Quasi-Likelihood Theory in Full

Quasi-likelihood theory is a way of generalizing the likelihood-based approach to generalized linear models (GLMs). In particular, quasi-likelihood theory provides a way of estimating the regression parameters from a GLM (that is, the β) in such a way that we do not need to specify a distribution for our outcome (conditional on covariates). This is often useful when we have data that appears to be *over-dispersed*. We say that data are over-dispersed (relative to some underlying model) when the variance observed in the data is larger than what the model would have predicted. This is often the case in count data, which we are likely to use a Poisson regression model for.

Recall that in the Poisson distribution, we have that $E[Y] = \lambda$ and $\text{var}(Y) = \lambda$. As a result, if we are fitting a Poisson GLM (say with a log-link function) to the data, then we are implicitly assuming that, for these data, the mean and variance are approximately equal. Often times this is not the case in observed data, where there is far *more* variance in the data than is predicted by the model. In this case, it is useful to be able to specify a model which maintains the same general mean structure as-in a Poisson regression, but which is not constrained by the distributional limitation of $E[Y] = \text{var}(Y)$. One method for doing this is through the use of quasi-likelihood.

While quasi-likelihood is typically presented as a mechanism for accounting for over-dispersion, it actually is more broadly useful than that. Anytime that we do not want to make a firm distributional assumption regarding our outcome variable, quasi-likelihood can be applied. In fact, when we use the least-squares framing of ordinary least squares of linear regression, we are actually using a quasi-likelihood estimator!

Likelihood Summary for GLMs

Recall that for **any** Score function, $S(\theta)$ we have that $E[S(\theta)] = 0$. We also know that if we differentiate $S(\theta)$, giving us $S'(\theta)$, then $I(\theta) = -S'(\theta)$ is the information and $\text{var}(S(\theta)) = E[I(\theta)]$. These two properties capture the behaviour of the likelihood function in a similar sense to how (μ, σ^2) capture the behaviour of a random variable $X \sim N(\mu, \sigma^2)$.

When we use maximum likelihood estimation, we end up solving $S(\widehat{\theta}) = 0$, to generate our estimator $\widehat{\theta}$. In a generalized linear model, we are typically concerned with estimating $\widehat{\beta}$, which is connected to our distribution through a link function and the mean (namely, $g(\mu_i) = X_i'\beta$). When we are working with an exponential family distribution, we know that the log-likelihood is given by

$$\ell(\theta) = \frac{Y\theta - b(\theta)}{a(\phi)} + c(y; \phi),$$

and as a result we find that

$$S(\theta) = \frac{Y - b'(\theta)}{a(\phi)}.$$

This is a useful general form, but it is in relation to θ rather than to β. Note that, we can connect statements about β to statements about μ through the link function $g(\mu) = X\beta$. Then, if we can connect statements about μ to statements about θ we would be able to apply

the chain rule to derive $S(\beta)$. Knowing that, in general, we have $E[S(\theta)] = 0$, we can use this property to demonstrate that

$$\begin{aligned} E[S(\theta)] = 0 &\implies 0 = \frac{E[Y] - b'(\theta)}{a(\phi)} \\ &\implies E[Y] = \mu = b'(\theta). \end{aligned}$$

As a result, this gives us that $\mu = b'(\theta)$.

If we want $S(\beta)$ we note that

$$S(\beta) = \frac{\partial}{\partial \beta}\ell(\theta) = \frac{\partial \mu}{\partial \beta} \cdot \frac{\partial \theta}{\partial \mu} \cdot \frac{\partial}{\partial \theta}\ell(\theta).$$

We already have that $\frac{\partial}{\partial \theta}\ell(\theta) = S(\theta)$. Since we know that $\mu = b'(\theta)$, then we can say that

$$\begin{aligned} \frac{\partial \mu}{\partial \theta} &= \frac{\partial}{\partial \theta} b'(\theta) = b''(\theta) \\ \frac{\partial \theta}{\partial \mu} &= \left(\frac{\partial \mu}{\partial \theta}\right)^{-1} = \left[b''(\theta)\right]^{-1}. \end{aligned}$$

Taking these values, and plugging in, we get

$$\begin{aligned} S(\beta) &= \frac{\partial \mu}{\partial \beta} \cdot \frac{\partial \theta}{\partial \mu} \cdot S(\theta) \\ &= \frac{\partial \mu}{\partial \beta} \cdot \frac{1}{b''(\theta)} \cdot \frac{Y - b'(\theta)}{a(\phi)} \\ &= \frac{\partial \mu}{\partial \beta} \cdot \frac{Y - \mu}{a(\phi) b''(\theta)}. \end{aligned}$$

Now, using the property that $E[-S'(\theta)] = \text{var}(S(\theta))$, we can simplify this expression somewhat further.

$$\begin{aligned} -S'(\theta) &= -\frac{\partial}{\partial \theta}\frac{Y - b'(\theta)}{a(\phi)} \\ &= \frac{b''(\theta)}{a(\phi)} \\ \text{var}(S(\theta)) &= \text{var}\left(\frac{Y - b'(\theta)}{a(\phi)}\right) \\ &= \frac{\text{var}(Y)}{a(\phi)^2} \\ \implies \frac{b''(\theta)}{a(\phi)} &= \frac{\text{var}(Y)}{a(\phi)^2} \\ \implies \text{var}(Y) &= a(\phi) b''(\theta). \end{aligned}$$

Thus we can write the score $S(\beta)$ as

$$S(\beta) = \frac{\partial \mu}{\partial \beta}\left[\text{var}(Y)\right]^{-1}\left\{Y - \mu\right\}.$$

Quasi-Likelihood Theory

The major issue with the above derivation is that it started from an assumption that Y follows some particular (exponential family) distribution. While this is a reasonable assumption on occasion, often our best rationale for selecting some distribution to work with is mathematical convenience rather than underlying truth. As a result methods which are *robust* to distributional assumptions are desirable. A method is called robust if it is still admissible or otherwise valid even when a particular assumption is not made. Put simply: we want a way to estimate GLMs without needing to assume particular distributions.

The idea with quasi-likelihood theory is quite simple: the above likelihood derivation only used the properties of the specific distribution to derive the form of $\ell(\theta)$. From there, the remainder of the discussion followed from our knowledge of Score functions, calculus, and simple algebraic manipulation. If we define a function which is not truly a Score function, but which behaves as though it were one, we might be able to derive a similar estimation procedure.

Consider an arbitrary random variable, Y, taken such that $E[Y] = \mu$ and $\text{var}(Y) = \phi V(\mu)$, for some function $V(\cdot)$. If we define the function

$$U(\mu; Y) = \frac{Y - \mu}{\phi V(\mu)},$$

then it is not difficult to show that $E[U(\mu; Y)] = 0$ and $E\left[-U'(\mu; Y)\right] = \text{var}(U(\mu; Y))$. Indeed,

$$\begin{aligned}
E\left[U(\mu; Y)\right] &= E\left[\frac{Y - \mu}{\phi V(\mu)}\right] \\
&= \frac{E[Y] - \mu}{\phi V(\mu)} = 0 \\
U'(\mu; Y) &= \frac{\partial}{\partial \mu}(Y - \mu)(\phi V(\mu))^{-1} \\
&= -(\phi V(\mu))^{-1} - (Y - \mu)(\phi V(\mu))^{-2} V'(\mu) \\
\implies E[-U'(\mu; Y)] &= E\left[(\phi V(\mu))^{-1} + (Y - \mu)(\phi V(\mu))^{-2} V'(\mu)\right] \\
&= (\phi V(\mu))^{-1} + 0 = \frac{1}{\phi V(\mu)}.
\end{aligned}$$

We can also see that, $\text{var}(U(\mu; Y)) = (\phi V(\mu))^{-1}$ since

$$\text{var}\left(\frac{Y - \mu}{\phi V(\mu)}\right) = \frac{\text{var}(Y)}{\phi^2 V(\mu)^2} = \frac{\phi V(\mu)}{\phi^2 V(\mu)^2} = \frac{1}{\phi V(\mu)}.$$

As a result, we find that both $E[U(\mu; Y)] = 0$ and that $E[-U'(\mu; Y)] = \text{var}(U(\mu; Y))$, just as with the Score functions. In this sense, we can think of $U(\mu; Y)$ as a *quasi-Score function*. We say that it's a quasi-Score function since we got to this point without making any distributional assumption on Y, outside of $E[Y] = \mu$ and $\text{var}(Y) = \phi V(\mu)$.

Now, if we integrate the Score function we get back the log-likelihood function. For this reason, if we integrate the quasi-Score function, we define the resulting quantity to be the

quasi-loglikelihood function. In particular, we write

$$Q(\mu; Y) = \int_y^\mu U(u; Y)du = \int_y^\mu \frac{Y-u}{\phi V(u)}du.$$

This gives us the desired relationship of $\frac{\partial}{\partial \mu}Q(\mu; Y) = U(\mu; Y)$ (through the FTC). We could go one step further, and consider the exponential of Q to denote the quasi-likelihood function (but this is not commonly done).

The quasi-likelihood and quasi-Score functions are both with respect to μ. Once again, we are typically interested not in the mean directly, but rather in a vector of regression parameters, β, which are connected (through a linear predictor) to our mean. Once more we take $g(\mu) = X\beta$. Motivated by the discussion of moving from $S(\theta)$ to $S(\beta)$ above, we can take a similar task for the quasi functions. In particular, we can take $U(\beta) = \frac{\partial}{\partial \beta}Q(\mu; Y)$. Applying a similar chain rule argument we find that

$$U(\beta) = \frac{\partial}{\partial \beta}Q(\mu; Y) = \frac{\partial \mu}{\partial \beta} \cdot \frac{\partial}{\partial \mu}Q(\mu; Y) = \frac{\partial \mu}{\partial \beta} \cdot [\phi V(\mu)]^{-1}(Y - \mu).$$

If we solve $U(\widehat{\beta}) = 0$, then the result $\widehat{\beta}$ are considered the quasi-Maximum Likelihood Estimator (QMLE) of β.

This entire discussion so far has centered around a single data point, where we have ignored our sample. In the event that we have an (iid) sample, say indexed by $i = 1, \ldots, n$, we can derive the QMLE for β, by solving

$$U(\beta) = \sum_{i=1}^n \frac{\partial \mu_i(\beta)}{\partial \beta} \left[\phi V\left(\mu_i(\beta)\right)\right]^{-1} \left[Y_i - \mu_i(\beta)\right] = 0.$$

Here we have emphasized the relationship of μ_i on β, through the notation $\mu_i(\beta) = g^{-1}(X_i'\beta)$, for whatever link function we have selected. Note that, ϕ does not actually impact the estimation of β above: we can factor out a term of $\frac{1}{\phi}$, which multiplying through leaves us with the estimating equation

$$U(\beta) = \sum_{i=1}^n \frac{\partial \mu_i(\beta)}{\partial \beta} V\left(\mu_i(\beta)\right)^{-1} \left[Y_i - \mu_i(\beta)\right] = 0.$$

In general, once we have estimated $\widehat{\beta}$, we can compute $\widehat{\phi}$ through a method of moments estimator (the details are fairly unimportant!).

Some Examples

Generally, in the context of generalized linear models, we take models which are informed by the underlying parametric GLM (that is, with a distributional assumption) to derive relevant quasi-likelihood models. The reason is: **whenever a distribution is correctly specified, quasi-likelihood models are equivalent to likelihood methods for GLM estimation**. This should not be surprising if we go back through our derivation: in the case of proper likelihood, we started with the log-likelihood of an exponential family distribution,

derived $S(\theta)$ from it. We then essentially copied this form, and worked in the opposite direction for quasi-likelihood. What this means is that if you use quasi-likelihood estimators, and the data do truly follow an exponential family distribution, then the MLE and QMLE will be equivalent. Taking motivation from this, we consider the following (commonly used) QMLE models.

Type/Distribution	Link	Quasi-Score	Quasi-(log)likelihood
Continuous/Normal	$g(\mu) = \mu$	$\mathbf{X}\left(\frac{Y-\mathbf{X}\beta}{\phi}\right)$	$-\frac{(Y-\mu)^2}{2}$
Count/Poisson	$g(\mu) = \log(\mu)$	$\mathbf{X}\exp(\mathbf{X}\beta)\left(\frac{Y-\mu}{\phi\mu}\right)$	$y\log(\mu) - \mu$
Binary/Binomial	$g(\mu) = \log\left(\frac{\mu}{1-\mu}\right)$	$\mathbf{X}\frac{\exp(-\mathbf{X}\beta)}{(1+\exp(-\mathbf{X}\beta))^2}\left(\frac{Y-\mu}{\phi\mu(1-\mu)}\right)$	$y\log\left(\frac{\mu}{1-\mu}\right) + \log(1-\mu)$

We can essentially view these estimators as maximizing the quasi-likelihood across the sample. This has a particularly useful interpretation for the continuous/normal distribution, where we see that we are optimizing a function proportional to $-(Y-\mu)^2$. Recall that the OLS estimators for regression coefficients in linear regression can be obtained either by solving for the MLE (assuming a normal distribution) *or* by minimizing the MSE, which is $(Y - X\beta)^2 = (Y-\mu)^2$. Minimizing the MSE is equivalent to maximizing the negative of it, and so we can see that the quasi-Score function here really corresponds to the least squares estimator we are familiar with (for continuous variables).

Key Properties/Why do We Care?

We have walked through the process of using quasi-likelihood methods without every *really* justifying their use. The fact that the QMLEs corresponds to MLEs when the distributional assumptions are correct is comforting, but is not enough to say that quasi-likelihood is useful without making a distributional assumption (which was the entire purpose of this discussion!). Fortunately for us, quasi-likelihood estimators *are* useful when distributional assumptions are not made.

Key Property #1: As long as $\mu_i = g^{-1}(X_i\beta)$ is correctly specified for the distribution, the quasi-likelihood estimators will be consistent for β. Note, this is true whether or not we have specified the variance function $V(\mu_i)$ correctly! This means that, without any need for distributional assumptions, and even without getting the variance structure correct, QMLE are consistent estimators for the parameters of interest.

Key Property #2: Even if $V(\mu_i)$ is incorrectly specified, we can estimate the variance of our estimators ($\widehat{\beta}$) in such a way so as to produce valid inference for $\widehat{\beta}$. This means that **with only the correct specification of** μ_i, we have consistent estimators which we can conduct valid inference on.

It is worth taking a moment to underscore how powerful these two properties really are. We have gone from an estimator which requires a full distributional assumption to be valid (GLMs) and moved towards one which we can use confidently only by knowing the mean

model! This is particularly useful as, often times distributional assumptions are the hardest to check and be confident in.

Now, if you are anything like me, upon learning this you will wonder: "why do we even need GLMs then?". That's a fair question! The reason has to do with efficiency: while these estimators will be consistent for the true underlying parameters, consistency is a large sample property (as $n \to \infty$). Moreover, while we can conduct valid inference, the inference is based on the asymptotic distribution of our estimators (again, as $n \to \infty$) and can perform quite poorly in small samples. Even in the event that we have sufficiently large n, it will generally be the case that an estimator which does not make any distributional assumptions will have higher variance than an estimator which makes a **correct** distributional assumption. That is, we say that these quasi-likelihood estimators are *less efficient* than corresponding likelihood estimators.

Coming up in this course we will understand some of the underlying theory that produces these properties in these quasi-likelihood estimators, and we will explore a technique for constructing such estimators in general. Moreover, we will see how these estimators allow us to produce estimates in longitudinal models (without the need for distributional assumptions).