Generalized, Linear, and Mixed Models. By Charles E. McCulloch and Shayle R. Searle
 ISBN: 0-471-19364-X

Chapter 5

GENERALIZED LINEAR MODELS (GLMs)

5.1 INTRODUCTION

Models for the analysis of non-normal data using nonlinear models have a long history. The use of probit regression for a binary response is a classic example. The word *probit* was traced by David (1995) as far back as Bliss (1934). Finney (1952) attributes the actual origin of probit regression to psychologists in the late 1800s.

In an early example of probit regression, Bliss (1934) describes an experiment in which nicotine is applied to aphids and the proportion killed is recorded (how is that for an early antismoking message?). As an appendix to a paper Bliss wrote a year later (Bliss, 1935), Fisher (1935) outlines the use of maximum likelihood to obtain estimates of the probit model.

However it was years before maximum likelihood estimation for probit models caught on. Finney (1952), in an appendix entitled "Mathematical basis of the probit method" gives some of the rational for maximum likelihood and motivates a computational method that he spends six pages describing in a different appendix.

More specifically, if we let p_i denote the probability of a success for the ith observation, the probit model is given by

$$\begin{aligned} y_i &\sim \text{indep. Bernoulli}(p_i) \\ p_i &= \Phi\left(\mathbf{x}_i'\boldsymbol{\beta}\right), \end{aligned} \tag{5.1}$$

where $\mathbf{x}_i'$ denotes the ith row of a matrix of predictors and $\Phi(\cdot)$ is the standard normal c.d.f. Considering the scalar functions applied elementwise to the vectors, we can rewrite (5.1) as

$$\begin{aligned} \mathbf{y} &\sim \text{indep. Bernoulli}(\mathbf{p}) \\ \mathbf{p} &= \Phi(\mathbf{X}\boldsymbol{\beta}) \end{aligned} \tag{5.2}$$

or equivalently

$$\Phi^{-1}(\mathbf{p}) = \mathbf{X}\boldsymbol{\beta},$$

where $\mathbf{X}$ is the model matrix. The use of the inverse standard normal c.d.f., known as the probit, to transform the mean of $\mathbf{y}$ to the linear predictor is attractive on two counts. First, it expands the range of $\mathbf{p}$ from [0,1] to the whole real line, making it more reasonable to assume a model of the form $\mathbf{X}\boldsymbol{\beta}$. Second, in many problems, the sigmoidal form of $\mathbf{p}$ as a function of the covariates is often observed in practice.

Finney (1952) suggested calculating an estimate of $\boldsymbol{\beta}$ via an iteratively weighted least squares algorithm. He recommended using *working probits* which he defined (ignoring the shift of five units historically used to keep all the calculations positive) as

$$t_i = \mathbf{x}_i'\boldsymbol{\beta} + \frac{y_i - \Phi(\mathbf{x}_i'\boldsymbol{\beta})}{\phi(\mathbf{x}_i'\boldsymbol{\beta})}, \tag{5.3}$$

where $\phi(\cdot)$ is the standard normal probability density function (p.d.f.). The working probits for a current value of $\boldsymbol{\beta}$ were regressed on the predictors using weights given by $\dfrac{[\phi(p_i)]^2}{\Phi(p_i)[1-\Phi(p_i)]}$ (see E 5.1) in order to get the new value of $\boldsymbol{\beta}$. This algorithm was iterated until convergence (or at least until the computer – a person! – got tired of performing the calculations).

Nelder and Wedderburn (1972) recognized that the working probits could be generalized in a straightforward way to unify an entire collection of maximum likelihood problems. This *generalized linear model* (GLM) could handle probit or logistic regression, Poisson regression, log-linear models for contingency tables, variance components estimation from ANOVA mean squares and many other problems in the same way.

They replaced $\Phi^{-1}(\cdot)$ with a general *link* function, $g(\cdot)$, which transforms (or links) the mean of y_i to the linear predictor. With $g_\mu(\mu)$ representing $\partial g(\mu)/\partial\mu$, they then defined a *working variate* via

$$\begin{aligned} t_i &\equiv g(\mu_i) + g_\mu(\mu_i)(y_i - \mu_i) \\ &= \mathbf{x}_i'\boldsymbol{\beta} + g_\mu(\mu_i)(y_i - \mu_i). \end{aligned} \tag{5.4}$$

Since the second term on the right-hand side of (5.4) has expectation zero it can be regarded as an *error term* so that t_i follows a linear model, albeit with unequal variances which depend on the unknown $\boldsymbol{\beta}$. This suggests using (5.4) just like (5.3): regress $\mathbf{t}$ on $\mathbf{X}$ using a weighted linear regression (more details are given in Section 5.4e) and iterate until the estimates of $\boldsymbol{\beta}$ stabilize.

More important, it made possible a style of thinking which freed the data analyst from necessarily looking for a transformation which simultaneously achieved linearity in the predictors and normality of the distribution (as in Box and Cox, 1962).

What advantages does this have? First, it unifies what appear to be very different methodologies, which helps us to understand, use and (for those of us in the business) teach the techniques. Second, since the right-hand side of the model equation is a linear model after applying the link, many of the standard ways of thinking about linear models carry over to GLMs.

5.2 STRUCTURE OF THE MODEL

Building a generalized linear model involves three decisions:

1. What is the distribution of the data (for fixed values of the predictors and possibly after a transformation)?

2. What function of the mean will be modeled as linear in the predictors?

3. What will the predictors be?

a. Distribution of y

Typically the vector $\mathbf{y}$ is assumed to consist of independent measurements from a distribution with density from the exponential family or

similar to the exponential family:

$$\begin{aligned} y_i &\sim \text{indep. } f_{Y_i}(y_i) \\ f_{Y_i}(y_i) &= \exp\{[y_i\gamma_i - b(\gamma_i)]/\tau^2 - c(y_i,\tau)\}, \end{aligned} \tag{5.5}$$

where, for convenience, we have written the distribution in what is called *canonical form.* For example, for the probit model, the data would be independent Bernoulli so that $f_{Y_i}(y_i)$ would be $p_i^{y_i}(1-p_i)^{1-y_i}$, where p_i is the probability of a success and $\gamma_i = \log[p_i/(1-p_i)]$. Most commonly-used distributions can be written in the form (5.5) (see E 5.2).

b. Link function

We typically want to relate the parameters of the distribution to various predictors. We do so by modeling a transformation of the mean, μ_i, which would be some function of γ_i, as a linear model in the predictors:

$$\begin{aligned} \mathrm{E}[y_i] &= \mu_i \\ g(\mu_i) &= \mathbf{x}_i'\boldsymbol{\beta}, \end{aligned} \tag{5.6}$$

where $g(\cdot)$ is a known function, called the *link function* (since it links together the mean of y_i and the linear form of predictors), $\mathbf{x}_i'$ is the ith row of the model matrix, and $\boldsymbol{\beta}$ is the parameter vector in the linear predictor. In the probit example $g(\mu) = \Phi^{-1}(\mu)$ and $\mu = 1/(1+\exp[-\gamma])$.

c. Predictors

In practice, of course, one must make decisions as to which predictors to include on the right-hand side of (5.6) and in what form to include them. For example, in the classic paper of Bliss (1934) the suggested predictor of survival is log nicotine dose as opposed to nicotine itself.

A key point in using GLMs is that many of the considerations in modeling are the same as for LMMs since the right-hand sides of the model equations for the mean are the same. For example, issues of how to represent predictors and interactions, whether and how to model nonlinear relationships and (as we will see in Chapter 8) the incorporation of random factors.

d. Linear models

This generalized class of models subsumes the linear model of Chapter 4 as a special case. The normal distribution can be written in the form (5.5) by defining:

$$\begin{aligned} \gamma_i &= \mu_i \\ b(\gamma_i) &= \tfrac{1}{2}\mu_i^2 \\ \tau^2 &= \sigma^2 \\ c(y_i, \tau) &= \tfrac{1}{2}\log 2\pi\sigma^2 + \tfrac{1}{2}y_i^2/\sigma^2. \end{aligned} \tag{5.7}$$

With $g(\mu_i) = \mu_i$ and $\mu_i = \mathbf{x}_i'\beta$ we generate the linear model of Section 4.3.

5.3 TRANSFORMING VERSUS LINKING

In its earliest incarnations, probit analysis was little more than a transformation technique. It was realized that the frequent sigmoidal shape in plots of observed proportions of successes plotted against a predictor x could be made into a straight line by applying a transformation corresponding to the inverse of the normal c.d.f. However, one of the main ideas of GLMs is to get away from the idea of transforming the data. The strategy, then, is to apply a link function to the mean of the response and fit the resulting model by the method of maximum likelihood.

5.4 ESTIMATION BY MAXIMUM LIKELIHOOD

a. Likelihood

The log likelihood for (5.5) is given by

$$l = \sum_{i=1}^{n}[y_i\gamma_i - b(\gamma_i)]/\tau^2 - \sum_{i=1}^{n} c(y_i, \tau). \tag{5.8}$$

b. Some useful identities

Before we derive the maximum likelihood equations it is useful to establish some identities. These flow from the results

$$\mathrm{E}\left[\frac{\partial \log f_{Y_i}(y_i)}{\partial \gamma_i}\right] = 0, \tag{5.9}$$

and

$$\mathrm{var}\left(\frac{\partial \log f_{Y_i}(y_i)}{\partial \gamma_i}\right) = -\mathrm{E}\left[\frac{\partial^2 \log f_{Y_i}(y_i)}{\partial \gamma_i^2}\right], \tag{5.10}$$

which require regularity conditions (Casella and Berger, 1990, p. 308). Using (5.5) in (5.9) gives

$$\mathrm{E}\left[\left\{y_i - \frac{\partial b(\gamma_i)}{\partial \gamma_i}\right\} \Big/ \tau^2\right] = 0 \tag{5.11}$$

or

$$\mathrm{E}[y_i] = \mu_i = \frac{\partial b(\gamma_i)}{\partial \gamma_i}. \tag{5.12}$$

And using (5.5) in (5.10) we obtain

$$\mathrm{var}\left(\left\{y_i - \frac{\partial b(\gamma_i)}{\partial \gamma_i}\right\} \Big/ \tau^2\right) = -\mathrm{E}\left[-\frac{1}{\tau^2}\frac{\partial^2 b(\gamma_i)}{\partial \gamma_i^2}\right], \tag{5.13}$$

which, using (5.12) gives

$$\mathrm{var}\left(\frac{y_i - \mu_i}{\tau^2}\right) = \frac{1}{\tau^2}\frac{\partial^2 b(\gamma_i)}{\partial \gamma_i^2}$$

or

$$\begin{aligned} \mathrm{var}(y_i) &= \tau^2 \frac{\partial^2 b(\gamma_i)}{\partial \gamma_i^2} \qquad (5.14) \\ &\equiv \tau^2 v(\mu_i), \end{aligned}$$

wherein we define $v(\mu_i)$ as $\partial^2 b(\gamma_i)/\partial \gamma_i^2$. Note that $v(\mu_i)$ is often called the *variance function*, since it indicates how the variance of y_i depends on the mean of y_i. Two other useful identities are

$$\frac{\partial \gamma_i}{\partial \mu_i} = \left(\frac{\partial \mu_i}{\partial \gamma_i}\right)^{-1} = \left(\frac{\partial^2 b(\gamma_i)}{\partial \gamma_i^2}\right)^{-1} = \frac{1}{v(\mu_i)} \tag{5.15}$$

and, using the chain rule and (5.6),

$$\frac{\partial \mu_i}{\partial \boldsymbol{\beta}} = \frac{\partial \mu_i}{\partial g(\mu_i)} \frac{\partial g(\mu_i)}{\partial \boldsymbol{\beta}} = \left(\frac{\partial g(\mu_i)}{\partial \mu_i}\right)^{-1} \frac{\partial \mathbf{x}_i' \boldsymbol{\beta}}{\partial \boldsymbol{\beta}}$$

$$= \left(\frac{\partial g(\mu_i)}{\partial \mu_i}\right)^{-1} \mathbf{x}_i' \ . \tag{5.16}$$

As an illustration of these results, consider the linear model in Section 5.1d. With subscripts denoting derivatives we have $b_\gamma(\gamma_i)$ equal to μ_i, the mean, and $b_{\gamma\gamma}(\gamma_i) = 1$ so that, from (5.14), $\text{var}(y_i) = \tau^2 b_{\gamma\gamma}(\gamma_i) = \sigma^2$, as expected. Also, $\partial\gamma_i/\partial\mu_i = \partial\mu_i/\partial\mu_i = 1 = v(\mu_i)^{-1}$, verifying (5.15) and, with $g_\mu(\mu_i) = 1$, $\partial\mu_i/\partial\boldsymbol{\beta} = \mathbf{x}_i'$ as in (5.16). Note that the normal distribution has an unusual feature among distributions given by (5.5): its variance is a constant and not a function of the mean.

c. Likelihood equations

We are now in a position to derive the maximum likelihood equations for $\boldsymbol{\beta}$. From (5.8) we have

$$\frac{\partial l}{\partial \boldsymbol{\beta}} = \frac{1}{\tau^2} \sum \left[y_i \frac{\partial \gamma_i}{\partial \boldsymbol{\beta}} - \frac{\partial b(\gamma_i)}{\partial \gamma_i} \frac{\partial \gamma_i}{\partial \boldsymbol{\beta}} \right]$$

$$= \frac{1}{\tau^2} \sum (y_i - \mu_i) \frac{\partial \gamma_i}{\partial \boldsymbol{\beta}} \qquad \text{using (5.12)}$$

$$= \frac{1}{\tau^2} \sum (y_i - \mu_i) \frac{\partial \gamma_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \boldsymbol{\beta}} \qquad \text{using the chain rule}$$

$$= \frac{1}{\tau^2} \sum \frac{(y_i - \mu_i)}{v(\mu_i) g_\mu(\mu_i)} \mathbf{x}_i' \qquad \text{using (5.15) and (5.16)}$$

$$= \frac{1}{\tau^2} \sum (y_i - \mu_i) w_i g_\mu(\mu_i) \mathbf{x}_i', \tag{5.17}$$

upon defining $w_i = [v(\mu_i) g_\mu^2(\mu_i)]^{-1}$.

We can write this in matrix notation as

$$\frac{\partial l}{\partial \boldsymbol{\beta}} = \frac{1}{\tau^2} \mathbf{X}' \mathbf{W} \boldsymbol{\Delta} (\mathbf{y} - \boldsymbol{\mu}), \tag{5.18}$$

with $\mathbf{W} = \left\{ {}_d w_i \right\}$ and $\mathbf{\Delta} = \left\{ {}_d g_\mu(\mu_i) \right\}$.

The ML equations are thus given by

$$\mathbf{X}'\mathbf{W}\mathbf{\Delta}\mathbf{y} = \mathbf{X}'\mathbf{W}\mathbf{\Delta}\boldsymbol{\mu}, \tag{5.19}$$

where $\mathbf{W}, \mathbf{\Delta}$ and $\boldsymbol{\mu}$ involve the unknown $\boldsymbol{\beta}$. Typically these are nonlinear functions of $\boldsymbol{\beta}$ and so (5.19) cannot be solved analytically.

For example, for the probit model of (5.2), the log likelihood and its derivative are

$$l = \sum \left(y_i \left\{ \log \Phi(\mathbf{x}_i'\boldsymbol{\beta}) - \log[1 - \Phi(\mathbf{x}_i'\boldsymbol{\beta})] \right\} + \log[1 - \Phi(\mathbf{x}_i'\boldsymbol{\beta})] \right) \tag{5.20}$$

and

$$\begin{aligned} \frac{\partial l}{\partial \boldsymbol{\beta}} &= \sum \left[y_i \left(\frac{\phi(\mathbf{x}_i'\boldsymbol{\beta})}{\Phi(\mathbf{x}_i'\boldsymbol{\beta})} \mathbf{x}_i' + \frac{\phi(\mathbf{x}_i'\boldsymbol{\beta})}{1 - \Phi(\mathbf{x}_i'\boldsymbol{\beta})} \mathbf{x}_i' \right) - \frac{\phi(\mathbf{x}_i'\boldsymbol{\beta})}{1 - \Phi(\mathbf{x}_i'\boldsymbol{\beta})} \mathbf{x}_i' \right] \\ &= \sum \frac{[y_i - \Phi(\mathbf{x}_i'\boldsymbol{\beta})]\phi(\mathbf{x}_i'\boldsymbol{\beta})}{\Phi(\mathbf{x}_i'\boldsymbol{\beta})[1 - \Phi(\mathbf{x}_i'\boldsymbol{\beta})]} \mathbf{x}_i' \\ &= \sum \frac{(y_i - \mu_i)\phi(\mathbf{x}_i'\boldsymbol{\beta})}{\mu_i(1 - \mu_i)} \mathbf{x}_i'. \end{aligned} \tag{5.21}$$

Identifying $b(\gamma_i)$ of (5.5) as $\log(1+e^{\gamma_i})$ so that $b_\gamma(\gamma_i) = (1+e^{-\gamma_i})^{-1} = \mu_i$ and $b_{\gamma\gamma}(\gamma_i) = \mu_i(1 - \mu_i)$, it is straightforward (see E 5.4) to show that (5.21) is of the form of (5.18).

For solving the ML equations or for deriving the large-sample variance of $\hat{\boldsymbol{\beta}}$, it is useful to have the expected value of the second derivative of the log likelihood:

$$\frac{\partial^2 l}{\partial \boldsymbol{\beta}\, \partial \boldsymbol{\beta}'} = -\frac{1}{\tau^2} \mathbf{X}'\mathbf{W}\mathbf{\Delta} \frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\beta}'} + \frac{1}{\tau^2} \mathbf{X}' \frac{\partial \mathbf{W}\mathbf{\Delta}}{\partial \boldsymbol{\beta}'} (\mathbf{y} - \boldsymbol{\mu}) \tag{5.22}$$

so that

$$\begin{aligned} -\mathrm{E}\left[\frac{\partial^2 l}{\partial \boldsymbol{\beta}\, \partial \boldsymbol{\beta}'} \right] &= \frac{1}{\tau^2} \mathbf{X}'\mathbf{W}\mathbf{\Delta} \frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\beta}'} + \mathbf{0} \\ &= \frac{1}{\tau^2} \mathbf{X}'\mathbf{W}\mathbf{\Delta}\mathbf{\Delta}^{-1}\mathbf{X} \quad \text{using (5.16)} \\ &= \frac{1}{\tau^2} \mathbf{X}'\mathbf{W}\mathbf{X}, \end{aligned} \tag{5.23}$$

where, again, $\mathbf{W} = \left\{ {}_d w_i \right\} = \left\{ {}_d [v(\mu_i) g_\mu^2(\mu_i)]^{-1} \right\}$.

d. Large-sample variances

To derive the large-sample variance of $\hat{\boldsymbol{\beta}}$ we first note that

$$\begin{aligned} -\mathrm{E}\left[\frac{\partial^2 l}{\partial \boldsymbol{\beta}\, \partial \tau^2}\right] &= -\mathrm{E}\left[\frac{\partial}{\partial \tau^2}\frac{1}{\tau^2}\mathbf{X}'\mathbf{W}\boldsymbol{\Delta}(\mathbf{y}-\boldsymbol{\mu})\right] \\ &= \frac{1}{\tau^4}\mathbf{X}'\mathbf{W}\boldsymbol{\Delta}\mathrm{E}[\mathbf{y}-\boldsymbol{\mu}] \qquad (5.24) \\ &= \mathbf{0}, \end{aligned}$$

so that estimation of τ^2 does not affect the large-sample variance of $\hat{\boldsymbol{\beta}}$. The usual large-sample arguments (see Section S.4c of Appendix S), along with (5.23) and (5.24), show that (see E 5.6)

$$\mathrm{var}_\infty(\hat{\boldsymbol{\beta}}) = \tau^2(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}, \qquad (5.25)$$

where var_∞ indicates the limiting or asymptotic variance.

e. Solving the ML equations

Solution of the ML equations, (5.19), for $\boldsymbol{\beta}$ is usually performed by an iterative weighted least squares method. This can be derived as an example of the use of Fisher scoring (Searle et al., 1992, p. 295). Fisher scoring is an iterative method for maximizing a likelihood and it takes the form

$$\boldsymbol{\theta}^{(m+1)} = \boldsymbol{\theta}^{(m)} + \mathbf{I}(\boldsymbol{\theta}^{(m)})^{-1}\frac{\partial l}{\partial \boldsymbol{\theta}}\bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(m)}}, \qquad (5.26)$$

where (m) indicates the mth iteration, $\mathbf{I}(\boldsymbol{\theta})$ is the information matrix and $\boldsymbol{\theta}$ is the entire parameter vector.

Using (5.24), (5.23), and (5.18), the portion of the equation for $\boldsymbol{\beta}$ (see E 5.7) is of the form

$$\boldsymbol{\beta}^{(m+1)} = \boldsymbol{\beta}^{(m)} + (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\boldsymbol{\Delta}(\mathbf{y}-\boldsymbol{\mu}), \qquad (5.27)$$

where it is understood that $\mathbf{W}$, $\boldsymbol{\Delta}$, and $\boldsymbol{\mu}$ are evaluated at $\boldsymbol{\beta}^{(m)}$.

How does this relate to the working variate of (5.4)? We have

$$\mathbf{t} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\Delta}(\mathbf{y}-\boldsymbol{\mu}) \qquad (5.28)$$

so that, with the use of (5.14)

$$\mathrm{var}(\mathbf{t}) = \mathrm{var}[\boldsymbol{\Delta}(\mathbf{y}-\boldsymbol{\mu})] = \left\{{}_d\tau^2 v(\mu_i) g'^2_\mu(\mu_i)\right\} = \tau^2\mathbf{W}^{-1}, \qquad (5.29)$$

so a weighted regression of $\mathbf{t}$ on $\mathbf{X}$ using weights equal to the inverse of the variance of $\mathbf{t}$ gives

$$\begin{aligned}\boldsymbol{\beta}^{(m+1)} &= (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}[\mathbf{X}\boldsymbol{\beta}^{(m)} + \boldsymbol{\Delta}(\mathbf{y}-\boldsymbol{\mu})] \\ &= \boldsymbol{\beta}^{(m)} + (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\boldsymbol{\Delta}(\mathbf{y}-\boldsymbol{\mu}), \qquad (5.30)\end{aligned}$$

which is the same as (5.27).

f. Example: Potato flour dilutions

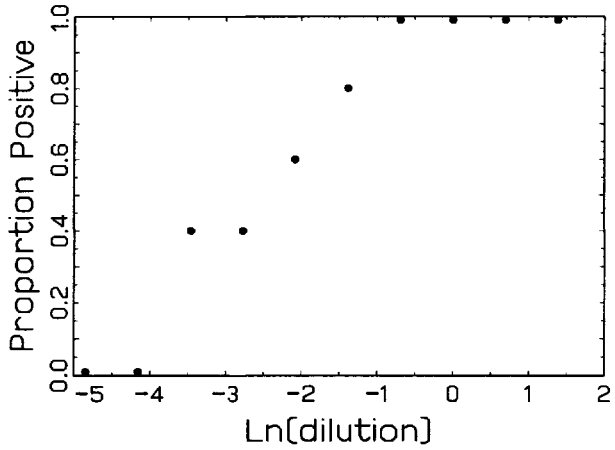
Finney (1971) gives an example of the growth of spores in a potato flour suspension. For each of 10 dilutions, five plates are tested for positive growth. The data are given in Table 5.1. As the flour suspensions get more concentrated, the probability of growth (i.e., proportion of positive plates) increases. Figure 5.1 shows that the probability of response, as a function of the natural logarithm of dilution, follows a roughly sigmoidal shape, so we might entertain a logistic regression model. Let y_i denote the number of plates out of five that show a positive response. A possible model is

Table 5.1: Potato Flour Data

Dilution	Spore Growth		Proportion
(g/100 ml	No. of Plates	No. Positive	of Residual Plates
1/128	5	0	0.0
1/64	5	0	0.0
1/32	5	2	0.4
1/16	5	2	0.4
1/8	5	3	0.6
1/4	5	4	0.8
1/2	5	5	1.0
1	5	5	1.0
2	5	5	1.0
4	5	5	1.0

$$\begin{aligned}\mathrm{E}[y_i] &= 5\pi(x_i) = 5\frac{1}{1+e^{-(\alpha+\beta x_i)}} \qquad (5.31) \\ y_i &\sim \text{indep. binomial } [5, \pi(x_i)].\end{aligned}$$

Figure 5.1: Proportion of positive spore growth plotted against log dilution for the potato flour data.



The log likelihood for this model is given by

$$l = \sum \left[\log \binom{5}{y_i} + y_i(\alpha + \beta x_i) - 5\log(1 + e^{\alpha+\beta x_i}) \right]$$

$$= c + \alpha \sum y_i + \beta \sum y_i x_i - 5 \sum \log(1 + e^{\alpha+\beta x_i}), \qquad (5.32)$$

where $c = \sum \binom{5}{y_i}$ is a function of the y_i but not of α and β. The log likelihood is shown as a function of α and β in Figure 5.2. The ML equations are thus given by

$$\sum y_i = \sum \frac{5}{1 + e^{-(\hat{\alpha}+\hat{\beta}x_i)}}$$

$$\sum y_i x_i = \sum \frac{5x_i}{1 + e^{-(\hat{\alpha}+\hat{\beta}x_i)}}. \qquad (5.33)$$

With $\sum y_i = 31$ and $\sum y_i x_i \doteq -17.329$ it is merely tedious arithmetic to verify that $\hat{\alpha} = 4.17$ and $\hat{\beta} = 1.62$ solve these equations to within rounding error. Figure 5.3 plots the data and fitted values.

To illustrate the large-sample variance calculation note that

$$\tau^2 = 1$$

Figure 5.2: Log likelihood plotted against parameters for the potato flour data.

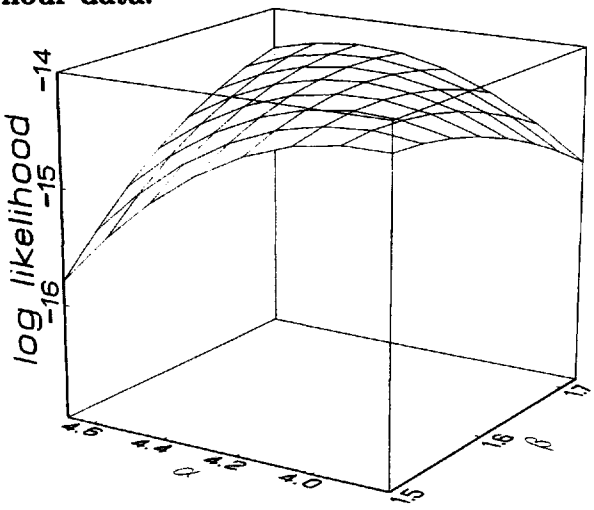
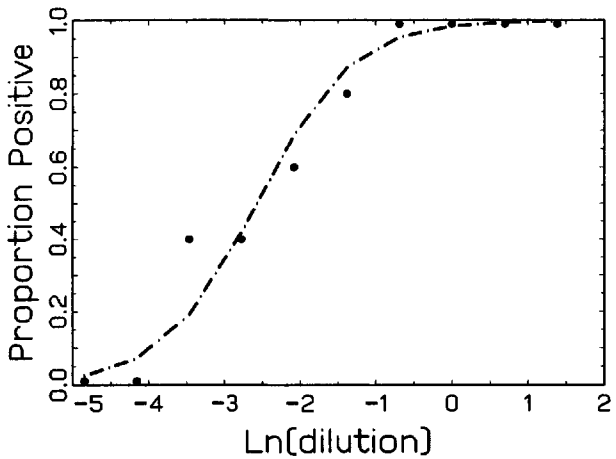


Figure 5.3: Proportion positive versus log dilution for the potato flour data.



$$v(\mu_i) = \mu_i(1-\mu_i)$$

$$g_\mu(\mu_i) = 1/v(\mu_i)$$

so that $\mathbf{W} = \left\{{}_d\mu_i(1-\mu_i)\right\}$. We thus have

$$\mathbf{X'WX} = \begin{bmatrix} \sum \mu_i(1-\mu_i) & \sum x_i\mu_i(1-\mu_i) \\ \sum x_i\mu_i(1-\mu_i) & \sum x_i^2\mu_i(1-\mu_i) \end{bmatrix}$$

$$= \begin{bmatrix} 4.38306 & -11.09943 \\ -11.09943 & 32.89365 \end{bmatrix}$$

with inverse

$$(\mathbf{X'WX})^{-1} = \begin{bmatrix} 1.56805 & 0.52911 \\ 0.52911 & 0.20894 \end{bmatrix}.$$

This gives

$$\begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \end{pmatrix} \sim \mathcal{AN}\left[\begin{pmatrix} \alpha \\ \beta \end{pmatrix}, \begin{pmatrix} 1.56805 & 0.52911 \\ 0.52911 & 0.20894 \end{pmatrix} \right].$$

5.5 TESTS OF HYPOTHESES

a. Likelihood ratio tests

Likelihood ratio tests follow the usual prescription of comparing the maximized values of the log likelihood both under H_0 and not restricted to H_0. If the difference is large (i.e., the unrestricted model fit is much better) then H_0 is rejected.

When there are multiple parameters we will often be interested in hypotheses concerning only a subset of the parameters. Accordingly, let the parameter vector $\boldsymbol{\theta}$ be partitioned into two components $\boldsymbol{\theta}' = (\boldsymbol{\theta}_1', \boldsymbol{\theta}_2')$ and suppose interest focuses on $\boldsymbol{\theta}_1$ while $\boldsymbol{\theta}_2$ is left unspecified. $\boldsymbol{\theta}_2$ is often called a *nuisance parameter*. Either or both of $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ could be vector-valued and, if the entire parameter vector is of interest, $\boldsymbol{\theta}_2$ could be null.

Suppose our hypothesis is of the form $H_0: \boldsymbol{\theta}_1 = \boldsymbol{\theta}_{1,0}$, where $\boldsymbol{\theta}_{1,0}$ is a specified value of $\boldsymbol{\theta}_1$, and let $\hat{\boldsymbol{\theta}}_{2,0}$ be the MLE of $\boldsymbol{\theta}_2$ under the restriction that $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_{1,0}$. The likelihood ratio test statistic is given by

$$-2\log\Lambda = -2\left[l(\boldsymbol{\theta}_{1,0}, \hat{\boldsymbol{\theta}}_{2,0}) - l(\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2)\right], \tag{5.34}$$

where $\hat{\boldsymbol{\theta}}' = (\hat{\boldsymbol{\theta}}_1', \hat{\boldsymbol{\theta}}_2')$ and the large-sample critical region of the test is to reject H_0 in favor of the alternative when

$$-2\log\Lambda > \chi^2_{\nu,1-\alpha}, \tag{5.35}$$

where ν is the dimension of $\boldsymbol{\theta}_1$.

b. Wald tests

An alternative method of testing is to use the large-sample normality of the ML estimator in order to form a test. From standard results (Appendix S)

$$\hat{\boldsymbol{\theta}} \sim \mathcal{AN}[\boldsymbol{\theta}, \mathbf{I}^{-1}(\boldsymbol{\theta})], \tag{5.36}$$

where $\mathbf{I}(\boldsymbol{\theta})$ is the Fisher information for $\hat{\boldsymbol{\theta}}$. Again, if we write $\boldsymbol{\theta}' = (\boldsymbol{\theta}_1', \boldsymbol{\theta}_2')$, and write conformably

$$\mathbf{I}(\boldsymbol{\theta}) = \begin{bmatrix} \mathbf{I}_{11} & \mathbf{I}_{12} \\ \mathbf{I}_{21} & \mathbf{I}_{22} \end{bmatrix} \tag{5.37}$$

then standard matrix algebra for partitioned matrices (Searle, 1982, p. 354) and multivariate normal calculations show that the large-sample variance of $\hat{\boldsymbol{\theta}}_1$ is given by

$$\text{var}_\infty(\hat{\boldsymbol{\theta}}_1) = \left(\mathbf{I}_{11} - \mathbf{I}_{12}\mathbf{I}_{22}^{-1}\mathbf{I}_{21}\right)^{-1}. \tag{5.38}$$

To test $H_0: \boldsymbol{\theta}_1 = \boldsymbol{\theta}_{1,0}$ we form the Wald statistic

$$W = (\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_{1,0})'[\text{var}_\infty(\hat{\boldsymbol{\theta}}_1)]^{-1}(\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_{1,0}), \tag{5.39}$$

which, under H_0, has the same large-sample χ^2 distribution as the LRT with degrees of freedom equal to the dimension of $\boldsymbol{\theta}_1$. More explicitly we would reject the $H_0: \boldsymbol{\theta}_1 = \boldsymbol{\theta}_{1,0}$ if

$$W > \chi^2_{\nu,1-\alpha}. \tag{5.40}$$

Both the LRT and the Wald tests are available to test the same hypotheses and have the same limiting distribution. What are the differences? For large samples, and if the deviation from the null hypothesis is not too extreme, the two test statistics will give similar, though not identical results (Bishop et al., 1975, Sec. 14.9). However, for small samples or for extreme deviations, they can differ. Generally, investigations have shown (Cox and Hinkley, 1974; McCullagh and Nelder,

1989) that use of the large sample-distribution for the LRT gives a more accurate approximation for small and moderate-sized samples than for the Wald test. The LRT is thus to be preferred. The Wald test does, however, have a computational advantage since it does not require calculation of $\hat{\boldsymbol{\theta}}_{2,0}$.

c. Illustration of tests

We use the potato flour data to illustrate these tests for the null hypothesis $H_0: \beta = 0$, i.e., no relationship between spore growth and log dilution. To perform the likelihood ratio test we must maximize the likelihood under the null hypothesis, that is, when the probability of growth is constant. Under $H_0, \hat{\alpha} \doteq 0.4896$ (see E 5.5). We thus have

$$l(\boldsymbol{\theta}_{1,0}, \hat{\boldsymbol{\theta}}_{2,0}) = l(0.50, 0) = -33.20$$

while

$$l(\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2) = l(\hat{\alpha}, \hat{\beta}) = l(4.17, 1.62) = -14.214.$$

The LRT statistic is thus $-2 \log \Lambda = -2[-33.20 - (-14.21)] = 37.88$. The statistic has 1 degree of freedom, which is the dimension of β. So we easily reject H_0 at any usual level of significance and the p-value is $\mathrm{P}\{\chi_1^2 \geq 37.88\} \doteq 0$.

The Wald test statistic uses $\hat{\beta} = 1.62$ from below (5.33) and $\mathrm{var}(\hat{\beta})_\infty = 0.2089$ from the end of Section 5.4. Substituting in (5.39) we then have $W = (1.62)(0.2089)^{-1}(1.62) = 1.62^2/0.2089 = 12.6$. This has a p-value of $\mathrm{P}\{\chi_1^2 \geq 12.6\} \doteq 0.0004$, which again corresponds to rejection of the null hypothesis at the usual significance levels. This illustrates that the two test statistics need not be numerically similar for large deviations from the null hypothesis. Of course, in such situations the same qualitative conclusion would ordinarily be reached.

d. Confidence intervals

Either the LRT or Wald test can be used to construct large-sample confidence intervals for $\boldsymbol{\theta}_1$. For the LRT we include in the confidence set all values $\boldsymbol{\theta}_1$ such that

$$-2\left[l(\boldsymbol{\theta}_1, \hat{\boldsymbol{\theta}}_{2,1}) - l(\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2)\right] \leq \chi^2_{\nu,1-\alpha}. \tag{5.41}$$

In (5.41) $\hat{\boldsymbol{\theta}}_{2,1}$ represents the MLE of $\boldsymbol{\theta}_2$ for each value of $\boldsymbol{\theta}_1$ checked for inclusion in the set.

For the Wald test we include in the confidence set all values of $\boldsymbol{\theta}_1$ such that

$$(\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_1)'[\text{var}_\infty(\hat{\boldsymbol{\theta}}_1)]^{-1}(\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_1) \leq \chi^2_{\nu,1-\alpha}. \tag{5.42}$$

The computational burden of the likelihood-based confidence interval is thus larger than that for the Wald-based interval. However, the small and moderate-sized sample performance of the LRT-based confidence region has generally been found to be better.

e. Illustration of confidence intervals

The likelihood-based confidence interval solves for the values of β such that

$$-2[l(\beta, \hat{\alpha}_\beta) - l(\hat{\alpha}, \hat{\beta})] \leq 3.84,$$

where $\hat{\alpha}_\beta$ denotes the MLE of α when β is fixed at some value. Numerical calculations give the interval as (0.90, 2.76).

The Wald-based confidence interval for β is straightforward since it is based on

$$\hat{\beta} \sim \mathcal{AN}(\beta, 0.2089),$$

which gives a confidence interval of $1.62 \pm 1.96(0.2089)^{1/2} = (0.72, 2.52)$. The LR based interval is approximately the same length as the Wald interval but is not symmetrically placed about the MLE, an indication of the non-normality of the sampling distribution.

5.6 MAXIMUM QUASI–LIKELIHOOD

a. Introduction

In some statistical investigations, such as the potato flour example of Section 5.5, we know the distribution of the data (binomial with $n = 5$ in that instance). In others we are less certain. For example, in analyzing data on costs of hospitalization we know the data are positive (though it would be nice to be paid for some hospital ordeals!) and they are invariably skewed right. With a little more experience with such data we would know that the variance increases with the mean and we might have a rough idea as to how quickly it increases. However, we are unlikely to know *a priori* exactly what distributional form is correct or even likely to fit well. But not knowing the distribution makes it impossible to construct a likelihood and thus to use such techniques as maximum likelihood and likelihood ratio tests.

It would therefore be useful to have inferential methods which work as well or almost as well as ML but without having to make specific distributional assumptions. This is the basic idea behind *quasi-likelihood*: to derive a likelihood-like quantity whose construction requires few assumptions.

What are the important characteristics of likelihood which are required to generate workable estimators? It turns out to be easier to mimic the properties of the derivative of the log likelihood (also called the *score function*) rather than the likelihood itself.

b. Definition

We define an analog of likelihood using (5.9) and (5.10), except that we differentiate with respect to μ_i instead of γ_i. First, from (5.9) we want

$$\mathrm{E}\left[\frac{\partial \log f_{Y_i}(y_i)}{\partial \mu_i}\right] = 0. \tag{5.43}$$

Then we observe that by the chain rule, what we will denote as v^* is

$$\begin{aligned} v^* = \mathrm{var}\left(\frac{\partial \log f_{Y_i}(y_i)}{\partial \mu_i}\right) &= \mathrm{var}\left(\frac{\partial \log f_{Y_i}(y_i)}{\partial \gamma_i}\frac{\partial \gamma_i}{\partial \mu_i}\right) \\ &= \left[\mathrm{var}\left(\frac{\partial \log f_{Y_i}(y_i)}{\partial \gamma_i}\right)\right]\left(\frac{\partial \gamma_i}{\partial \mu_i}\right)^2 \end{aligned} \tag{5.44}$$

and using (5.10)

$$= \left(-E\left[\frac{\partial^2 \log f_{Y_i}(y_i)}{\partial \gamma_i^2}\right]\right)\left(\frac{\partial \gamma_i}{\partial \mu_i}\right)^2.$$

Now, by the nature of $f_{Y_i}(y_i)$ in (5.5), with $b(\gamma_i)$ containing no data this is

$$v^* = \frac{1}{\tau^2}\left[\frac{\partial^2 b(\gamma_i)}{\partial \gamma_i^2}\right]\left(\frac{\partial \gamma_i}{\partial \mu_i}\right)^2, \tag{5.45}$$

and from the definition of $v(\mu_i)$ below (5.14) this becomes

$$\begin{aligned} v^* &= \frac{v(\mu_i)}{\tau^2}\left(\frac{\partial \gamma_i}{\partial \mu_i}\right)^2 \\ &= \frac{v(\mu_i)}{\tau^2}\frac{1}{v(\mu_i)^2} \qquad \text{from (5.15).} \end{aligned} \tag{5.46}$$

Thus

$$\text{var}\left(\frac{\partial \log f_{Y_i}(y_i)}{\partial \mu_i}\right) = \frac{1}{\tau^2 v(\mu_i)}, \tag{5.47}$$

or, by (5.10) and using $\partial\mu_i$ in place of $\partial\gamma_i$,

$$\text{var}\left(\frac{\partial \log f_{Y_i}(y_i)}{\partial \mu_i}\right) = -\text{E}\left[\frac{\partial^2 \log f_{Y_i}(y_i)}{\partial \mu_i^2}\right] = \frac{1}{\tau^2 v(\mu_i)}. \tag{5.48}$$

Observe that (5.43) and (5.48) are the analogs of (5.9) and (5.10).

We thus seek a quantity in place of $\partial \log f_{Y_i}(y_i)/\partial\mu_i$ which has properties (5.43) and (5.48). It is straightforward to verify (see E 5.8) that

$$q_i = \frac{y_i - \mu_i}{\tau^2 v(\mu_i)} \tag{5.49}$$

satisfies these same conditions, where we assume that $\text{var}(y_i) \propto v(\mu_i)$. The τ occurring in (5.49) is merely the (unspecified) constant of proportionality relating $\text{var}(y_i)$ to $v(\mu_i)$, which is not exactly the same as the τ that appears in the density (5.5). However, we will use the same notation since, as we see below, they play the same role.

Since the contribution to the log likelihood from y_i is the integral with respect to μ_i of $\partial \log f_{Y_i}(y_i)/\partial\mu_i$, we define the log quasi–likelihood via the contribution y_i makes to it:

$$Q_i = \int_{y_i}^{\mu_i} \frac{y_i - t}{\tau^2 v(t)} dt, \tag{5.50}$$

which, by definition, has derivative with respect to μ_i equal to q_i. Finally, to find the *maximum quasi–likelihood* (MQL) estimator of $\boldsymbol{\beta}$ we solve the *maximum quasi–likelihood equations*

$$\frac{\partial}{\partial \boldsymbol{\beta}} \sum Q_i = \mathbf{0}. \tag{5.51}$$

Evaluating the derivative in (5.51) gives

$$\sum \frac{y_i - \mu_i}{\tau^2 v(\mu_i)} \frac{\partial \mu_i}{\partial \boldsymbol{\beta}} = \mathbf{0},$$

which, using (5.16), is the same as

$$\sum \frac{y_i - \mu_i}{\tau^2 v(\mu_i) g_\mu(\mu_i)} \mathbf{x}_i' = \mathbf{0}, \tag{5.52}$$

or, in matrix notation,

$$\frac{1}{\tau^2}\mathbf{X}'\mathbf{W}\boldsymbol{\Delta}(\mathbf{y}-\boldsymbol{\mu})=\mathbf{0}, \tag{5.53}$$

the same as (5.18). Note that by defining maximum quasi–likelihood estimators as solutions to the maximum quasi–likelihood equations, (5.51), we avoid a true maximization problem or even the definition of a quasi–likelihood or log quasi–likelihood itself.

In some ways this is a remarkable result. Q_i is constructed using only information about how the variance changes with the mean and nothing more. And, it is often the case that if we specify a mean-to-variance relationship, we obtain maximum quasi–likelihood equations which are exactly the same as those corresponding to a legitimate likelihood.

For example, suppose we are willing to assume the mean and variance are equal, so that what we build into quasi–likelihood is the fact that $v(\mu_i)=\mu_i$. Note that this allows the variance to be merely proportional to the mean rather than exactly equal to it, so that

$$\begin{aligned} Q_i &= \int_{y_i}^{\mu_i}\frac{y_i-t}{\tau^2 t}dt, \\ &= \frac{1}{\tau^2}(y_i\log t - t)\Big|_{y_i}^{\mu_i} \\ &= \frac{1}{\tau^2}(y_i\log\mu_i-\mu_i-y_i\log y_i+y_i) \end{aligned} \tag{5.54}$$

and the MQL equations for $\boldsymbol{\beta}$ are

$$\frac{\partial}{\partial\boldsymbol{\beta}}\sum(y_i\log\mu_i-\mu_i)=\mathbf{0} \tag{5.55}$$

(the other terms dropping out).

Instead of merely assuming that $v(\mu_i)=\mu_i$ suppose we make the assumption that $y_i\sim\text{Poisson}(\mu_i)$, which would actually force $\text{var}(y_i)=\mu_i$ as well. Then $\log f_{Y_i}(y_i)=y_i\log\mu_i-\mu_i-\log(y_i!)$ and the ML equations would be

$$\frac{\partial}{\partial\boldsymbol{\beta}}\sum(y_i\log\mu_i-\mu_i)=\mathbf{0}, \tag{5.56}$$

which are the same as the MQL equations, (5.55)! In this case MQL and ML would give exactly the same estimates and hence MQL would

be fully efficient. In other cases (see E 5.3) ML does not give equations of the form (5.19) and, in those cases, MQL may not be fully efficient. See exercise E 5.10 for some simple calculations and Firth (1987) for more detail.

MQL has important advantages over ML. To explain, consider again the specific situation of regression with a Poisson-distributed response. ML would assume $\text{var}(y_i) = v(\mu_i)$. However, in practice it is often true that data appear selected from a distribution in which the variance is larger than the mean. If the variance is proportional to the mean, the specification of the model under quasi–likelihood is still correct because the assumption is only that $\text{var}(y_i) = \tau^2 v(\mu_i)$; that is, $\text{var}(y_i)$ is proportional to $v(\mu_i)$, not necessarily equal.

Thus MQL affords us two degrees of robustness. First, we need not make a distributional assumption and second, we have only to specify the mean-to-variance relationship up to a proportionality constant which can be estimated from the data (see below).

Inference using MQL proceeds much as ML for β. Under mild conditions (McCullagh, 1983) it can be shown that

$$\tilde{\beta} \sim \mathcal{AN}\left[\beta, \tau^2(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\right], \tag{5.57}$$

with $\tilde{\beta}$ being the MQL estimator of β and, as we defined before, $\mathbf{W} = \left\{{}_d\,[v(\mu_i)g_\mu^2(\mu_i)]^{-1}\right\}$.

However, τ is usually handled differently and estimated via a moment estimator (McCullagh and Nelder, 1989, p. 328):

$$\hat{\tau}^2 = \frac{1}{n-p}\sum \frac{(y_i - \hat{\mu}_i)^2}{v(\mu_i)}, \tag{5.58}$$

where n is the number of observations and p is the dimension of β.

5.7 EXERCISES

E 5.1 Show that $\dfrac{\phi(p_i)^2}{\Phi(p_i)[1-\Phi(p_i)]}$ is the inverse of an estimate of $\text{var}(t_i)$, where t_i is defined in (5.3).

E 5.2 Show that the binomial, Poisson and gamma distributions can be written in the form (5.5). *Hint for the gamma distribution*: Write the density in terms of the mean and coefficient of variation.

E 5.3 Suppose $y \sim \mathcal{N}(e^\theta, e^\theta)$, i.e., y is normal with equal mean and variance. Show that the distribution of y is *not* of the form (5.5).

E 5.4 Show that (5.21) can be written in the form (5.18).

E 5.5 Suppose $y_i \sim$ indep. Binomial(n,p) for $i = 1,2,\ldots,m$, where $p = 1/(1+e^{-\alpha})$. Show that the MLE of α is $\log[\sum y_i/(mn-\sum y_i)]$.

E 5.6 Using (5.24) verify that the large-sample variance of $\hat{\beta}$ is given by (5.25).

E 5.7 Derive (5.27) from (5.26).

E 5.8 Show that q_i of (5.49) satisfies (5.51), (5.52), and (5.53).

E 5.9 For binary (Bernoulli) and Poisson distributed data, in (5.19) show that $\mathbf{W\Delta} = \mathbf{I}$ and hence it simplifies to

$$\mathbf{X'y} = \mathbf{X'}\boldsymbol{\mu}.$$

E 5.10 *Efficiency of MQL*: Suppose that $y_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$ for $i = 1,2,\ldots n$, where $\log \mu_i = x_i\beta$ and $v(\mu_i) = \mu_i$. Calculate the ratio of the large-sample variances of $\tilde{\beta}$, the MQL estimator of β and $\hat{\beta}$, the MLE of β. For concreteness, assume that $n/2$ of the observations have $x_i = 5$ and $n/2$ are 10. Do the calculations for β equal to 0.1, 1, and 10.