - Examine relationship of each variable to target
- Specify Model
- Evaluate Output
- Validate Model
- “Productize” Model
- Maintain Model
- Rebuild Model
Simple Linear Model

• $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \epsilon$
  – $Y$ is the *target* or *response* variable – it is what we are trying to predict (e.g. pure premium)
  – $X_1, X_2$, etc are the *explanatory* (e.g. age of driver, type of vehicle) variables – we use them to predict $Y$
  – $\epsilon$ is the *error* or *noise* term – it is the portion of $Y$ that is unexplained by $X$

• $\mu = E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n$
• In general, we are modeling the mean of $Y$
Simple Linear Model Assumptions

• Assumptions of simple linear models
  – Target variable Y does not depend on the value of Y for any other record, only the predictors
  – Y is normally distributed
  – Mean of Y depends on the predictors, but all records have same variance
  – Y is related to predictors through simple linear function

• Unfortunately, these assumptions are often unrealistic
  – Target variables of interest, such as pure premium, frequency, and severity, are not normally distributed and have non-constant variance
Generalized Linear Models

• Generalized linear model: \( g(\mu) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n \)

• Assumptions of generalized linear models
  – Target variable \( Y \) does not depend on the value of \( Y \) for any other record, only the predictors
  – Distribution of \( Y \) is a member of the exponential family of distributions
  – Variance of \( Y \) is a function of the mean of \( Y \)
  – \( g(\mu) \) is linearly related to the predictors. The function \( g \) is called the link function

• The exponential family of distributions include the following: Normal, Poisson, Gamma, Binomial, Negative Binomial, Inverse Gaussian, Tweedie
• \( \text{Var}(Y) = \phi \cdot V(\mu) / w \)

• \( \phi \) is the dispersion coefficient, which is estimated by the GLM

• \( w \) is the weight assigned to each record
  – GLMs calculate the coefficients that maximize likelihood, and \( w \) is the weight that each record gets in that calculation

• \( V(\mu) \) is the GLM Variance Function, and is determined by the distribution
  – Normal: \( V(\mu) = 1 \)
  – Poisson: \( V(\mu) = \mu \)
  – Gamma: \( V(\mu) = \mu^2 \)
GLM Link Function

\[ g(\mu) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots \]

- **Common choices for link function**
  - Identity: \( g(\mu) = \mu \)
  - Log: \( g(\mu) = \ln(\mu) \)
  - Logit: \( g(\mu) = \ln\left[\frac{\mu}{1-\mu}\right] \)

- **Log link commonly used to model rating plans because it produces multiplicative relativities**
  - \( \ln(\mu) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \)
    - \( \mu = \exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2) \)
    - \( \mu = \exp(\beta_0)\exp(\beta_1 X_1)\exp(\beta_2 X_2) \)

- **Logit link used to model probability of event occurring**
Offsets

• An effect in a model that is fixed by the modeler
• Variable offsets – fix the effect of variables that are not being modeled
  – Example: Constructing a rating plan and not modeling base territory rates
  – Solution: Offset for current territory rates
• Volume offsets – reflect fact that different records have different volumes of data and thus have different expected values
  – Example: Modeling claim counts. Some records have a single exposure, other have many exposures
  – Solution: Offset for exposure volume of each observation
• Discrete variables: exponentiate GLM coefficient
  – Example: coefficient for youth drivers is 0.52
    → Rating factor = \( \exp(0.52) = 1.68 \)
    → Youth drivers have 68% surcharge relative to base level of adult drivers (who, by definition, have rating factor of 1.00)

• Continuous variables with no transformation
  – Example: modeling pure premium, and annual miles driven is a continuous variable
  – As miles driven increases by 1 unit, expected pure premium is scaled by a factor of \( \exp(\beta) \), regardless of whether mileage goes from 1,000 to 2,000 or 20,000 to 21,000

• Continuous variables with log transformation
  – Pure premium ~ \( (\text{annual mileage})^{\beta} \)
  – If \( \beta < 1 \), then as mileage increases, pure premium increases at decreasing rate
GLMs allow us to quantify uncertainty in parameter estimates.

Wald 95% confidence interval for mean of parameter estimate = Mean +/- 1.96*(Standard Error)

Test for the significance of an individual parameter
- Wald Chi Square = (Parameter Estimate/Standard Error)^2
  Approximately follows a Chi Squared distribution with 1 degree of freedom
- P-value is probability of obtaining a Chi Square statistic of given magnitude by pure chance
  Lower p-value → more significant
Two Modeling Approaches

• Pure Premium Approach: Build a single model for pure premium
  – Generally straightforward to implement

• Frequency-Severity Approach: Build one model for claim frequency and another for claims severity
  – Additional work for additional insight
Pure Premium Approach

• Advantages:
  – Only a single model needs to be built
  – No need to split variable offsets
  – Results often very similar to frequency-severity approach

• Disadvantages:
  – Yields less insight than frequency-severity
  – Tweedie distribution is only good choice
    • Relatively new and mathematically complex
    • Includes implicit assumptions that may not hold
Frequency-Severity Approach

• Advantages:
  – May yield meaningful insights about data
  – Can choose from several well-known and well-understood distributions

• Disadvantages:
  – Two models to build, run, and validate
  – Requires splitting variable offsets
  – Often produces limited additional benefit
Tweedie Distribution

• Mixed Poisson-Gamma process – number of claims follow a Poisson distribution, and the size of each claim follows a Gamma distribution

• The Tweedie is a 3-parameter distribution:
  – Mean ($\mu$), equal to the product of the means of the underlying Poisson and Gamma distributions
  – Power ($p$), which depends on the coefficient of variation of the underlying Gamma distribution
  – Dispersion ($\phi$), a measure of variance
Frequency Distribution Options

• Poisson
  – The Coca Cola of claim count distributions

• Overdispersed frequency distributions
  – Overdispersed Poisson
  – Zero-Inflated Poisson
  – Negative Binomial
  – Zero-Inflated Negative Binomial
Severity Distribution Options

• Several reasonable distributions
• Criteria
  – Member of exponential family
  – $p \geq 2$, where $V(\mu) = \mu^p$
• In order of increasing variance:
  – Gamma ($p=2$)
  – Tweedie ($2 < p < 3$)
  – Inverse Gaussian ($p=3$)
  – Tweedie ($p>3$)
Three Pillars of Model Validation

• Tests of Fit
• Tests of Lift
• Tests of Stability
Fit Statistics

• Traditional: Absolute/Squared Error
• Alternatives: Likelihood, Deviance, Pearson’s Chi-Squared
• Penalized: AIC, BIC
• Per Observation: Residuals, Leverage
Absolute/Squared Error

• Only appropriate if data is normally distributed

• Inappropriate to use on disaggregate claim frequency, severity, or pure premium data

• Useful to assess model fit within buckets
  – Bucket data into percentiles, or similar quantiles, and calculate squared difference between actual and predicted for each bucket
Better Alternatives to Squared Error

• Likelihood: chance of observation, given model
  – Always increases as parameters are added to model
• Deviance: twice the difference in loglikelihoods between the saturated and fitted models
  – GLMs are fit so as to minimize deviance
  – Accounts for the shape of the distribution
• Pearson’s chi-squared: squared error divided by the variance function of the distribution
  – Accounts for the skew of the distribution
Penalized Measures

• Akaike Information Criterion (AIC): Penalizes loglikelihood for additional model parameters

• Bayesian Information Criterion (BIC): Penalizes loglikelihood for additional model parameters, and this penalty increases as the number of records in the dataset increases
  – Can be too restrictive

• Used primarily for variable selection
Per Observation

• Traditional residual: actual minus predicted

• Deviance residual: square root of weighted deviance times sign of actual minus predicted
  – Reflects the shape of the distribution
  – Plotting deviance residual against weight or any predictor should yield an uninformative cloud
  – Should be approximately normally distributed

• Leverage: used to identify extreme outliers
  – Does not necessarily measure impact
Model Lift

• Lift is meant to approximate economic value
  – Fit has no relationship with economic value

• Economic value is produced by comparative advantage in avoidance of adverse selection
  – Lift is a *comparative* measure, i.e. the lift of one model over another, or the lift of a model over status quo

• Lift should always be measured on holdout data
Lift Measures

- Simple Quantile Plot
- Double Lift Chart
- Loss Ratio Chart
- Gini Index
Double Lift Chart

![Double Lift Chart](image-url)
Loss Ratio Chart

![Loss Ratio Chart Image]
Economic Gini Index
Gini Index of Rating Plan

• Model should differentiate lowest and highest loss cost policyholders

• Creation of Gini index:
  – Order policyholders by model prediction, from best to worst
  – X-axis is cumulative percent of exposures
  – Y-axis is cumulative percent of losses

• Had model produced Gini index in prior slide, would have identified 60% of exposures that contribute only 20% of losses
Methods for Testing Model Stability

• Cross-validation
  – Split data into subsets (e.g. by time period)
  – Refit model on each subset
  – Compare model parameter estimates

• Bootstrapping
  – Refit model on many bootstrapped samples
  – Calculate variability of parameter estimates

• Deletion of influential records
Measures of Influence

• Cook’s Distance: Statistical measure of the impact each record has on the overall model
  – Excellent tool for identifying errors or anomalies
  – Deletion of records with high Cook’s Distance may significantly change model results, and so this procedure can be used to test stability

• DFBETA: Influence on a certain parameter

• Influence is not to be confused with leverage
For Further Reference

For Further Reference

For Further Reference

For Further Reference

Introduction to Predictive Modeling Using GLMs

A Practitioner’s Viewpoint

Dan Tevet, FCAS, MAAA
Anand Khare, FCAS, MAAA, CPCU