Quasi and pseudo likelihood

Patrick Breheny

December 10, 2025

Introduction

- In our previous lecture, we introduced the idea of carrying out inference without even specifying a full probability model, only the score or estimating equations
- Today, we will see what this looks like in the context of GLMs, where it is referred to as quasi-likelihood
- We also saw that it is possible to obtain correct standard errors via sandwich estimators even if the model is wrong
- This makes it possible to carry out inference for *deliberately incorrect* models (often referred to as pseudo-likelihood)

Advantages

- Why would we do this? Why would we not specify a full model, or deliberately specify an incorrect model?
- The main reason is simplicity: in many applications such as longitudinal data, spatial statistics, and time series analysis, complex correlation structures are present and the full likelihood can be very complicated (both challenging to specify and difficult to optimize)

Disadvantages

Obviously, there are also potential disadvantages:

- Our estimates may be less efficient (higher SE for a given sample size)
- Many likelihood tools will be inaccessible, such as AIC and likelihood ratio tests
- Small-sample inference may be problematic; without an actual probability model, we depend entirely on asymptotic approximations

Exponential dispersion families

- The term *quasi-likelihood* is typically used to refer to the application of the estimating equation idea in the context of GLMs
- Recall that for an exponential dispersion family

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

we have

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

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

$$\boldsymbol{\psi}_i(\boldsymbol{\beta}) = \phi^{-1}(\frac{\partial \mu_i}{\partial \boldsymbol{\beta}})v_i^{-1}(y_i - \mu_i),$$

with $\boldsymbol{\psi}(\boldsymbol{\beta}) = \sum_i \boldsymbol{\psi}_i(\boldsymbol{\beta})$

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

- This "quasi-score" $\boldsymbol{\psi}(\boldsymbol{\beta})$ has the same theoretical properties as the usual score:

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

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

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

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

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 $\widehat{\boldsymbol{\beta}}$
- 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	0.791	1.082
Quasi	0.954	1.760
Sandwich	0.943	1.713

Note that:

- All three models have the same estimating equations, and produce the same $\widehat{\beta}$
- Since Poisson and Quasipoisson have the same estimating equations, their sandwich versions are identical

Response-biased sampling

- Changing topics, let's consider response-biased sampling: instead of a simple random sample, observations are sampled conditional on the outcome, with the case-control study being the most common
- In such situations, the prospective likelihood (the one based on the simple random sample) is usually straightforward and easy to work with, but isn't the actual likelihood based on the study design . . . is it OK to use it anyway?
- This idea of replacing the true likelihood with a simpler likelihood is known as *pseudo-likelihood*

Binomial example: Setup

- Let's start with the simplest case: $Y_i \overset{\text{iid}}{\sim} \text{Bern}(\pi)$ for $i = 1, \ldots, N$
- However, we do not get to observe all N observations; instead, if $Y_i = 1$, the observation is sampled with (known) probability p_1, while if $Y_i = 0$, it is sampled with (known) probability p_0
- Introducing some extra notation, let N_1 and N_0 denote the unobserved number of events, with n_1 and n_0 the observed number of cases and controls in our sample

Binomial example (cont'd)

- As a concrete example, let's suppose $\pi = 0.2$, $p_1 = 1$, and $p_0 = 1/2$ (we get to see all the cases, but only half of the controls)
- In this scenario, if $N = 100$, we would expect to see $n_1 = 20$ cases and $n_0 = 40$ controls; the naïve estimate $n_1/(n_1 + n_0)$ would produce the biased estimate $\hat{\pi} = 0.333$
- Clearly, we must make adjustments for the sampling frequencies p_1 and p_0

Likelihood?

- Let's say we attempted to carry out a likelihood-based analysis of this problem with

$$\begin{aligned} L_i &= \mathbb{P}(Y_i \cap S_i) \\ &= \begin{cases} \pi p_1 & \text{if } Y_i = 1 \\ (1-\pi)p_0 & \text{if } Y_i = 0 \end{cases} \end{aligned}$$

  where S_i denotes the event that the observation was sampled
- Unfortunately, this produces the "MLE" of $\hat{\pi} = n_1/(n_1 + n_0)$, exactly what we said we didn't want
- What went wrong?

Correct likelihood

- This likelihood is incorrect, as we have ignored the unsampled data
- The correct likelihood is $\mathbb{P}(Y_i \cap S_i | S_i)$, the probability of Y_i *conditional* on the fact that the observation made it into the sample
- With this likelihood, the score is now

$$u(\pi) = \frac{n_1}{\pi} - \frac{n_0}{1-\pi} - \frac{(n_0 + n_1)(p_1 - p_0)}{\pi p_1 + (1-\pi)p_0}$$

- The good news is that this score is now "correct", in that the MLE is now sensibly adjusted for sampling fraction:

$$\hat{\pi} = \frac{n_1 p_0}{n_1 p_0 + n_0 p_1}$$

Remarks

- The bad news is that the likelihood is far more complicated and difficult to work with
- In this simplest of scenarios, it is still possible to work through the algebra, but messy enough that I chose to skip it during class time
- One can imagine that this approach is not going to scale up particularly well with more complex probability models

An "estimated" likelihood

- Perhaps there's a simpler way
- In terms of N_1 and N_0, the likelihood for π is simply that of a binomial distribution
- Unfortunately, N_1 and N_0 are unobserved; however, they can easily be *estimated*: $\hat{N}_j = n_j / p_j$
- Thus, perhaps a reasonable way to proceed is to simply plug in these estimates into the binomial likelihood

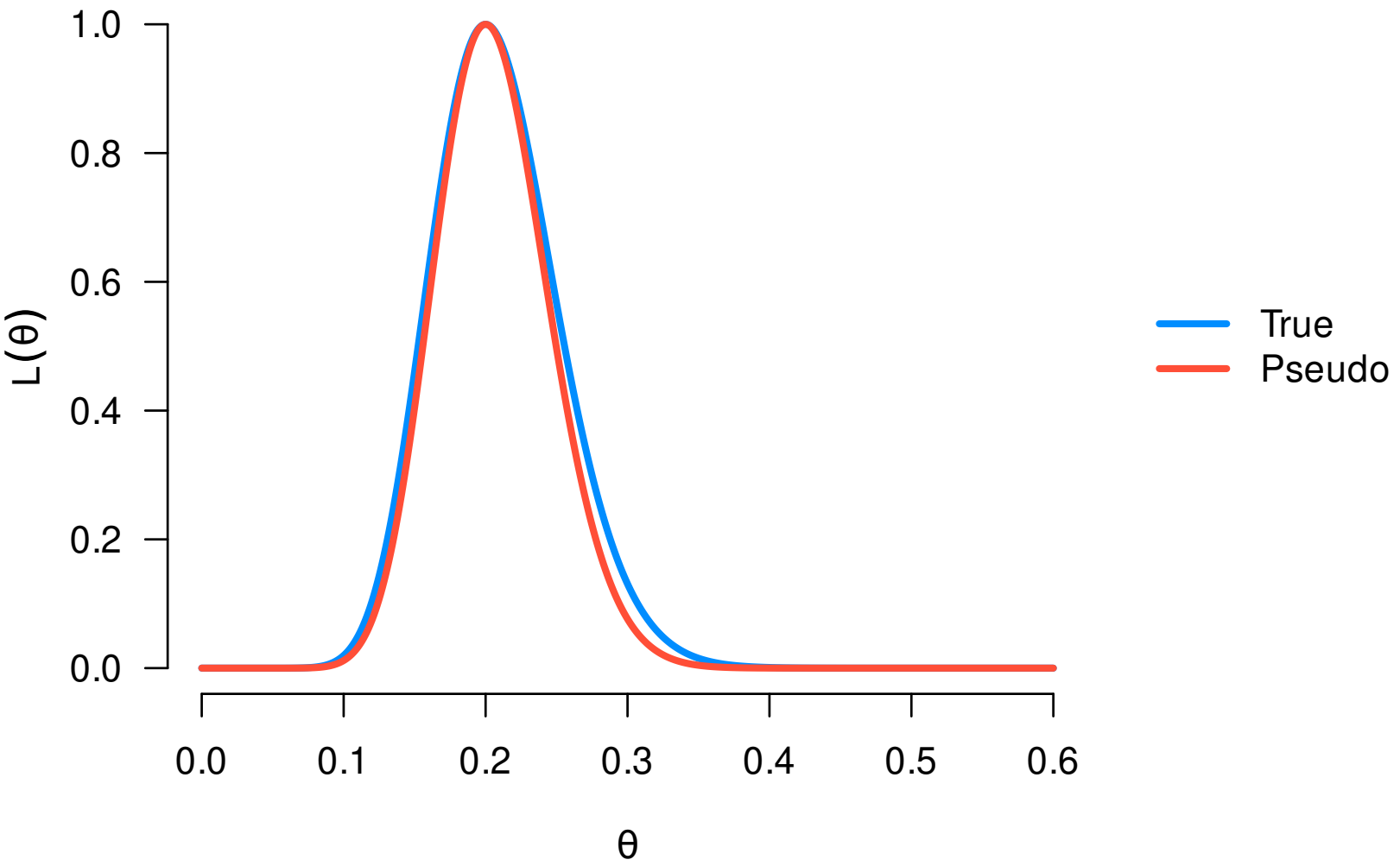
Inverse probability weighting

- Doing so, we obtain the log-likelihood

$$\ell(\pi) = \frac{n_1}{p_1} \log \pi + \frac{n_0}{p_0} \log(1 - \pi)$$

- Note that this is the original, "naïve" likelihood, but where the observations have been weighted by $1/p_1$ and $1/p_0$
- This idea, known as inverse probability weighting, comes up often in statistics, in a variety of contexts

Connection with true likelihood



Remarks

- As the figure illustrates, the pseudo-likelihood is roughly similar to the true likelihood, and the pseudo-MLE is the same as the true MLE
- However, the likelihoods are not the same – in particular, the pseudo-likelihood is narrower
- Treating the pseudo-likelihood as an ordinary likelihood, therefore, is going to produce variance estimates that are too small

Variance estimation

- This is exactly the kind of thing that one would use a sandwich estimator for:

$$\sqrt{n}(\hat{\pi} - \pi^*) \xrightarrow{\mathrm{d}} \mathrm{N}(0, B^{-1} M B^{-1}),$$

  where $B = -\mathbb{E}\nabla^2 \ell_i(\pi^*)$ is the pseudo-information and $M = \mathbb{V}u_i(\pi^*)$ is the variance of the pseudo-score
- These approaches yield the following 95% Wald CIs for π:
  - True likelihood: [0.114, 0.286]
  - Pseudo-likelihood (no adjustment): [0.122, 0.278]
  - Pseudo-likelihood (corrected): [0.114, 0.286]

Case-control studies

- Response-biased sampling arises is in the application of logistic regression to case-control studies
- In this experimental design, a fixed number of cases (n_1) and controls (n_0) are sampled
- The disease status, therefore, is not random; rather it is the exposure(s) that are random
- The true likelihood, therefore, is

$$L = \prod_i p(\mathbf{x}_i | y_i)$$

A pseudo-likelihood

- This is an inconvenient likelihood for several reasons; perhaps most importantly, it requires us to specify a (multivariate) distribution on the predictors, something that is not required in regression approaches
- Suppose we instead treat the data as prospectively acquired, with the likelihood

$$L = \prod_i p(y_i|\mathbf{x}_i);$$

this is obviously much more convenient, as this is just the usual likelihood from a logistic regression model

Comparing the two likelihoods

- This is a pseudo-likelihood in the sense that it does not correspond to the actual likelihood from the experiment
- However, in the special case of logistic regression — this is *not* true for other models — the score of the pseudo-likelihood is equal to the score of the true likelihood
- In this special case, no adjustment is needed to either the estimators or the standard errors, even though we're not actually using the true likelihood (only the estimate of the intercept is affected)

Derivation

- Specifically, letting $s = 1$ denote the event that a subject was sampled, if we assume

$$\frac{\mathbb{P}(s = 1 \,|\, \mathbf{x}, y = 1)}{\mathbb{P}(s = 1 \,|\, \mathbf{x}, y = 0)} = c,$$

  then

$$\mathbb{P}(y = 1 \,|\, \mathbf{x}, s = 1) = \frac{e^{\tilde{\eta}}}{1 + e^{\tilde{\eta}}},$$

  where $\tilde{\eta} = \eta + \log c$
- Note that requiring the case-control sampling ratio to be unaffected by covariates is not a trivial assumption, and is a common source of bias in retrospective studies

Composite likelihood

- Another type of pseudo-likelihood arises from multiplying together separate small components of the likelihood; this is known as *composite likelihood*:

$$L_{\text{cl}}(\boldsymbol{\theta}|\mathbf{y}) = \prod_{k=1}^{K} L_k(\boldsymbol{\theta}|\mathbf{y})$$

- Typically, this is done when the components are simple to derive but the full likelihood is very complicated

One-dimensional lattice

- For example, suppose we have ordered observations $y_1, y_2, \ldots, y_n$ (perhaps ordered with respect to time, or along a genome)
- We might specify a model for how each observation depends on its neighbors: $p(y_k|y_{k-1}, y_{k+1})$
- Multiplying these probabilities together, however

$$p(y_2|y_1, y_3) \times p(y_3|y_2, y_4) \ldots$$

does not actually result in the correct likelihood:

$$p(y_2) \times p(y_3|y_2) \times p(y_4|y_2, y_3) \ldots$$

Ising model

- For example, suppose $y_k \in \{0, 1\}$ and let $n_k = y_{k-1} + y_{k+1}$
- One way to model the dependence of a point on its neighbors is with the *Ising model*

$$p(y_2, \ldots, y_{n-1} | y_1, y_n) = \exp\left\{ \alpha \sum_{k=2}^{n-1} y_k + \beta \sum_{k=2}^{n-1} y_k n_k - h(\alpha, \beta) \right\},$$

  where positive values of β reflect positive dependence (1s and 0s tend to cluster together)
- This true likelihood is intractable, however, since the normalizing constant $h(\alpha, \beta)$ is very complicated

Ising model with composite likelihood

- The composite likelihood, however, is quite convenient:

$$p(y_k|y_{k-1}, y_{k+1}) = \frac{\exp(\alpha + \beta n_k)}{1 + \exp(\alpha + \beta n_k)},$$

  in other words, simple logistic regression
- The parameters α and β are then estimated by maximizing

$$\ell_{\mathsf{cl}}(\alpha, \beta) = \sum_k \ell_k(\alpha, \beta|y_k);$$

  derivatives, Hessians, etc., are straightforward
- The same idea can be extended to higher dimensions as well as continuous outcomes

Example

A simple example in one dimension:

```
# Generate data (in this case, pure noise)
y_all <- rbinom(200, size = 1, prob = 0.5)

# Set up as composite likelihood
idx = 2:(length(y_all) - 1)
y <- y_all[idx]
n <- y_all[idx - 1] + y_all[idx + 1]

# Fit
fit <- glm(y ~ n, family = binomial)
```

Standard errors

- Since the likelihood is misspecified, the ordinary standard error is incorrect and we must use the sandwich estimator $\mathbf{B}^{-1}\mathbf{M}\mathbf{B}^{-1}$
- This is a case where estimating $\mathbf{M}$ is not trivial due to lack of independence, but a variety of alternative estimators exist:

```
coeftest(fit)
#             Estimate Std. Error z value Pr(>|z|)
# (Intercept) -0.51582    0.24720 -2.0866  0.03692
# n            0.47031    0.20690  2.2732  0.02302
coeftest(fit, vcov = vcovHAC)
#             Estimate Std. Error z value Pr(>|z|)
# (Intercept) -0.51582    0.30418 -1.6958  0.08993
# n            0.47031    0.27526  1.7086  0.08753
```

Remarks

- Composite likelihood methods have found many uses in analyzing longitudinal, time series, genetic, and spatio-temporal data
- They are also used in network analysis, where it is (relatively) easy to model how an individual depends on their neighbors, but hard to specify the full likelihood of an entire network
- The idea of taking a valid likelihood for an individual observation but then combining these likelihoods in a way that is *not* the full likelihood also appears in a variant called *partial likelihood*, which is used extensively in survival analysis

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