Population Data Analysis

Atikur R. Khan
&
Sumaiya Abedin
Monash University & University of Rajshahi

Last Update: October, 2013

Contents

Chapter 1

Generalized Linear Model

Linear regression model assumes that the conditional expectation of $\mathbf{y}$ is a linear function $\mathbf{X}\boldsymbol{\beta}$ and equivalently the relationship is stated with an equation $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ where $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$. Because of the continuous distribution of $\boldsymbol{\epsilon}$, this model is only suitable for continuous response variable $\mathbf{y}$ but not for binary or count response variables. Generalised linear model has a fascinating property of dealing with response varibles from different distributions. GLM consists of three components:

1. A random component that specifies the conditional distribution of the response variable, given the values of the explanatory variables in the model. Distribution of response variables are widely adopted from exponential family such as Binomial and Poisson distributions.

2. A systematic component that is a linear predictor and is essentially a linear function of regressors $\eta_i = \mathbf{x}_i^T\boldsymbol{\beta}$

3. A parametric link component that is a smooth and invertible linearizing link function $g(\mu_i) = \eta_i = \mathbf{x}_i^T\boldsymbol{\beta}$, which transforms the expectation of the response variable, $\mu_i = E(y_i)$, to the linear predictor η_i.

1.1 Estimation of parameters from score equation

Let us assume that the response y_i is an observation from a family of exponential distributions. Thus the log-likelihood function for y_i is

$$\ell_i(\theta, \psi, y) = logf(y_i, \theta_i, \psi) = \frac{y_i\theta_i - b(\theta_i)}{a(\psi)} + c(y_i, \psi) \tag{1.1}$$

and that for the sample $y_1, \ldots, y_n$ is

$$\ell(\theta, \psi, y) = \sum_{i=1}^{n} \left(\frac{y_i\theta_i - b(\theta_i)}{a(\psi)} + c(y_i, \psi) \right), \tag{1.2}$$

where $\ell_i(\theta, \psi, y)$ denotes the individual log-likelihood contribution for the ith observation.

Let us assume that the expectation of response is $\mu_i = E(y_i) = h^{-1}(\theta_i)$ and $g(\mu_i) = \eta_i$ is the link function that transforms the expectation of response to the linear predictor $\eta_i = \mathbf{x}_i^T\boldsymbol{\beta}$. Thus the log-likelihood for GLM estimation process can be expressed as

$$\ell(\boldsymbol{\beta}, \psi, y) = \sum_{i=1}^{n} \left(\frac{y_i h(\mu_i) - b(h(\mu_i))}{a(\psi)} + c(y_i, \psi) \right), \tag{1.3}$$

From (1.3) we may note that $c(y_i, \psi)$ and $a(\psi)$ are independent of θ_i and so these are independent of parameters of our interest. To get estimating equations for the regression parameters, we have to differentiate the log-likelihood with respect to each coefficient in turn. Let ℓ_i represent the ith component of the log likelihood. Then, by the chain rule,

$$\frac{\delta\ell_i}{\delta\beta_j} = \frac{\delta\ell_i}{\delta\theta_i} \times \frac{\delta\theta_i}{\delta\mu_i} \times \frac{\delta\mu_i}{\delta\eta_i} \times \frac{\delta\eta_i}{\delta\beta_j}$$

for $j = 0, 1, \ldots, k$, which can be rewritten as

$$\frac{\delta\ell_i}{\delta\beta_j} = \frac{y_i - \mu_i}{v(\mu_i)} \frac{\delta\mu_i}{\delta\eta_i} \times x_{ij} \tag{1.4}$$

where $v(\mu_i) = a(\psi)b''(\theta_i)$. Summing the above (1.4) across all observation, we may write the score equation

$$s_j(\beta) = \sum_{i=1}^{n} \frac{y_i - \mu_i}{v(\mu_i)} \times \frac{d\mu_i}{d\eta_i} \times x_{ij} = 0 \text{ for } j = 0, 1, \ldots, k \tag{1.5}$$

which is known as the estimating equations for the generalised linear model and the maximum likelihood estimates are obtained by solving the score equations. A general method of solving score equations is the iterative algorithm based on Fisher's method of scoring derived from a Taylor's expansion of $\mathbf{s}(\beta)$. In the rth iteration , the new estimate $\beta^{(r+1)}$ is obtained from the previous estimate $\beta^{(r)}$ from the equation

$$\beta^{(r+1)} = \beta^{(r)} + \mathbf{s}(\beta^{(r)})E(H(\beta^{(r)}))^{-1}, \tag{1.6}$$

where H is the Hessian matrix, a matrix of second derivatives of the log-likelihood. The Hessian matrix is

$$H(\beta) = \begin{bmatrix} \frac{\delta s_1(\beta)}{\delta \beta_1} & \cdots & \frac{\delta s_1(\beta)}{\delta \beta_p} \\ \vdots & \ddots & \vdots \\ \frac{\delta s_p(\beta)}{\delta \beta_1} & \cdots & \frac{\delta s_p(\beta)}{\delta \beta_p} \end{bmatrix} = \begin{bmatrix} \frac{\delta^2 \ell}{\delta \beta_1 \delta \beta_1} & \cdots & \frac{\delta^2 \ell}{\delta \beta_1 \delta \beta_p} \\ \vdots & \ddots & \vdots \\ \frac{\delta^2 \ell}{\delta \beta_p \delta \beta_1} & \cdots & \frac{\delta^2 \ell}{\delta \beta_p \delta \beta_p} \end{bmatrix}.$$

where $\frac{\delta \ell}{\delta \beta_j} = \sum_{i=1}^{n} \frac{\delta \ell_i}{\delta \beta_j} = \sum_{i=1}^{n} \left(\frac{y_i - \mu_i}{v_i} \right) \frac{\delta \mu_i}{\delta \eta_i} x_{ij}$ and

$$\begin{aligned} \frac{\delta^2 \ell}{\delta \beta_j \delta \beta_k} &= \frac{\delta \ell}{\delta \beta_k} \left[\sum_{i=1}^{n} \left(\frac{y_i - \mu_i}{v_i} \right) \frac{\delta \mu_i}{\delta \eta_i} x_{ij} \right] \\ &= \sum_{i=1}^{n} (y_i - \mu_i) \frac{\delta}{\delta \beta_k} \left(v_i^{-1} \frac{\delta \mu_i}{\delta \eta_i} x_{ir} \right) + \sum_{i=1}^{n} v_i^{-1} \frac{\delta \mu_i}{\delta \eta_i} x_{ir} \frac{\delta}{\delta \beta_k} (y_i - \mu_i). \end{aligned}$$

Since $\frac{\delta}{\delta \beta_k}(y_i - \mu_i) = -\frac{\delta \mu_i}{\delta \eta_i} \frac{\delta \eta_i}{\delta \beta_k} = -\frac{\delta \mu_i}{\delta \eta_i} x_{ik}$ we may note that

$$E\left(\frac{\delta^2 \ell}{\delta \beta_j \delta \beta_k} \right) = -E\left[\sum_{i=1}^{n} v_i^{-1} \left(\frac{\delta \mu_i}{\delta \eta_i} \right)^2 x_{is} x_{ir} \right] = -\sum_{i=1}^{n} w_i x_{is} x_{ir}$$

where $\mathbf{W} = diag(w_1, \ldots, w_n)$ is a diagonal matrix of weights. Thus $A(\beta) = E(H(\beta)) = -E\left(\frac{\delta^2 \ell}{\delta \beta_j \delta \beta_k} \right) = \mathbf{X}^T \mathbf{W} \mathbf{X}$ and by Fisher scoring method

$$\beta^{(r+1)} = \beta^{(r)} + A^{-1}(\beta^{(r)}) s(\beta^{(r)}).$$

It can be deduced from the above equation that

$$A(\beta^{(r)})\beta^{(r+1)} = \sum w_i^{(r)x_{ij}} \left[\sum_{s=1}^{p} x_{is} b_s^{(r)} + (y_i - \mu_i^{(r)}) \frac{\delta \eta_i^{(r)}}{\delta \mu_i^{(r)}} \right]$$

and further that

$$A(\beta^{(r)})\beta^{(r+1)} = \sum w_i^{(r)x_{ij}} \sum_{s=1}^{p} x_{is} b_s^{(r+1)} = \sum w_i^{(r)x_{ij}} \eta_i^{(r+1)}$$

Thus we may write

$$\sum_{i=1}^{n} w_i^{(r)} x_{ij} Z_i^{(r)} = \sum_{i=1}^{n} w_i^{(r)} x_{ij} \eta_i^{(r+1)}$$

and in matrix form this is

$$\mathbf{X}^T \mathbf{W}^{(r)} \mathbf{X} \beta^{(r+1)} = \mathbf{X}^T \mathbf{W}^{(r)} \mathbf{Z}^{(r)}$$

and it turns out that the updates can be written as

$$\beta^{(r+1)} = (\mathbf{X}^T \mathbf{W}^{(r)} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W}^{(r)} \mathbf{Z}^{(r)}$$

with similar kind of iteratively reweighted least squares of z_i on $\mathbf{X}$ with weights $\mathbf{W}^{(r)} = diag(w_i^{(r)})$ and $z_i^{(r)} = \eta_i^{(r)} + (y_i - \mu_i^{(r)}) g'(\mu_i^{(r)})$.

Since $A(\boldsymbol{\beta}^{(r)})\boldsymbol{\beta}^{(r+1)} = \mathbf{X}^T \mathbf{W} \mathbf{X} \boldsymbol{\beta}^{(r+1)}$ and $A(\boldsymbol{\beta}^{(r)})\boldsymbol{\beta}^{(r)} + s(\boldsymbol{\beta}^{(r)}) = \mathbf{X}^T \mathbf{W} \mathbf{Z}^{(r)}$, the iterative algorithm for confidentialised estimates is

$$\mathbf{X}^T \mathbf{W} \mathbf{X} \boldsymbol{\beta}^{(r+1)} = \mathbf{X}^T \mathbf{W} \mathbf{Z}^{(r)} + \mathbf{e}^*$$

and the iteration

$$\boldsymbol{\beta}^{(r+1)} = (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W} \mathbf{Z}^{(r)} + (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{e}^* \tag{1.7}$$

continues until a pre-selected level of accuracy in estimation, say for example,

$$\sqrt{\frac{(\widehat{\boldsymbol{\beta}}^{(r+1)} - \widehat{\boldsymbol{\beta}}^{(r)})^T (\widehat{\boldsymbol{\beta}}^{(r+1)} - \widehat{\boldsymbol{\beta}}^{(r)})}{\widehat{\boldsymbol{\beta}}^{(r)^T} \widehat{\boldsymbol{\beta}}^{(r)}}} \leq 10^{-10}$$

is satisfied. An iteratively reweighted algorithm for parameter estimation can be stated as:

1. Start with the initial estimates μ_i, compute $\hat{\eta}_i = g(\hat{\mu}_i)$, denote them by $\hat{\mu}_i^{(0)}$ and $\hat{\eta}_i^{(0)}$. Simple start option for initialisation could be $\hat{\mu}_i^{(0)} = y_i + 0.5$ (but this can be different based on the data type and estimation complexity)

2. Calculate working responses
$$z_i^{(r)} = \eta_i^{(r)} + (y_i - \mu_i^{(r)})g'(\mu_i^{(r)})$$
and working weights
$$w_i^{(r)} = \frac{1}{\left[g'(\mu_i^{(r)})\right]^2 a_i(\phi)v(\mu_i^{(r)})}$$

3. Calculate $\beta^{(r+1)}$ by weighted least squares estimates
$$\beta^{(r+1)} = (X^T W^{(r)} X)^{-1} X^T W^{(r)} z^{(r)}$$

4. Repeat 2 and 3 until the regression coefficients stabilise, at which stage the estimates converge to the maximum likelihood estimate of β.

1.2 Estimation of Dispersion Parameter

The maximum-likelihood estimating equations for generalized linear models take the common form
$$\sum_{i=1}^{n} \frac{y_i - \mu_i}{a_i v(\mu_i)} \times \frac{d\mu_i}{d\eta_i} \times x_{ij} = 0 \text{ for } j = 0, 1, \ldots, k$$

These equations are generally nonlinear and therefore have no general closed-form solution, but they can be solved by iterated weighted least squares (IWLS). The estimating equations for the coefficients do not involve the dispersion parameter, which (for models in which the dispersion is not fixed) then can be estimated as
$$\tilde{\phi} = \frac{1}{n-k-1} \sum_{i=1}^{n} \frac{(Y_i - \hat{\mu}_i)^2}{a_i v(\hat{\mu}_i)}$$

The estimated asymptotic covariance matrix of the coefficients is
$$\hat{V}(\mathbf{b}) = \tilde{\phi}(\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1}$$
where $\mathbf{b}$ is the vector of estimated coefficients and $\mathbf{W}$ is a diagonal matrix of weights from the last IWLS iteration.

1.3 Asymptotic distribution of maximum likelihood estimator

The estimates $\hat{\beta}$ have the usual properties of maximum likelihood estimators. In particular, $\hat{\beta}$ is asymptotically normal:

$$\hat{\beta} \sim N(\beta, \phi(X^T W X)^{-1})$$

where ϕ is unknown and should be estimated in order to estimate the covariance matrix $Cov(\hat{\beta}) = \phi(X^T W X)^{-1}$. Given the results for W in the final iteration and estimated dispersion parameter $\hat{\phi}$ (as has been estimated in the previous section), we may estimate the covariance matrix $\widehat{Cov}(\hat{\beta}) = \hat{\phi}(X^T W X)^{-1}$.

1.4 Binary Response Model

Suppose that the response variable Y_i is a 0-1 random variable (binary response 0 or 1), then the link function

$$g(\mu_i) = logit(\mu_i) = log\left(\frac{\mu_i}{1-\mu_i}\right)$$

and the variance function is

$$V(\mu_i) = \mu_i(1-\mu_i).$$

Since $\eta_i = \mathbf{x}_i^T \boldsymbol{\beta}$, we may note for $\eta_i = g(\mu_i)$ that $\mu_i = \frac{e^{\mathbf{x}_i^T \boldsymbol{\beta}}}{e^{\mathbf{x}_i^T \boldsymbol{\beta}_{+1}}}$. Also, we may deduce that $g'(\mu_i) = \frac{1}{\mu_i(1-\mu_i)} = \frac{1}{V(\mu_i)}$. Thus the estimation process could be continued with the initial set up:

$$\begin{aligned}
\mu_i^{(0)} &= (y_i + 0.5)/2 \\
\eta_i^{(0)} &= g(\mu_i^{(0)}) = log\left(\frac{\mu_i^{(0)}}{1-\mu_i^{(0)}}\right) \\
z_i^{(0)} &= \eta_i^{(0)} + (y_i - \mu_i^{(0)}) g'(\mu_i^{(0)}) = \eta_i^{(0)} + \frac{y_i - \mu_i^{(0)}}{\mu_i^{(0)}(1-\mu_i^{(0)})} \\
w_i^{(0)} &= \frac{1}{[g'(\mu_i^{(0)})]^2 a(\psi) V(\mu_i^{(0)})} = \mu_i^{(0)}(1-\mu_i^{(0)})
\end{aligned}$$

By employing this initialisation, we may estimate $\widehat{\boldsymbol{\beta}}$ by using the iterative algorithm of solving the score function. Once the parameter $\widehat{\boldsymbol{\beta}}$, dispersion parameter ϕ, and covariance matrix $\phi(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}$ are estimated it is straight forward to compute $\boldsymbol{\Omega}$ for some pre-selected $e\ell(\widehat{\boldsymbol{\beta}}_{PSE})$ vector. Finally, we estimate the coefficients by solving perturbed score equation described in Section 3.

Remark 1. For a probit model, $g(\mu_i) = \Phi^{-1}(\mu_i)$ and the variance function $V(\mu_i) = \mu_i(1-\mu_i)$ are used, where Φ stands for the CDF of $N(0,1)$. So the perturbation process described in this section can be easily extended to the probit model.

1.5 Count Response Model

Assuming that the number of occurances of an event (say for example, number of visits to see a doctor or number of medicare claims in a year) Y_i follows Poisson distribution with parameter μ_i, a Poisson model is fitted with a suitable link function. The most commonly adopted link function for count response models is $\log(\mu_i)$. Thus for a GLM of Poisson distribution of counts, we consider the link function $g(\mu_i) = log(\mu_i)$, variance function $V(\mu_i) = \mu_i$, and the linear predictor $\eta_i = g(\mu_i) = \mathbf{x}_i^T\boldsymbol{\beta}$. Since $g'(\mu_i) = \frac{1}{\mu_i}$, $\mu_i = g^{-1}(\eta_i) = e^{\mathbf{x}_i^T\boldsymbol{\beta}}$, and $a(\psi) = 1$, we may start with an initialisation with $\mu_i^{(0)} = (y_i + 0.5)/2$, $\eta_i^{(0)} = log\left(\mu_i^{(0)}\right)$, $z_i^{(0)} = \eta_i^{(0)} + (y_i - \mu_i^{(0)})/\mu_i^{(0)}$, and $w_i^{(0)} = \mu_i^{(0)}$ to obtain $\widehat{\boldsymbol{\beta}}$ from an iteratively reweighted algorithm described earlier.

1.6 Practical experimentation

For practical experimentation for binary and count response models we consider two datasets: Heart and Estrogen/Progestin Study (HERS) data and US National Medical Expenditure Survey (NMES) data. The HERS dataset is used to demonstrate the proposed perturbation process for binary response models and the NMES dataset is used to fit a count response model.

1.6.1 Heart and Estrogen/Progestin Study Data

The HERS dataset comes from a clinical trial of hormone therapy for prevention of recurrent heart attacks and deaths among 2,763 post-menopausal women with existing coronary heart disease (**?**). **?** used this dataset in their book to fit a generalized linear model and the dataset can be downloaded from the website `http://www.epibiostat.ucsf.edu/biostat/vgsm/data.html`. Preexisting medical condition (`medcond: yes or no`) is a very sensitive personal information and sometimes very important to assess the amount of rebate from private health insurance. A binary response model can be fitted to predict whether the individuals had a medical condition (`medcond`), given the information on other variables: the age of the respondents (`age`), regular exercise (`exercise: yes or no`), having diabetes (`diabetes: yes or no`), and drinking habit (`drinkany: yes or no`). `R commands` to produce results presented in Table 1.2 are provided in the following box.

```
HERS <- read.csv("C:/Users/atikur/hersdata.csv", na.strings="")
hers.data<- data.frame( HERS$age, HERS$raceth, HERS$nonwhite,
                  HERS$smoking,  HERS$drinkany,  HERS$exercise,
                  HERS$diabetes, HERS$statins, HERS$medcond )
colnames(hers.data)<- c("age", "raceth", "nonwhite", "smoking",
                       "drinkany", "exercise", "diabetes",
                       "statins", "medcond")
nrow(hers.data)
ncol(hers.data)
hers.clean<- na.omit(hers.data)
summary(glm(hers.clean$medcond ~ hers.clean$age + hers.clean$diabetes
                + hers.clean$exercise + hers.clean$drinkany,
                  family=binomial(link="logit")))$coef
```

1.6.2 US National Medical Expenditure Survey Data

This dataset comes from the US National Medical Expenditure Survey (NMES) for 1987, and is available at `http://www.jstatsoft.org/v16/i09/`. There are 4,406 individuals in the dataset who are aged 66 and over, and are covered by the public insurance program (Medicare). **?** used this dataset to model the count