Quasi-likelihood

Patrick Breheny

December 6

Introduction

- By this point in the course, you've certainly noticed that likelihood-based inference follows a general pattern of likelihood $\rightarrow$ log-likelihood $\rightarrow$ score $\rightarrow$ information
- With the exception of likelihood ratio tests, however, for estimation and inference, we only ever use the score and information
- This brings up the question ... do we even need a likelihood? Can we just start with a score?

Definition

- Let's formalize this idea: given data $y_1, \ldots, y_n$, suppose we intend to estimate parameters $\boldsymbol{\theta}$ by solving the equation

$$\sum_i \psi(\boldsymbol{\theta}|y_i) = \mathbf{0},$$

  where $\psi : \mathbb{R}^d \to \mathbb{R}^d$ is a known function
- Presumably this function ψ is the score function of some likelihood, but we are not bothering to find or specify it
- This unspecified likelihood associated with the above estimating equation is known as a *quasi-likelihood*
- The resulting estimate is called an "M-estimate" (because it's kind of like an MLE)

Quasi-likelihood: Advantages

- Why might we choose to take this approach?
- One main reason is simplicity: in many applications such as longitudinal data, spatial statistics, and time series analysis, complex correlation structures are present and specifying a full likelihood is rather complex; specifying only the score is quite a bit easier
- The other reason involves robustness: there may be different models that give rise to the same score function; by focusing only on properties of the score, our results may hold for a wider class of models

Quasi-likelihood: Disadvantages

Obviously, there are also potential disadvantages:

- Our estimates may be less efficient (higher SE for a given sample size)
- Certain likelihood tools may be inaccessible, such as AIC and likelihood ratio tests
- Small-sample inference may be problematic; without an actual probability model, we have to rely on asymptotic approaches (bootstrapping may be useful as well)

Exponential dispersion families

- A common setting in which quasi-likelihood arises is in the context of GLMs
- Recall that for an exponential dispersion family

$$\ell(\theta) \propto \frac{y\theta - \psi(\theta)}{\phi},$$

we have

$$\mathbb{E}(y) = \nabla\psi(\theta) \equiv \mu$$
$$\mathbb{V}(y) = \phi\nabla^2\psi(\theta) \equiv \phi v$$

(note that ψ here on this slide only represents the cumulant generating function, as opposed to what it represents in the rest of the lecture)

GLMs

- If we are in the modeling context where μ_i depends on a set of predictors $\mathbf{x}_i$ through coefficients $\boldsymbol{\beta}$, we have the score function

$$\sum_i \frac{\partial \theta_i}{\partial \boldsymbol{\beta}} \frac{\partial \ell_i}{\partial \theta_i}$$

- Setting this equal to zero, we can rewrite the estimating equation so that it is solely a function of the mean and variance of y:

$$\phi^{-1} \sum_i \frac{\partial \mu_i}{\partial \boldsymbol{\beta}} v_i^{-1} (y_i - \mu_i) = \mathbf{0}$$

Mean-variance modeling

- The appeal of this approach is that we can model

$$\mathbb{E}Y_i = \mu_i(\boldsymbol{\beta})$$
$$\mathbb{V}Y_i = \phi v(\mu_i)$$

  without worrying about the full distribution of Y
- In other words, we can focus on modeling the mean and the only real distributional assumption we make is the mean-variance relationship $v(\mu_i)$

Generalized estimating equations

- These derivations are the same for multivariate outcomes, in which the estimating equations are

$$\sum_i \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}} \mathbf{V}_i^{-1}(\mathbf{y}_i - \boldsymbol{\mu}_i) = \mathbf{0}$$

- In the multivariate context, this idea is known as *generalized estimating equations*, or GEE
- This is a popular approach for analyzing longitudinal data, and you will learn more about how it works in practice when you take Longitudinal Data Analysis

Properties of the "quasi-score"

- Does our usual likelihood theory hold for these quasi-likelihood models?
- Not by our previous arguments; recall that we needed a true likelihood (and some regularity conditions) to establish that $\mathbb{E}\mathbf{u}(\boldsymbol{\beta}^*) = \mathbf{0}$ and $\mathbb{V}\mathbf{u}(\boldsymbol{\beta}^*) = -\mathbb{E}\nabla\mathbf{u}(\boldsymbol{\beta}^*)$
- Let $\psi_i(\boldsymbol{\beta}) = \phi^{-1}(\partial\mu_i/\partial\boldsymbol{\beta})v_i^{-1}(y_i - \mu_i)$, with $\psi(\boldsymbol{\beta}) = \sum_i \psi_i(\boldsymbol{\beta})$; what properties does this "quasi-score" statistic have?

Properties of the "quasi-score" (cont'd)

- As it turns out, $\psi(\boldsymbol{\beta})$ has the same theoretical properties as the usual score:

$$\mathbb{E}\psi(\boldsymbol{\beta}^*) = \mathbf{0}$$
$$\mathbb{V}\psi_i(\boldsymbol{\beta}^*) = -\mathbb{E}\nabla\psi_i(\boldsymbol{\beta}^*)$$

- Thus, we can apply our previous theoretical arguments (again, assuming Lindeberg condition, an interior neighborhood, and a suitably smooth ψ) to obtain the asymptotic distribution

$$(\mathbf{X}^\top\mathbf{W}\mathbf{X})^{1/2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) \xrightarrow{\text{d}} \text{N}(\mathbf{0}, \mathbf{I}),$$

where $\mathbf{W}$ is a diagonal matrix with entries $(\partial\mu_i/\partial\eta_i)^2/(\phi v_i)$

- One can also use a robust/sandwich estimator for the variance, as we have seen previously

Poisson and quasi-Poisson

- To see an example of how this works, let's consider the Poisson distribution
- As you may have seen in other courses, the Poisson distribution is a convenient distribution for modeling counts, but in practice there are usually extra sources of variability such that the relationship $\mathbb{V}Y_i = \mathbb{E}Y_i$ often does not hold in practice
- A simple remedy is a quasi-Poisson model in which $\mathbb{V}Y_i = \phi\mu_i$

Quasi-Poisson: Estimates and standard errors

- Note that ϕ cancels out of the estimating equation – Poisson and quasi-Poisson models give the exact same estimates $\widehat{\beta}$
- The standard errors, however, are different
- The variance-covariance matrix is $(\mathbf{X}^\top \mathbf{W} \mathbf{X})^{-1}$ in both cases, although
  - Poisson: $w_i = \mu_i$
  - Quasi-Poisson: $w_i = \mu_i/\phi$
- The dispersion parameter ϕ can be estimated with

$$\hat{\phi} = \frac{\sum_i (y_i - \mu_i)^2/\mu_i}{n},$$

although typically $n - d$ is used to account for degrees of freedom

Simulation: Setup

- To see how this works, let's simulate some data in which the mean model is correct, but the variance is incorrect
- Specifically, let

$$\begin{aligned} g_i &\sim \text{Exp}(1) \\ \log(\mu_i) &= x_i\beta \\ Y_i|g_i &\sim \text{Pois}(\mu_i g_i) \end{aligned}$$

- Note that the quasi-Poisson model is also wrong here, but at least it has a dispersion parameter ϕ that allows for extra variability beyond what the model can account for

Simulation: Results

Over 1,000 independent replications, for 95% confidence intervals:

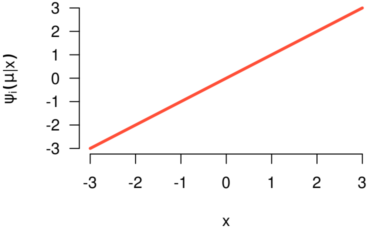
	Coverage	Average SE
Poisson	0.776	0.278
Quasi	0.946	0.451

General quasi-likelihood

- Thus far, we have considered quasi-likelihood exclusively as it pertains to regression models of the mean
- In the time we have left, let's look at quasi-likelihood more generally, without assuming that $\psi(\boldsymbol{\theta})$ can be written in a form involving $y_i - \mu_i$
- To make the discussion a bit more specific, we'll focus on the use of quasi-likelihood as it pertains to robust estimation

That non-robust mean

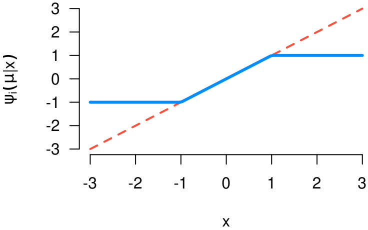
- As you know, the mean is not robust to outlying observations
- One way of visualizing this is to look at it as an M-estimate, with $\psi(\mu|x) = x - \mu$:



the influence that x has over the solution grows without bound as x becomes far from μ

The Huber function

- Consider instead the idea of "capping" the influence of x:



- This quasi-score function was proposed by Peter Huber in 1964

Remarks

- As you would imagine, the resulting M-estimate is much more accurate than the mean when outliers or contamination is present
- So is the median, of course, but one big advantage of the Huber estimate is that unlike the median, it is continuous in the sense that small changes to the data produce small changes in the estimate (unless we're in the capped region)
- The ψ function on the preceding slide would be the score function of a distribution that was normal near the mean, but at some point the tails of the distribution became exponential

Theoretical setup

- With this in mind as a potentially motivating example (there are many, many other examples of robust location estimators and ψ functions for a wide variety of problems), let's consider the theoretical properties of these estimators
- First, some notation:

$$\lambda_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_i \psi(\boldsymbol{\theta}|\mathbf{x}_i)$$

$$\lambda(\boldsymbol{\theta}) = \mathbb{E}\psi(\boldsymbol{\theta}|X)$$

- Note, of course, that $\lambda_n(\boldsymbol{\theta}) \xrightarrow{\text{P}} \lambda(\boldsymbol{\theta})$ for all θ
- Now, let $\boldsymbol{\theta}^*$ be the unique solution to $\lambda(\boldsymbol{\theta}) = \mathbf{0}$

Main result

- **Theorem:** Let $\{x_i\}_{i=1}^n$ be an iid sample, with $\hat{\boldsymbol{\theta}}$ satisfying $\sum_i \psi(\boldsymbol{\theta}|x_i) = \mathbf{0}$. Suppose
  (i) $\psi(\boldsymbol{\theta}|x_i)$ is monotone
  (ii) $\lambda(\boldsymbol{\theta})$ is differentiable at $\boldsymbol{\theta}^*$ and $-\nabla\lambda(\boldsymbol{\theta}^*)$ is positive definite
  (iii) $\mathbb{E}\psi(\boldsymbol{\theta})\psi(\boldsymbol{\theta})^\top$ is finite and continuous in a neighborhood of $\boldsymbol{\theta}^*$

  Then $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \xrightarrow{\text{d}} \text{N}(\mathbf{0}, \mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1})$, where

$$\begin{aligned}\mathbf{A} &= -\nabla\lambda(\boldsymbol{\theta}^*)\\ \mathbf{B} &= \mathbb{E}\psi(\boldsymbol{\theta}^*|X)\psi(\boldsymbol{\theta}^*|X)^\top.\end{aligned}$$

- Condition (i) ensures that solutions to $\lambda(\boldsymbol{\theta}) = \mathbf{0}$ and $\lambda_n(\boldsymbol{\theta}) = \mathbf{0}$ are unique; as with the corresponding MLE theorem, this can be relaxed to a conclusion about the existence of an asymptotically normal solution

Some final thoughts

- Hopefully by this point in the course you feel that you've seen the wide applicability of likelihood, along with many useful extensions, modifications, and applications
- Certainly, there are others we didn't cover, but hopefully you've gained enough experience and familiarity with the tools we have derived and used that you could read and understand how they work on your own