DRAFT

Appendix I – Generalized Linear Model

The Poisson Regression Model

The Poisson regression model is a specific type of generalized linear model (GLM). A comprehensive reference for GLMs is McCullagh, P. and Nelder, J.A. (1989) *Generalized Linear Models*. Second Edition. London: Chapman and Hall[1].

A GLM is described by the following assumptions:

1. There is a response variable, y, observed independently for specific values of the predictor variables, $x_1, x_2, \ldots, x_p$.
2. The predictor variables influence the distribution of y through a single linear function called the *linear predictor* $\eta = \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$.
3. The distribution of y has a density function of the form $f(y_i;\theta_i,\phi) = \exp\left[A_i\{y_i\theta_i - g(\theta_i)\}/\phi + \tau(y_i,\phi/A_i)\right]$, where ϕ is a scale parameter, A_i is a known prior weight, and parameter θ_i depends upon the linear predictor.
4. The mean, μ, is a smooth invertible function of the linear predictor: $\mu = m(\eta)$, $\eta = m^{-1}(\mu) = l(\mu)$. The inverse function, $l(\bullet)$, is called the **link function**.

For a Poisson distribution with mean μ, we have $\ln f(y) = y\ln(\mu) - \mu - \ln(y!)$ so $\theta = \ln(\mu)$, $\phi = 1$ and $g(\theta) = \mu = e^{\theta}$.[2]

Given n observations from a GLM, the log-likelihood function is

$$l(\theta,\phi;Y) = \sum_{i=1}^{n}\left[A_i\{y_i\theta_i - g(\theta_i)\}/\phi + \tau(y_i,\phi/A_i)\right],$$

which has a score function for θ of

$$U(\theta) = A_i\{y_i - g'(\theta_i)\}/\phi.$$

From this it can be shown that

$$E(y_i) = \mu_i = g'(\theta_i) \text{ and } VAR(y_i) = \frac{\phi}{A_i}g''(\theta_i).$$

The score function is only provided here for reference purposes; we do not make use of it in subsequent sections of this report. For a derivation of this, including use of the score function, see McCullagh and Nelder, 1989, section 2.2.

---

[1] Another source for a quick review of generalized linear models is Christensen, R. 1997. *Log-Linear Models and Logistic Regression*, 2nd Edition. Springer-Verlag. Chapter 9, Generalized Linear Models.

[2] For a review of homogeneous and non-homogeneous Poisson processes see: Ross, S. 1997. *Introduction to Probability Models*, 6th Edition, Academic Press. Section 5.3 and 5.4. Also see Kao, E. 1997. *An Introduction to Stochastic Processes*, Duxbury Press. Chapter 2.

DRAFT

We assume that the number of loans that will "claim" during a given year, out of the loans that are active[3] at the beginning year, is a function of a number of predictor variables. We further assume that the mean or expected number of claims during a given year is the parameter of a Poisson distribution. The Poisson distribution models the probability of *y* events, or claims, according to a Poisson process with the probability distribution function given by:

$$p(y;\lambda) = \frac{e^{-\lambda}\lambda^{y}}{y!}, \text{ for } y = 0,1,2,\ldots \qquad \textbf{(I.1).}$$

The mean or expected value of the Poisson distribution is λ; this is known as the Poisson parameter.

The Poisson parameter is dependent on a specified unit or period of time. For our model we assume that the basic unit of time is one year and that a given Poisson distribution only applies to this period. For example, it would be incorrect to assume that a specific Poisson parameter applies for a period of two or more years since each year will have its own "unique" Poisson distribution.

The mean number of claims, or the Poisson parameter, is a function of the predictor variables. Suppose the data takes the form:

$$\begin{matrix} y_1 & x_{11}x_{21}\cdots x_{k1} \\ \vdots & \vdots\cdots\cdots\vdots \\ y_n & x_{1n}x_{2n}\cdots x_{kn} \end{matrix},$$

where the y_i represent *n* observations of the response variable and the x_{ij} are the corresponding observed values of *k* predictor variables.

The model is then written $y_i = \lambda_i + \varepsilon_i$ for $i = 1, 2, \ldots, n$. The probability as a function of the predictor variables is:

$$p(y_i; x_i, \beta) = \frac{e^{-\lambda(x_i,\beta)}\left[\lambda(x_i,\beta)\right]^{y_i}}{y_i!}, \text{ for } y_i = 0, 1, 2, \ldots \qquad \textbf{(I.2).}$$

(In this notation β and x_i are vectors.) Here $\lambda(x_i,\beta)$ replaces our earlier λ. The function $\lambda(x_i,\beta)$ must always be non-negative. A candidate for this function is $e^{x_i'\beta}$, where $x_i'\beta$ is a linear function. $\lambda(x_i,\beta)$ relates the predictor variables to the mean. Then equation (A.1) is of the form $\ln(\lambda) = x_i' \cdot \beta$. Transforming the log link function we get the following expression for our response variable:

$$\lambda_t = e^{(a+b_1\cdot t+b_2\cdot t^2+b_3\cdot INT.RT_t+b_4\cdot R.GT1_t+b_5\cdot R.LT1_t+b_6\cdot CUMDIFF_t+b_7\cdot LTV_0+b_8\cdot LTV.AGE3_t+b_9\cdot ANN.HPA_t+b_{10}\cdot HPA_t} \qquad \textbf{(I.3).}$$
$$+b_{11}\cdot NEGEQ.RGT1_t+b_{12}\cdot NEGEQ.RLT1_t+b_{13}\cdot RHP+b_{14}\cdot UNEMP.L0_t+b_{15}\cdot PAY.INC.AGE4_t+b_{16}\cdot SR)$$

[3] By "active" we mean loans that enter a given year and have not claimed, prepaid or have been otherwise terminated.

DRAFT

Since $\text{var}(y_i) = e^{x_i' b}$ is not homogeneous from observation to observation, standard least squares does not apply. We use maximum likelihood methods. For the Poisson model the log-likelihood function is given by:

$$l(\mathbf{y}, \boldsymbol{b}) = \frac{\left\{\prod_{i=1}^{n} l(\mathbf{x_i}, \boldsymbol{b})^{y_i}\right\} e^{-\sum_{i=1}^{n} l(\mathbf{x_i}, b)}}{\prod_{i=1}^{n} y_i!}$$

We employ a generalized linear model (GLM) because our link function is non-linear and the error variance is not homogeneous. Since explicit expressions for the maximum likelihood estimators are not generally available, estimates are calculated using an iterative approach. As mentioned in Appendix A, a commonly used approach is *iteratively re-weighted least squares (IRWLS).*[4]

An outline of the IRWLS procedure is given below.[5]

1. Obtain an initial estimate of the coefficients and from this result obtain an initial estimate of the residuals. The initial estimate of the linear predictor is obtained using a standard linear model that checks for problems such as negative logarithms.
2. From the initial residuals, compute a variance estimate, $\hat{\sigma}_0^2$ (equal to the squared residual), and the initial weights, $w_{i,0} = \psi(e^*_{i,0}) / (e^*_{i,0})$. Here $\psi(\bullet)$ is the influence function.[6]
3. Use weighted least squares to obtain new robust parameter estimates.
4. Let the parameter estimates from step (3) take the role of the initial weights in step (1) and obtain new residuals, a new variance estimate, and new weights.
5. Return to and repeat step (3).
6. Repeat until the estimates converge. The convergence criterion is to stop if $\left|deviance^{i} - deviance^{i-1}\right| < \varepsilon$. In our model we set ε equal to 10^{-4}. The deviance for iteration i is defined as twice the log-likelihood ratio statistic; this is given by $2\sum_{i=1}^{n} A_i \left[\left\{y_i \theta(y_i) - g(\theta(y_i))\right\} - \left\{y_i \hat{\theta}_i - g(\hat{\theta}_i)\right\}\right]$.

---

[4] For a complete description of the IRWLS procedure please see McCullagh and Nelder, 1989, section 2.5. Another source is Stokes and Koch, 1983; *A Macro for Maximum Likelihood Fitting of Log-Linear Models to Poisson and Multinomial Counts*; Proceedings of the Eighth Annual SAS Users Group International; Cary, North Carolina: SAS Institute, pp. 795-800.

[5] Note that, although the procedure is conventionally known as iteratively re-weighted least squares, it is a maximum likelihood technique.

[6] An influence function estimates how individual data points affect regression results. We use a Huber influence function which is bounded: $\psi(e_i^*) = e_i^*$ if $|e_i^*| \le r$ and r if $e_i^* > r$ and $-r$ if $e_i^* < -r$. We set r =1 and $e_i^* = e_i / \sigma_i$. In OLS the influence function is the identity function. See Huber, P.J. 1973. Robust Regression: Asymptotics, Conjectures, and Monte Carlo. *Annals of Statistics* 1: 799-821.

DRAFT

Supposed Bias in GLM

It has been argued that log-linear models (like the Poisson model used in our analysis of conditional claim and prepayment rates) are biased, and therefore the results of such a model would need to be adjusted by some factor to correct for that bias. In the following text, we explain why this is not the case and that no adjustment is needed.[7]

The argument for log-linear bias begins with the statement that the log-linear model is specified as:

$$\ln(\lambda) = \beta \cdot x + \varepsilon$$

where epsilon represents the error term. However, the correct specification of a log-linear model is:

$$e^{\ln(\lambda)} = \beta \cdot x$$

The differences between the two specifications are:

1. The intention is to model the expected value of the response, in this case, the Poisson parameter, and

2. There is no error term because we are modeling the expected value.

In the Poisson regression model, the only variability is around the Poisson counts. There is no automatic bias in this estimate because of the "errors" in the linear model; no such errors are assumed to be present.

The suggestion of bias in the argument flows from the idea that $\ln(\lambda)$ is determined via a linear model with errors and that the observation is taken from a Poisson model with a λ so determined. While this "linear-model-with-errors-for-$\ln(\lambda)$" view could be appropriate, it does not correspond to the usual GLM. With the latter, observations are taken from a Poisson model where the log of the mean is exactly linear in the explanatory variables. No errors enter into the population values of $\ln(\lambda)$. The key modeling assumption under the usual GLM view of things is that $\ln(\lambda)$ corresponding to different values of x lie exactly on a straight line. Departures from straight line behavior in actual counts arise solely out of the Poisson variability at each λ.

If it is believed that the populations, cells in this case, being modeled are comprised of groups with varying Poisson parameters, then one would build a Mixed Effects model. In this type of model, we would introduce additional variability around the Poisson parameter - the most popular approach to doing this is to assume that the Poisson parameters are Gamma distributed in which case the model turns out to be a Negative Binomial.

[7] The explanation that follows was enhanced by discussions with several statisticians and econometricians from various academic institutions and professional service firms, including the Wharton School of the University of Pennsylvania, Oxford University, Virginia Tech, and Deloitte & Touche. In addition, a more introductory text is Dobson, Annette J. 1990. *An Introduction to Generalized Linear Models*, CRC Press.

DRAFT

With respect to the interpretation of $e^{b \cdot x}$ at a given value of x, and where b is estimated, this is both the estimated mean at the given x and a prediction of a given count at that x. The latter is of course "predicted" with much less certainty since actual counts deviate from the mean according to Poisson variability (which is equal to the mean). Another sometimes important issue is that b, and possibly x, are estimated, and hence subject to error. This implies the estimate of the mean is subject to estimation error which could be taken into account when making inferences. However, there is no consensus on how this should be accomplished - some suggest that the confidence interval around the mean should be increased. As a practical matter, the estimation error associated with b and x is often small in relation to the Poisson variability of the counts given l (that is, assuming b and x are known exactly).