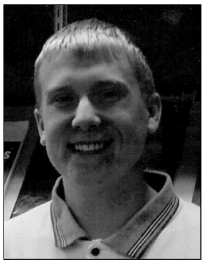
## Extended Thesis Abstracts

# Generalized Linear Models and Their Applications to Actuarial Modeling

*James Smith*



**James Smith** graduated Summa Cum Laude from Ball State in May 2004 with a major in Actuarial Science and a minor in Foundations of Business. Jimmy served as the president of the Actuarial Science Club for two years. He is an actuarial analyst for The Hartford in Southington, Connecticut. Curtis Gary Dean was his honors thesis advisor.

There are many phenomena in the world that, upon close observation, are clearly related. For example, when profits decline, unemployment rises. When weather conditions worsen, drivers have more accidents. The goal of an actuary is to model such relationships. Perhaps the most well-known and accepted method for modeling relationships is linear regression. The simplest version of this method is known as the standard linear model. It relates two variables $X$ and $Y$ by $Y = \alpha + \beta X$. The variable $X$ is taken as the independent variable, and $Y$ is a variable assumed to depend on $X$. Here, $\alpha$ and $\beta$ are constants whose values depend on the relationship between $X$ and $Y$. While such a model may be useful for predicting the outcome of $Y$, given a value for $X$, the prediction will not be exact. Instead, each observation will satisfy the equation $y_i = \alpha + \beta x_i + \epsilon_i$, where $\alpha + \beta x_i = \hat{y}_i$ represents the fitted (predicted) value, and $\epsilon_i$ is the error term that measures the difference between the fitted value and the actual outcome. In other words $\epsilon_i = y_i - \hat{y}_i$ for $i = 1, 2, \ldots, N$, where $N$ is the number of observations.

In the classical linear regression model, four important assumptions are made about the error term:

1. $\mathrm{E}[\epsilon_i] = 0$ for all $i$.
2. $\epsilon_i$ has constant variance $\sigma^2$ for all $i$ (i.e. $\mathrm{E}[\epsilon_i^2] = \sigma^2$ for all $i$).
3. All error terms are independent for different observations.
4. Each error term is normally distributed.

While these assumptions are useful for building a solid statistical foundation for testing the validity of the linear model, they are also quite restrictive. For example, the assumption that the error terms are normally distributed implies that $Y$ is also normally distributed. Unfortunately, many of the random variables in the real world that we would like to model are clearly not normally distributed. Claim counts, for instance, typically follow a Poisson distribution, or even a negative binomial distribution. For random variables such as these, a more accurate modeling tool is the generalized linear model.

## Generalized Linear Models

The generalized linear model is an extension of multiple linear regression, which is an extension of the standard linear model. In multiple linear regression, $Y$ is assumed to be dependent on several factors such as

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k.$$

This is commonly expressed in matrix form as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}.$$

Here, $\mathbf{Y}$ is an $N \times 1$ matrix of $N$ observations, $\boldsymbol{\beta}$ is a $(k+1) \times 1$ matrix of the beta coefficients, $\mathbf{X}$ is an $N \times (k+1)$ matrix containing $N$ observations for $k$ independent variables and a column of 1's corresponding to $\beta_0$, and $\boldsymbol{\epsilon}$ is an $N \times 1$ matrix of the error terms. We refer to $\mathbf{X}\boldsymbol{\beta}$ as the linear predictor.

The key to the generalized linear model lies in how $\mathbf{X}\boldsymbol{\beta}$ is related to the mean of $\mathbf{Y}$ (i.e. $\boldsymbol{\mu}$). In multiple linear regression, $\mathbf{X}\boldsymbol{\beta}$ is taken to be the mean itself. In other words,

$$\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\mu}.$$

In the case of the generalized linear model, however, each component of $\mathbf{X}\boldsymbol{\beta}$ is taken to be a (common) function of the corresponding mean. That is,

$$\mathbf{X}\boldsymbol{\beta} = g(\boldsymbol{\mu}).$$

The function $g$ is called the link function and is determined by the probability distribution of $\mathbf{Y}$. This link function models the nonlinear relationship between $\mathbf{Y}$ and $\mathbf{X}$.

If a random variable $Y$ is a member of the exponential family and is dependent on a parameter $\mu$, then the probability density function of $Y$ can be written in the following form

$$f(y;\mu) = \exp[a(y)b(\mu) + c(\mu) + d(y)].$$

Common properties of distributions in the exponential family allow a coherent theory for generalized linear models for those distributions. The exponential family includes many widely used distributions including the binomial, Poisson, normal, and gamma distributions. For a matrix $\mathbf{Y}$ of random variables, each component is assumed to be a member of the exponential family.

If $a(y) = y$, then the function is said to be in canonical form for the variable $Y$. Likewise, if $b(\mu) = \mu$, then it is in canonical form for the parameter $\mu$. If it is not in canonical form, a transformation can be made to force it into canonical form. For example, let $\theta$ represent a component of the linear predictor $\mathbf{X}\boldsymbol{\beta}$. If the substitutions $\theta = b(\mu)$ and $y = a(y)$ are made, the above equation becomes

$$f(y;\theta) = \exp[y\theta + c(\theta) + d(\theta)].$$

Note that the functions $c(\,.\,)$ and $d(\,.\,)$ will now be different from those in the previous equation depending on what transformations are made. Putting the distribution into canonical form makes the computations easier and allows the use of general formulas that will work for many distributions For example, using the canonical form, it can easily be shown that the general formula for the inverse of the link function is $\mu = -c'(\theta)$, where $\mu$ is any component of the matrix $\boldsymbol{\mu}$. This can be inverted to produce the original link function $g(\mu)$, but the inverse of the link function is where the value truly lies. It shows what transformation must be made to the linear predictor $\mathbf{X}\boldsymbol{\beta}$ in order to produce $\boldsymbol{\mu}$, the mean of $\mathbf{Y}$. For example, if $Y$ is Poisson, then the link function turns out to be the natural logarithm. Applying the inverse of this function to the linear predictor $\mathbf{X}\boldsymbol{\beta}$ will produce the desired mean of $\mathbf{Y}$. That is, $\boldsymbol{\mu} = \exp[\mathbf{X}\boldsymbol{\beta}]$. Thus, it is a rather simple task to find the proper link function for any distribution one would want to use. Being able to specify the link function gives one the freedom to model $Y$ as having any probability distribution that seems fit to assume. This relaxes many of the assumptions made by the standard linear model. The error terms no longer have to be assumed to be normally distributed, and their variance does not have to be assumed as constant. Because of this, generalized linear models are a more appropriate tool for modeling real world data.

This result is particularly important to actuaries. Actuaries constantly work with real world data and model real world phenomena that are known not to be normally distributed. An example is aggregate claims for automobile insurance. A common method for modeling aggregate claims is to separate it into two factors: frequency and severity. Frequency, the number of claims that occur in a period, is often most appropriately modeled using a Poisson distribution or even a negative binomial distribution. Likewise, severity, the average claim size, is commonly modeled using a gamma distribution or something very similar. These distributions all have important properties that differ from the normal distribution. Most notably, in each case the variance is not independent of the mean. Generalized linear models offer the flexibility needed to incorporate important properties such as this into a model.

The benefits of generalized linear models are already being realized in many areas of actuarial practice. Such areas include marine liability insurance pricing, loss reserving, estimating claim settlement values, territorial rating, and others. Yet, this is only the tip of the iceberg. As actuaries continue to become familiar with generalized linear models, more areas of actuarial practice will incorporate generalized linear models into their work. The generalized linear model is a powerful tool whose full benefit is yet to be realized.