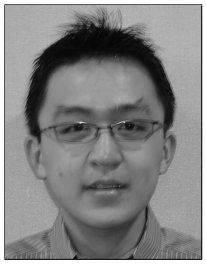
Student-Faculty Seminar

Introduction to Generalized Linear Models

Eugene Tan



Eugene Tan obtained his Bachelor's degree in Mathematics from Northern Arizona University in 2006, and graduated from Ball State with a Masters in Actuarial Science. He was a recipient of the Harold J. Gale Scholarship in 2008. Currently, he is a full-time Actuarial Analyst with Watson Wyatt Worldwide in Chicago

Abstract

Generalized Linear Models (GLM) provides a unifying framework for many commonly used statistical modeling techniques, particularly in the actuarial field. The purpose of this paper is to portray the basic ideologies behind the usefulness of generalized linear models.

Introduction

Advances in statistical theory and computational methodologies have enabled statisticians to use methods analogous to those that have been developed for linear models in a broader approach. Generalizations can be made about the distributions of response variables, which can be, but are not limited to the Normal distribution. In fact, the many properties of the Normal distribution which make it easy to understand and work with are shared by a wider class of distributions, known as the *exponential family of distributions*.

Consider a single random variable Y, with its probability distribution depending on a single parameter, θ. It belongs to the exponential family of distributions if it can be written in the form

$$f(y;\theta) = e^{a(y)b(\theta)+c(\theta)+d(y)}.$$

The more popular models in the exponential family besides the Normal distribution are Poisson and Binomial. In the case of the Poisson distribution, it can be used to model count data where the number of occurrences in a defined environment are probabilistically independent.

The probability distribution function of the Poisson is

$$f(y;\theta) = \frac{\theta^y e^{-\theta}}{y!}.$$

Exponentiating the distribution function yields

$$f(y;\theta) = e^{y\ln\theta - \theta - \ln y!}.$$

In this form, $a(y) = y,\ b(\theta) = \ln\theta,\ c(\theta) = \theta$, and $d(y) = \ln y!$.

In addition, we need and can find expressions for the expected value and variance of the derivatives of the log-likelihood function. From above, the log-likelihood function is

$$l(\theta;y) = a(y)b(\theta) + c(\theta) + d(y)$$

and its derivative with respect to θ is

$$U(\theta;y) = \frac{dl(\theta;y)}{d\theta} = a(y)b'(\theta) + c'(\theta),$$

where the function U is the score statistic and its expected value and variance are key components for inference about parameter values.

The idea of a generalized linear model is anchored upon the notion that for a set of independent random variables from the same exponential family, the point of analytical interest is the smaller set of parameters, where a monotone, differentiable *link function* exists as a transformation. This link function in question provides an explanatory relationship between the linear predictor and the mean of the original distribution function.

The identification of the link function is highly dependent on the study and the nature of the response variable. For example, if we have a binary response then one can show through the exponential family representation of Binomial distribution that the natural or canonical link for the linear combination of explanatory variables is the logit link. Other possible links for binary response which are not the canonical links are probit and complementary log-log links. However, it may prove beneficial should the domain of the link function be matched to the range of the mean of the distribution function. For example, a study modeling count data would logically be better served using Poisson regression where the natural link is a log link, again coming from the exponential family representation of the Poisson distribution

Recall the classical model

$$Y = x\beta + \epsilon$$

The link function is such that

$$g(E(Y_i)) = g(\mu_i) = x_i^T\beta + \epsilon.$$

In essence, the generalized linear model is comprised of three parts:

- The response variable(s) $Y_1, \ldots, Y_N$, which are assumed to be of the same distribution from the exponential family
- A set of parameters β and explanatory variables
- A monotone link function

Case Study

The Society of Actuaries puts forth a set of preliminary exams that cover a variety of material, including mathematical statistics and probability, interest theory, life contingencies, financial economics, estimation and credibility. These examinations average 3 hours and require plenty of preparation. SOA-recommended amount of study time is 100 hours per hour of exam. This study will attempt to answer the question of whether the amount of time spent (in hours) studying is significant to the probability of passing any given 3 hour exam.

Because of the nature of this study, initial thoughts were to model the study using the logistic function, since the response variable is of the simple binary form, Pass/Fail. If p is the probability of passing, then $\frac{p}{1-p}$ is the corresponding odds.

As such, the link function is

$$g(p) = \ln\left(\frac{p}{1-p}\right) = x^T\beta,$$

which is then equivalent to modeling the probability as

$$p = \frac{e^{x^T\beta}}{1+e^{x^T\beta}},$$

and in this particular case,

$$x^T\beta = \beta_0 + \beta_1 x,$$

with x being the number of hours spent studying. The likelihood function of a binomial model is

$$\binom{n}{y} p^y(1-p)^{n-y} \to \binom{n}{y}\left(\frac{e^{x^T\beta}}{1+e^{x^T\beta}}\right)^y\left(1-\frac{e^{x^T\beta}}{1+e^{x^T\beta}}\right)^{n-y}.$$

There is no closed form for the maximum likelihood estimates of the parameters β_0 and β_1, as each parameter is dependent on the other. In generalized linear models, the *Iterative re-Weighted Least Squares* (IRLS) method is used. This iterative numerical technique corresponds to the maximum likelihood criterion if experimental errors have a normal distribution (which is part of the driving factor behind generalized linear models). Under SAS, the default algorithm is known as the Fisher scoring method, which utilizes the expected information matrix in the acquisition of the maximum likelihood estimator (MLE).

Data for 100 participants, including the number of hours studied, pass/fail information, and exam score were simulated. Resulting data that was produced were close in line with passing rates released by the SOA after each exam sitting. The median number of hours studied among the participants was 174. Testing was to determine if studying more than 174 hours was significant in affecting the probability of passing.

The logistic procedure in SAS was used in analysis of data and the response variable Y was marked as 1 if a participant passed. As such, thirty-nine percent were successful. In the absence of further information about the population, the model fit statistics provided a good measure to a decent extent. Further criteria to consider may be how far in advance a participant began preparation, average weekly hours spent studying, or even the number of hours of employer-provided study time. At the default significance level, the null hypothesis, that there is no significant increase in the probability of passing with an increased number of hours studied, was rejected. Furthermore, the resulting parameter estimates were also not rejected. β_0, the intercept, was -1.9924 and β_1, the effects of the categorical hours spent studying that was greater than or equal to 174, was 2.6557. This led to

$$x^T\beta = -1.9924 + 2.6557x.$$

Finally, the point estimate for the odds ratio suggests a highly increased probability of passing beyond 174 hours of preparation at a value of 14.235, with the 95% Wald confidence interval being $(5.06, 40.049)$.

Conclusion

Additionally, suggestive approaches to taking this case study further have been proposed. Applications of other models may be considered, like the Poisson model for discrete analytical purposes. An example along the same lines of the case study would be counting the number of people passing any given test, with the parameters of interest being factors that affect their total hours of studying.

The study was a simple example of the utility that generalized linear models provide. As closely as the data was simulated, it is not reflective of actual data, and it could serve as a better test had data been collected instead. Furthermore, the test and data would have included more parameters of interest as stated prior. However, time constraints have deemed the collection of data not feasible.

References

[1] A. J. Dobson, An Introduction to Generalized Linear Models, Chapman & Hall/ CRC (2002).

[2] SAS Online Documentation.
⟨`http://support.sas.com/documentation/`⟩