# Quasi-Likelihood

So far we have been fitting models using maximum likelihood. This has meant assuming that there is a probability model for the data. This means that we have specified a data generation mechanism, for example that the data consists of counts of events in a Poisson process. In order to propose such a mechanism, we need knowledge of the physical processes that lead to the data, or substantial experience with similar data from previous studies.

The main purpose of many analyses is to show how the mean response is affected by several covariates. Sometimes there is insufficient information about the data for us to specify a model for the data. However, we may be able to specify some of the features of the data. For example,

- whether it is continuous or discrete
- how the mean or median is affected by external stimuli
- how the variability of the response changes with the average response
- whether the observations are independent
- whether the response distribution is skewed.

We can develop analyses based on approximations to the likelihood. We will concentrate on cases where the observations are independent, but extensions can be made to include correlations between the data points.

*The Quasi-likelihood*

Suppose we have a vector of responses, $\boldsymbol{Y}$, which are independent with mean $\mu$ and covariance matrix $\sigma^2$V($\mu$). We assume that $\mu$ is a function of covariates, $\boldsymbol{x}$, and some regression parameters, $\beta$. At the moment we will not need to limit the nature of this relationship, so we absorb the covariates into the regression function by writing $\mu(\beta)$.

Typically, $\sigma^2$ is unknown, and has to be estimated, and V($\mu$) is made up of known functions. As it is assumed that the components of $\boldsymbol{Y}$ are independent, V($\mu$) must be diagonal. Thus we write

$$V(\mu)=\mathrm{diag}\left(V_1(\mu),\ldots,V_n(\mu)\right) \quad .$$

We also need to assume that V$_i$($\mu$) only depends on the $i^{\text{th}}$ component of $\mu$. This seems a reasonable assumption, as it is difficult to see why the variance of an observation would depend on other components, even if the mean does not.

In most applications, the functions $V_1(\cdot)$, ... , $V_n(\cdot)$ may be the same, although their arguments could be different.

To construct the quasi-likelihood, we start by looking at a single component $Y$ of $\boldsymbol{Y}$. Under the conditions listed above, the function

$$U=u(\mu|Y)=\frac{Y-\mu}{\sigma^2\ V(\mu)}$$

has several properties in common with the log-likelihood derivative (i.e. the score). In particular,

$$E(U)=0$$
$$Var(U)=1/\left(\sigma^2 V(\mu)\right)$$
$$-E\left(\frac{\partial U}{\partial \mu}\right)=1/\left(\sigma^2 V(\mu)\right) \quad .$$

Most of the first order first-order asymptotic theory concerned with the likelihood is founded on these properties. It is therefore not surprising that

$$Q(\mu|y)=\int_y^\mu u(y|y)\,dt=\int_y^\mu \frac{y-t}{\sigma^2 V(t)}dt$$

behaves like a log-likelihood function. We refer to this as a quasi-likelihood, or more correctly as a log quasi-likelihood. Since the components of ***Y*** are independent by assumption, the quasi-likelihood for the complete data is the sum of the individual contributions:

$$Q(\mu|\boldsymbol{y})=\sum Q(\mu_i|y_i) \quad .$$

By analogy, the quasi-deviance function for a single observation is

$$D(y|\mu)=-2\ \sigma^2\ Q(\mu|y)=2\int_\mu^y \frac{y-t}{V(t)}dt$$

(note reversal of order of integration!). The total deviance, D(***y*** | $\mu$), is the sum of the individual components, and only depends on ***y*** and $\mu$, but not $\sigma^2$. It should also be noted that the complete quasi-likelihood only depends multiplicatively on $\sigma^2$, so that it does not affect the MLEs of $\mu(\beta)$ (and hence $\beta$).

*Examples*

The simplest example is when the variance function is 1. Then *U* is

$$U=\frac{Y-\mu}{\sigma^2}$$

so that the quasi-likelihood is

$$Q(\mu|y)=-\frac{Y-\mu^2}{2}$$

which is the same as the likelihood for a normal distribution.

Similarly, if our variance function is $\mu$, U is

$$U=\frac{Y-\mu}{\mu\sigma^2}$$

so the quasi-likelihood is

$$Q(\mu|y) = y\log\mu - \mu$$

which is the same as the likelihood for a Poisson distribution.

Other quasi-likelihoods are:

| V($\mu$) | $Q(\mu \mid y)$ | ***Distribution name*** | ***Canonical parameter*** | Range Restrictions |
|---|---|---|---|---|
| 1 | $-(y-\mu)^2/2$ | Normal | $\mu$ | - |
| $\mu$ | $y\log\mu-\mu$ | Poisson | log $\mu$ | $\mu > 0$<br>$y \geq 0$ |
| $\mu^2$ | $-y/\mu-\log\mu$ | Gamma | $-1/\mu$ | $\mu > 0$<br>$y \geq 0$ |
| $\mu^3$ | $-\frac{y}{2\mu^2}+\frac{1}{\mu}$ | Inverse Gaussian | $-1/2\mu^2$ | $\mu > 0$<br>$y \geq 0$ |
| $\mu^\zeta$ | $\mu^{-\zeta}\left(\frac{\mu y}{1-\zeta}-\frac{\mu^2}{2-\zeta}\right)$ | - | $\frac{1}{(1-\zeta)\mu^{\zeta-1}}$ | $\mu > 0$<br>$\zeta \neq 0, 1, 2$ |
| $\mu(1-\mu)$ | $y\log\left(\frac{\mu}{1-\mu}\right)+\log(1-\mu)$ | Binomial/*m* | $\log\left(\frac{\mu}{1-\mu}\right)$ | $0 < \mu < 1$<br>$0 \leq y \leq 1$ |
| $\mu + \mu^2/k$ | $y\log\left(\frac{\mu}{k+\mu}\right)+k\log\left(\frac{k}{k+\mu}\right)$ | Negative Binomial | $\log\left(\frac{k}{k+\mu}\right)$ | $\mu > 0$<br>$y \geq 0$ |

Note that there are now no restrictions on whether the data are discrete or continuous.

*Covariates*

Thus far we have written the models in terms of $\mu$. In practice, we need to make the vector $\mu$ a function of a smaller number of parameters, $\beta$. The obvious approach to this is to use a linear model for a function of $\mu$, indeed the clearest approach (by analogy with GLMs) is to make the canonical parameter a linear function. We can then simply use the same formulae for describing models as we use for GLMs.

That the theory that is developed here does not assume linearity - in principle it could be used for any model for the relationship between $\mu$ and $\beta$, linear or not.

*Estimation*

The quasi-likelihood estimating equations for the parameters $\beta$ are obtained by differentiating the function $Q(\mu|\,y)$, and can be written as

$$\boldsymbol{U}(\beta)=\boldsymbol{D}^{\boldsymbol{T}}\boldsymbol{V}^{-\mathbf{1}}(\boldsymbol{Y}-\mu)/\sigma^2=0 \quad .$$

This is called the quasi-score function. D is a $n \times p$ matrix with elements $\partial\mu_i/\partial\beta_r$ , the derivatives of $\mu(\beta)$ with respect to the parameters.

The covariance matrix of $\boldsymbol{U}(\beta)$ is also the negative expected value of $\partial\boldsymbol{U}(\beta)/\partial\beta$ , and is

$$\boldsymbol{i}_\beta=\boldsymbol{D}^{\boldsymbol{T}}\boldsymbol{V}^{-\mathbf{1}}\,\boldsymbol{D}/\sigma^2 \quad .$$

This matrix plays the same role as the Fisher information for likelihood functions. In particular, the asymptotic covariance matrix of $\hat{\beta}$ is

$$Cov(\hat{\beta})\simeq\boldsymbol{i}_\beta^{-\mathbf{1}}=\sigma^2\,(\boldsymbol{D}^{\boldsymbol{T}}\boldsymbol{V}^{-\mathbf{1}}\,\boldsymbol{D})^{-1} \quad .$$

We still have to estimate $\sigma^2$. There is no equivalent to an ML estimate, so the method of moments estimate, based on the residual vector $\boldsymbol{Y}-\hat{\mu}$ , is used:

$$\tilde{\sigma}^2=\frac{1}{n-p}\sum_i\frac{(Y_i-\mu_i)^2}{V_i(\hat{\mu}_i)}=X^2/(n-p) \quad .$$

*Residuals*

As there is no formal likelihood, deviance residuals are not defined. we could, of course, use contributions to the quasi-deviance as residuals. Of course, raw residuals and Pearson residuals

*Example: leaf blotch on barley*

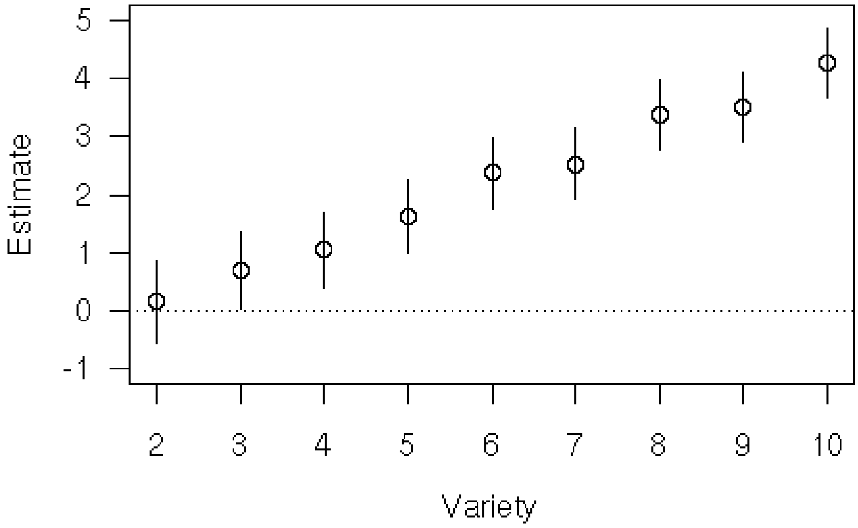
The data below comes from an experiment on leaf blotch (*Rhyncosporium secalis*) on barley. The response is the percentage of the leaf covered by leaf blotch.

| | ***Variety*** | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| ***Site*** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** |
| **1** | 0,05 | 0,00 | 0,00 | 0,10 | 0,25 | 0,05 | 0,50 | 1,30 | 1,50 | 1,50 |
| **2** | 0,00 | 0,05 | 0,05 | 0,30 | 0,75 | 0,30 | 3,00 | 7,50 | 1,00 | 12,70 |
| **3** | 1,25 | 1,25 | 2,50 | 16,60 | 2,50 | 2,50 | 0,00 | 20,00 | 37,50 | 26,25 |
| **4** | 2,50 | 0,50 | 0,01 | 3,00 | 2,50 | 0,01 | 25,00 | 55,00 | 5,00 | 40,00 |
| **5** | 5,50 | 1,00 | 6,00 | 1,10 | 2,50 | 8,00 | 16,50 | 29,50 | 20,00 | 43,50 |
| **6** | 1,00 | 5,00 | 5,00 | 5,00 | 5,00 | 5,00 | 5,00 | 10,00 | 50,00 | 75,00 |
| **7** | 5,00 | 0,10 | 5,00 | 5,00 | 50,00 | 10,00 | 50,00 | 25,00 | 50,00 | 75,00 |
| **8** | 5,00 | 10,00 | 5,00 | 5,00 | 25,00 | 75,00 | 50,00 | 75,00 | 75,00 | 75,00 |
| **9** | 17,50 | 25,00 | 42,50 | 50,00 | 37,50 | 95,00 | 62,50 | 95,00 | 95,00 | 95,00 |

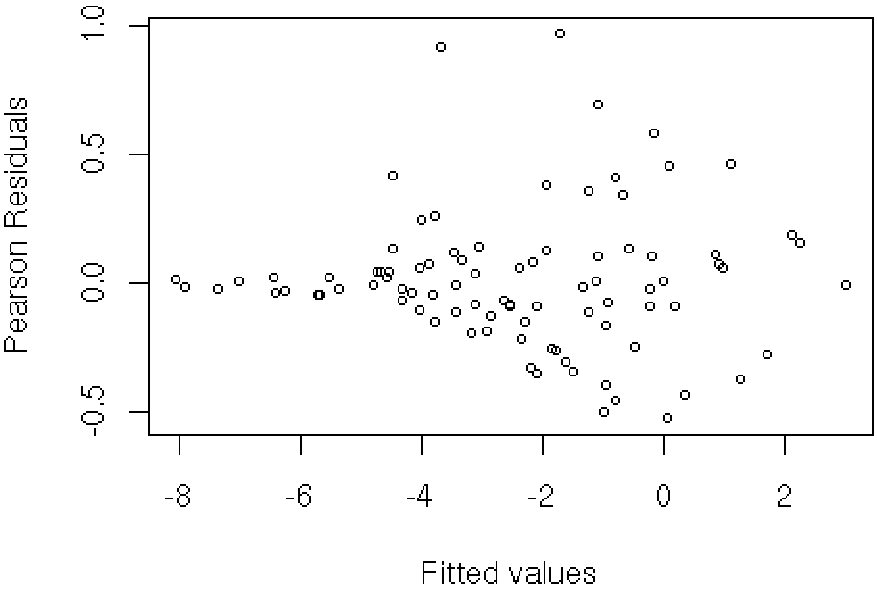
Although there is no formal model for the distribution of percentages, it seems reasonable that there would be a higher variance in the centre of the range, i.e. around 50%. The model might therefore be that the data are pseudo-binomial observations, with a $\sigma^2\mu(1-\mu)$ variance function.

The model can be fitted, and the dispersion parameter is 0.089. The residual deviance is 6.0447 on 72 df. Because the data does not involve counts, there is no reason to expect $\sigma^2$ to be near 1.

When we examine the parameter estimates, as contrasts against Variety=1 and Site =1, we see that they have been ordered by variety.



We can plot the Pearson residuals against the linear predictor, $\hat{\eta} = \log(\hat{\mu}/(1-\hat{\mu}))$ :



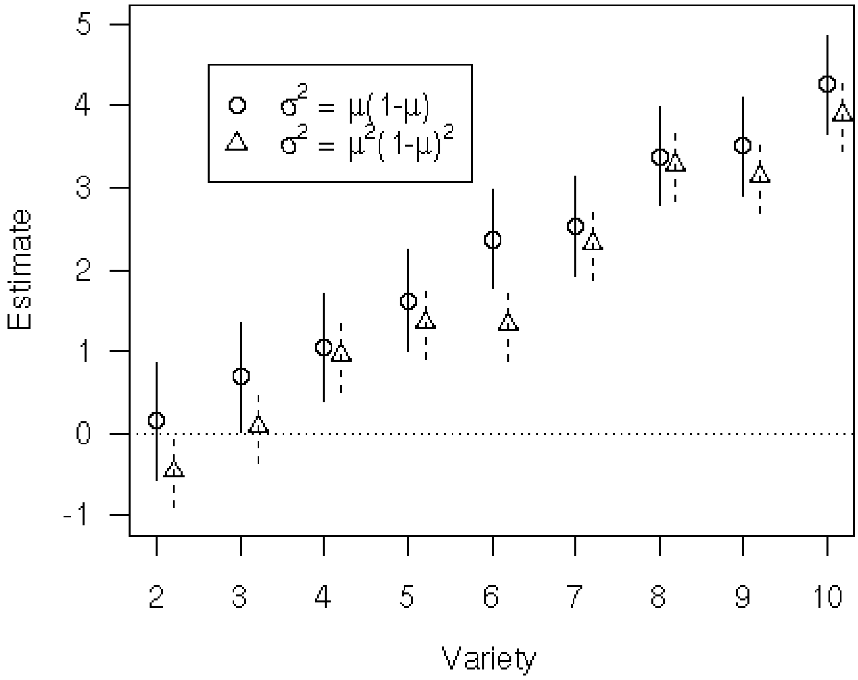
This shows us that there is still heteroscedasticity, with the variance being much larger for intermediate probabilities.

One solution to this problem is to use a different variance function, for example $\mu^2(1-\mu^2)$.

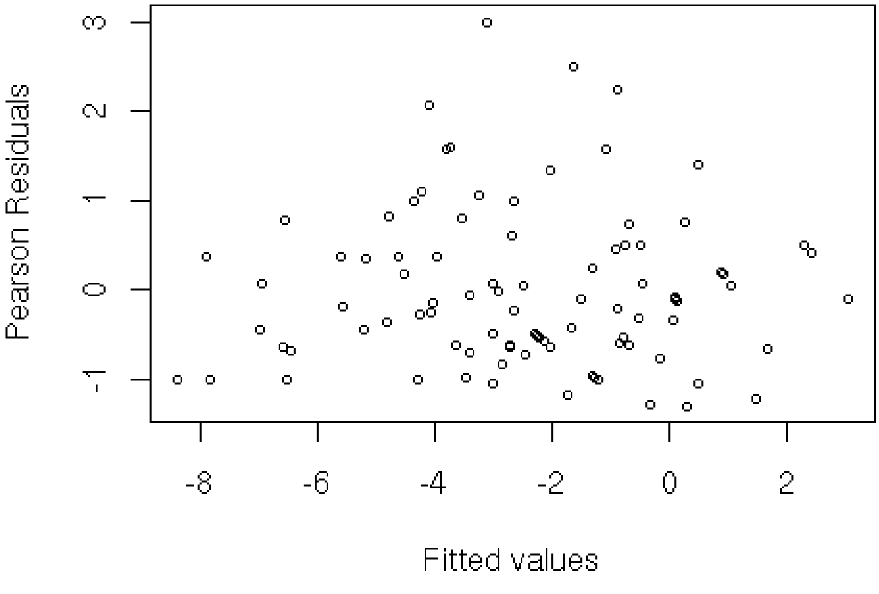
The resulting quasi-likelihood function is

$$Q(\mu|y)=(2y-1)\log\left(\frac{\mu}{1-\mu}\right)-\frac{y}{\mu}-\frac{1-y}{1-\mu} \quad .$$

This is not defined for either $\mu$=0 or $\mu$=1, so the saturated deviance, and hence the residual deviance cannot be calculated. However, this can still be fitted. When this is done, the parameter estimates are similar to those with the binomial variance:



The main difference is that Variety 1 now has a higher estimate. The residuals still show a similar pattern, but much less pronounced:



*Extended Quasi-likelihood: Estimating* $\sigma^2$

In the previous example, the different variance functions were compared graphically. However, they cannot be compared formally, as the properties of the quasi-likelihood that make it comparable to a likelihood refer only to derivatives of $\beta$, and not $\sigma^2$.

The quasi-likelihood can be extended to include terms for the variance. This will allow us to compare different variance functions, and opens up hte possibility of modelling the dispersion as a function of covariates.

For a single obervation, *y*, we want to construct a function $Q^+(\mu, \sigma^2 \mid y)$ that, for known $\sigma^2$, is the same as $Q(\mu \mid y)$, but which also has the properties of a log likelihood with respect to derivatives of $\sigma^2$. Thus we have to have

$$Q^+(\mu, \sigma^2 | y) = Q(\mu | y) + h(\sigma^2 | y) = -\frac{D(y|\mu)}{2\sigma^2} + h(\sigma^2 | y)$$

for some function h($\sigma^2 \mid y$). We will assume that it is of the form

$$h(\sigma^2 | y) = -\frac{1}{2} h_1(\sigma^2) - h_2(y) \quad .$$

If $Q^+$ is to behave like a log likelihood with respect to $\sigma^2$, we must have $E(\partial Q^+/\partial\sigma^2)=0$. Thus

$$0 = \frac{1}{2\,\sigma^4} E\left(D(Y|\mu)\right) - \frac{1}{2} h'_1(\sigma^2) \quad ,$$

implying that

$$\sigma^4 \; h'_1(\sigma^2) = E\left(D(Y|\mu)\right) \quad .$$

To a rough first order approximation we have $E\left(D(Y|\mu)\right) = \sigma^2$, giving $h_1(\sigma^2) = \log(\sigma^2)$ + const. Thus the extended quasi-likelihood function is given by

$$Q^+(\mu, \sigma^2 | y) \simeq -\frac{1}{2} D(y|\mu)/\sigma^2 - \frac{1}{2} \log \sigma^2 \quad .$$

If we have information about the higher order moments, we can improve the approximation. It can be shown that

$$E\left(D(Y|\mu)\right) \simeq \sigma^2 + \frac{1}{12\,V^2}\left(6\sigma^4\, V V'^2 - 3\,\sigma^4\, V^2\, V'' - 4V'\kappa_3\right)$$

where *V* is the variance function, and $\kappa_3$ is the third order cumulant. Members of the exponential family of distributions (and averages from these), have the property

$$\kappa_{r+1} = \kappa'_r \kappa_2 \quad , \text{for r} \geq 2$$

*Cumulants*

Cumulants are constants that, like moments, can be used to describe a probability distribution. Formally,

$$\exp\left(\sum_{r=1}^{\infty} \kappa_r t^r / r!\right) = \sum_{r=0}^{\infty} \mu'_r t^r / r!$$

where $\mu'_r$ is the $r$th moment (about the origin). The first four moments and their cumulants are related like this:

$$\mu'_1 = \kappa_1$$
$$\mu'_2 = \kappa_2 + \kappa_1^2$$
$$\mu'_3 = \kappa_3 + 3\kappa_2 \kappa_1 + \kappa_1^3$$
$$\mu'_4 = \kappa_4 + 4\kappa_3 \kappa_1 + 6\kappa_2 \kappa_1^2 + \kappa_1^4$$

where $\kappa_2 = \sigma^2 V(\mu)$, and the differentiation is w.r.t. $\mu$. From this, we get

$$\begin{aligned} E(D(Y|\mu)) &\simeq \sigma^2 \left(1 + \frac{5\,(\kappa_3^2/\kappa_2^3)^2 - 3\,(\kappa_4/\kappa_2^2)}{12}\right) \\ &= \sigma^2 \left(1 + \frac{\sigma^2\,(2V'^2/V - 3\,V'')}{12}\right) \end{aligned}$$

as well as

$$Var(D) \simeq 2\,\kappa_2^2/V^2 = 2\,\sigma^4$$
$$Cov(D,Y) \simeq (\kappa_3 - \kappa_2\,\kappa'_2)/V \quad .$$

The covariance obviously reduces to 0 under the property of exponential family cumulants above.

If we use the simpler assumption that $\sigma^2$ is sufficiently small that $E(D(Y|\mu)) \simeq \sigma^2$, we find that the derivatives

$$\frac{\partial Q^+}{\partial \mu} = \frac{Y - \mu}{\sigma^2\,V(\mu)} \quad \text{and} \quad \frac{\partial Q^+}{\partial \sigma^2} = \frac{D(Y|\mu)}{2\sigma^4} - \frac{1}{2\sigma^2}$$

have zero mean and approximate covariance matrix

$$\begin{pmatrix} \dfrac{1}{\sigma^2\,V(\mu)} & \dfrac{\kappa_3 - \kappa_2\,\kappa'_2}{2\,\sigma^6\,V^2} \\ \dfrac{\kappa_3 - \kappa_2\,\kappa'_2}{2\,\sigma^6\,V^2} & \dfrac{1}{2\sigma^4} \end{pmatrix}$$

The off diagonal terms are zero under the property above, and even if it does not hold, they are often negligible. The expected value of the second derivative matrix is the same as above, except that the off-diagonal terms are zero. Consequently, $Q^+$ has the properties of a quasi-likelihood with respect to both mean and dispersion parameter.