STAT 830
Likelihood Methods of Inference

0.1 Extensions of Likelihood Methods

There are many objects which include the word 'likelihood' in their name. Here is a partial list

- Likelihood: the basic notion we have already discussed.
- Conditional Likelihood
- Marginal Likelihood (warning: at least two distinctly different definitions are in use)
- Partial Likelihood
- Quasi Likelihood
- Empirical Likelihood
- Pseudo Likelihood
- Composite Likelihood

Many, even most, of these notions were invented to help deal with high dimensional parameter spaces. The typical situation is that we have a high dimensional (or even infinite dimensional) parameter vector θ but only some components of θ are of direct interest to the data aanaalyst.

For instance, in many research problems the regression coefficients are the focus and parameters affecting variability (such as the error variance and even the error distribution) are not. So we might write $\theta = (\beta, \sigma)$ and declare that σ is a nuisance parameter. Or we might really only care about one entry in β, like the one for the treatment effect.

The low-dimensional (usually) *parameter of interest* is distinguished from all the other *nuisance parameters*.

The concepts listed above often either seek to eliminate the nusiance parameters from the inference problem or to provide accurate approximations to the distribution of the estimate of the parameter of interest.

Example: Consider the one-way layout. We have K samples with sample sizes $n_1, \ldots, n_K$. The data are then Y_{ij} which are modelled as independent, and often normally distributed, with mean μ_i and common standard deviation σ. Let $T_i = \sum_j (Y_{ij} - \bar{Y}_i)^2$ be ESS in sample i. Then

$$S = (S_1, S_2) = (\bar{Y}_1, \ldots, \bar{Y}_K, T_1, \ldots, T_K)$$

where S_1 is the vector means, S_2 vector of Ts is aa The complete data is S together with the vector of standardized residuals

$$V = \left[\frac{Y_{11} - \bar{Y}_1}{s_1}, \cdots \frac{Y_{K,n_K} - \bar{Y}_K}{s_k} \right]$$

where each $s_i = \sqrt{T_i/(n_i - 2)}$. It is a fact thaat V is independent of the sufficient statistic S.

The joint density of S then factors aa aa product of a conditional and a marginal density:

$$f(s_1 \mid s_2, \mu, \sigma) f(s_2 \mid \sigma).$$

(I omit V from consideration because V is *ancillary* – its distribution does not depend on the parameters.) The second term in this factorization is called a *marginal likelihood*: it is the joint density of a part of the data.

ASIDE: the term *marginal likelihood* is also used for the object

$$\int f(x \mid \phi, \psi)\, d\psi$$

which eliminates ψ by integrating over a (uniform) prior. This strategy is certainly different from a frequency perspective.

In our example the key point is that there is no μs in the second term. In a homework problem I asked you to use this liklihood to estimate σ; you should see that somehow this corrects the problem with bias which arises from the full likelihood.

The first term in our factorizaation is called aa 'conditional likelihood' but it is not very useful here since it depends on all the parameters. In general, we can try to split the data X into X_1, X_2 and factor

$$f(x \mid \theta) = f(x_1 \mid x_2, \theta) f(x_2 \mid \theta).$$

Then the first tem is a conditional likelihood aand the second term a marginal likelihood. The factorization can be useful if one term or the other depends only on parameter of interest.

I should note that very occasionally $\theta = (\phi, \psi)$ and one term depends only on ϕ and the other only on ψ. In this case the Fisher Information matrix for the full likelihood is block diagonal and the estimates $\hat{\phi}$ and $\hat{\psi}$ depend only on the term they appear in. They are asymptotically independent.

0.1.1 Theoretical Ideas

Now suppose that $\theta = (\phi, \psi)$ and

$$f(x \mid \theta) = f(x_1 \mid \phi) f(x_2 \mid \phi, \psi)$$

Then $f(x_1 \mid \phi)$ is a likelihood with score

$$U_m(\phi) = \nabla \log(f(x_1 \mid \phi))$$

and Hessian which I denote H_m; subscript m for marginal. Usual likelihood theory will often apply to this smaller data set. But notice that typically the Fisher Information in the partial data set is *smaller* than in the complete data set.

Example: Markov Processes and Time Series. In one common data type values of a response Y are measured over time. This is often true for economic data such as the unemployment rate, the Gross Domestic Product, and so on.

Here is one simple example. We assume that Y_t is recorded at times $t = 0, 1, \ldots, T$. The most elementary model is an autoregression of order 1, an AR(1) process, which satisfies the model equation

$$Y_{t+1} - \mu = \rho(Y_t - \mu) + \epsilon_{t+1}$$

with $\epsilon_1, \ldots, \epsilon_T$ iid $\mathrm{N}(0, \sigma^2)$. This specification is supposed to mean the ϵ_{t+1} is independent of all previous Y values, that is, of $Y_t, Y_{t-1}, \ldots$. It specifies the conditional distribution of Y_{t+1} given Y_t. It does not specify the distribution of Y_0.

We can now factor the joint density of $Y_0, \ldots, Y_T$ in the form

$$f(Y_T \mid Y_{T-1}, \mu, \rho, \sigma) \cdots f(Y_1 \mid Y_0, \mu, \rho, \sigma) f(Y_0)$$

The last term here is complex in principle and may depend on other parameters. But

$$f(y_1, \ldots, y_T \mid y_0, \mu, \rho, \sigma) = \prod_1^T f(y_t \mid y_{t-1}, \mu, \rho, \sigma)$$

is a *conditional likelihood*.

Each conditional density is just a normal density so we have so

$$\ell_c(\mu, \rho, \sigma) = -\sum_{t=1}^{T} \frac{\{Y_t - \mu - \rho(Y_{t-1} - \mu)\}^2}{2\sigma^2} - T\log(\sigma).$$

This conditional log-likelihood can be maximized to find MLEs for μ, σ, and ρ. I want you to notice the similarity to the likelihood for simple linear regression. In particular, defining $\beta = \rho$ and $\alpha = \mu(1-\rho)$ we see that

$$\ell_c(\alpha, \beta, \sigma) = -\sum_{t=1}^{T} \frac{(Y_t - \alpha - \beta Y_{t-1})^2}{2\sigma^2} - T\log(\sigma).$$

This function is easily maximized analytically and it is a fact that standard likelihood properties hold for this conditional log-likelihood. In particular we can deduce that $\hat{\rho}$ is approximately normally distributed with a variance which depends on ρ and σ.

0.1.2 Quasi-likelihood

The idea of quasi-likelihood is really to take the likelihood equations for an exponential family regression model and use analogues without assuming you have correctly specified the distribution of the data.

Consider an exponential family regression model. In such a modelt there is a response Y_i which has density

$$f(y_i \mid \phi_i) = h(y_i) exp\left\{y_i\phi_i - b(\phi_i)\right\}$$

where $g(\phi_i) = \alpha + \beta x_i$ and x_i is a covariate. Poisson, Gamma, logistic, exponential and Gaussian regression all fit this framework.

Note that

$$\frac{\partial g(\phi_i)}{\partial \theta} = g'(\phi)\frac{\partial \phi_i}{\partial \theta}$$

by the change rule. As a result we find that

$$\frac{\partial \phi_i}{\partial \theta} = \frac{1}{g'(\phi_i)} \begin{bmatrix} 1 \\ x_i \end{bmatrix}$$

This allows us to see that if we have independent observations $Y_1, \ldots, Y_n$ with non-random covariate vectors $x_1, \ldots, x_n$ then the score function is

$$U(\alpha, \beta) = \begin{bmatrix} \sum_i w_i (y_i - b'(\phi_i)) \\ \sum_i w_i x_i (y_i - b'(\phi_i)) \end{bmatrix}$$

where

$$w_i = \frac{1}{g'(\phi_i)}.$$

Notice that in an exponential family $b'(\phi_i)$ must be the mean $\mu_i = \mu(\alpha + x_i\beta)$. The idea now is to consider replacing $b'(\phi)$ by some generic function $\mu(\phi)$ and replacing the weights w_i by a generic function $w(\phi)$. Then

$$U(\alpha, \beta) = \begin{bmatrix} \sum_i w_i(\phi_i)(y_i - \mu(\phi_i)) \\ \sum_i w_i x_i (y_i - \mu(\phi_i)) \end{bmatrix}$$

is called a *quasi-score*. And a function ℓ whose gradient is U is called a *quasi-likelihood* even if it is not the log of a density.

I warn you that not every quasi-score has a corresponding quasi-likelihood. If $h : \mathbb{R}^p \mapsto \mathbb{R}^p$ then in order for h to be the gradient of some function $g : \mathbb{R}^p \mapsto \mathbb{R}$ the $p \times p$ derivative matrix Dh of h must be the same as the Hessian of g. But the Hessian is necessarily symmetric if g is twice continuously differentiable so h must have a symmetric derivative matrix if it is to be a gradient.

0.1.3 The Cox Proportional Hazards Model

We now turn to the framework of clinical trials in which we follow each of a group of patients through time. For patient i we record the time T_i at which some well-defined thing called an *endpoint* event. This might be the time to recurrence of a disease, the time for a blood marker like T4 cell count to drop to a specified level or the time to death of the patient.

Let f_i be density of T_i and define the *hazard* function of T_i by

$$\begin{aligned} h_i(t) &= \lim_{\epsilon \to 0} \frac{P(t \le T_i \le t + \epsilon \mid T_i \ge t)}{\epsilon} \\ &= \lim_{\epsilon \to 0} \frac{\int_t^{t+\epsilon} f_i(u)\, du}{\epsilon \{1 - F(t)\}} \\ &= \frac{f_i(t)}{1 - F(t)}. \end{aligned}$$

In the proportional hazards model we assume that there is some *baseline* hazard $h_o(t)$ such that each individual has hazard h_i proportional to h_0:

$$h_i(t) = h_0(t) \exp(x_i \beta)$$

Some reality checks

This framework is too narrow for real clinical trials because a number of wrinkles arise. The first crucial problem is that in real medical trials some patients disappear before they experience the endpoint. We say they are *censored* or *lost to follow-up*. We do assume that you record C_i, the *censoring* time, for each such case.

Another potential wrinkle is that in real trials you don't get to start all the patients off simultaneoulsly at $t = 0$. Instead patients are recruited gradually over a period which is often several years. For each patient you must choose how to measure time. Sometimes time is measured from recruitment of that patient into the trial. Sometimes is is measured from date of treatment. Sometimes it might be some fixed patient age. It would be rare to measure time from the beginning of the study.

Finally another potential wrinkle is that the covariate X_i might depend on t: age, blood pressure, and income might all depend on time. It will turn out that Cox's partial likelihood can handle all these potential problems.

Parametric form example

I am going to do an exponential distribution example in the very simple case where there is no censoring and where the covariates are time independent and non-random. I will assume $h_0(t) = \alpha$, independent of t.

Each T_i has constant hazard $\alpha \exp(x_i \beta)$. So I will now consider a T with constant hazard, say α. In this case Thus

$$h(t) = -\frac{d}{dt} \log \{1 - F(t)\} = \alpha.$$

We have integrate this formula from 0 to t to see that

$$\log \{1 - F(t)\} - \log \{1 - F(0)\} = -\alpha t.$$

So the survival function of T is

$$1 - F(t) = \exp \{-\alpha t\}$$

which means that T has an exponential distribution with rate parameter α and mean $\mu = 1/\alpha$. Finally we see each T_i has an exponential distribution with rate

$$\alpha \exp(x_i\beta).$$

In what follows I will find rhe log-likelihood in two ways. First, the joint densty of $T_1, \ldots, T_n$ at $t_1, \ldots, t_n$ is

$$\prod_{i=1}^{n} \alpha\gamma_i \exp(-t_i\alpha\gamma_i)$$

where

$$\gamma_i = \exp(x_i\beta).$$

We can certainly do standard likelihood analysis on this likelihood but I want to rewrite it in a form which will help deal with censoring. To do so let $T_{(1)} < T_{(2)} < \cdots < T_{(n)}$ be the sorted *failure* times. Then let $J_1, \ldots, J_n$ be the indexes of the variables which fail at these times. So at time $T_{(1)}$ patient J_1 fails.

Here are some facts about expondential random variables. Suppose $T_i; i = 1, \ldots, n$ are independent exponential random variables and that T_i has rate β_i. Then $T_{(1)}$ is the minimum of $T_1, \ldots, T_n$. Fix t and compute

$$P(T > t) = P(T_1 > t, \ldots, T_n > t)$$

because in the smallest observation is more than t if and only if every observation is more than t. But then

$$\begin{aligned} P(T > t) &= \prod_{i=1}^{n} P(T_i > t) \\ &= \prod_{i=1}^{n} \exp(-\beta_i t) \\ &= \exp\left\{-(\beta_1 + \cdots + \beta_n)t\right\} \end{aligned}$$

This proves that T has an exponential distribution with rate which is the sum of all the rates in the minimum. That rate in our problem is $\alpha\lambda_{(1)}$ where

$$\lambda_{(1)} = \sum_i \gamma_i = \sum_i \exp(x_i\beta).$$

Fact:

$$P(J_1 = j \mid T_{(i)} = t_{(1)}) = \frac{\gamma_j}{\sum_i \gamma_i} = \frac{\gamma_j}{\lambda_{(1)}}.$$

Lack of Memory Property

If T is exponential with rate τ then conditional law of $T - t$ given $T \geq t$ is exponential with rate τ. So given that $T_{(1)} = t_{(1)}$ and $J_1 = j_1$ the time $T_{(2)} - T_{(1)}$ to the next failure has an exponential distribution. Conditional rate is $\alpha\lambda_{(2)} = \alpha\lambda_{(1)} - \gamma_{j_1}$. Then given $T_{(1)} = t_{(1)}$, $T_{(2)} = t_{(2)}$ and $J_1 = j_1$ we have

$$P(J_2 = j_2 \mid T_{(1)} = t_{(1)}, T_{(2)} = t_{(2)}, J_1 = j_1) = \frac{\gamma_{j_2}}{\lambda_{(2)}}.$$

Continuing this way we may write the likelihood in the (symbolic) form

$$P(T_{(1)} = t_{(1)})P(J_1 = j_1 \mid T_{(1)} = t_{(1)})P(T_{(2)} = t_{(2)} \mid T_{(1)} = t_{(1)}, J_1 = j_1) \cdots$$

The even numbered terms do not involve α, the baseline hazard. We form the *partial* likelihood by multiplying together these terms.

The Partial Likelihood

The log-partial likelihood depends on the sequence $J_1, J_2, \ldots, J_n$ but *NOT* on the actual times:

$$\ell_P = \log\left\{P(J_i = j_i \mid \text{History})\right\} = \sum_{i=1}^{n} \left\{\log(\gamma_{j_i}) - \log(\lambda_{(i)})\right\}.$$

which simplifies to

$$\ell_P = \sum_{i=1}^{n} \left\{x_{j_i}\beta - \log(\lambda_{(i)}\right\}.$$

In general: observe sequence of failure and censoring times. Think about time of ith failure. Instantly before time of ith failure, there is a set R_i of those subjects who have not yet failed or been censored. Set R_i is called the *risk set*. At this time individual j_i fails.

The log-partial likelihood is

$$\sum_i \left[x_{j_i}\beta - \log\left\{\sum_{j \in R_i} \exp(\gamma_i)\right\}\right]$$

General Structure

Think of data at $V_1, W_1, V_2, \ldots, V_n, W_n, V_{n+1}$ Write the likelihood in form

$$\begin{aligned}\ell &= P(V_{n+1} \mid W_n, \ldots, V_1)P(W_n \mid V_n, \ldots, V_1) \\ &\times P(V_n \mid W_{n-1} \ldots, V_1)P(W_{n-1} \mid V_{n-1} \ldots, V_1) \\ &\times \cdots \times P(W_1 \mid V_1)P(V_1)\end{aligned}$$

Keep all terms of form $P(W_i \mid \text{History})$ and call this L_P. Then L_P has many properties of a likelihood.

Theoretical concepts

We can analyze the partial likelihood

$$\ell_P(\beta) = \sum_i \left\{ x_{J_i}\beta - \log(\lambda_{(J_i)} \right\}.$$

The corresponding *partial score* is

$$\sum_i \left\{ x_{J_i} - \frac{\sum_{j \in R_i} x_j \gamma_j}{\sum_{j \in R_i} \gamma_j} \right\}$$

Define

$$D_k = x_{J_k} - \frac{\sum_{j \in R_k} x_j \gamma_j}{\sum_{j \in R_k} \gamma_j}$$

and

$$M_k = \sum_{i=1}^{k} D_i$$

Let $\mathcal{H}_{k-1}$ be the *history* up to the instant before the kth failure. This history contains information on which subjects have failed or been censored up to this failure. In particular this history contains R_k.

Given R_k the probability that subject $j \in R_k$ fails in the next ϵ time units is

$$h_j(t)\epsilon + o(\epsilon)$$

and the probability that some subject in R_k fails in this time interval is

$$\sum_{i \in R_k} h_i(t)\epsilon + o(\epsilon).$$

So given R_k and time $T_{(k)}$ of kth failure the conditional probability that $j \in R_k$ is the subject who fails is

$$\frac{h_j(t)\epsilon + o(\epsilon)}{\sum_{i \in R_k} h_i(t)\epsilon + o(\epsilon)}$$

Let $\epsilon \to 0$, to get the kth contribution to the partial likelihood and then compute

$$\mathrm{E}(D_k \mid \mathcal{H}_{k-1}) = \sum_{j \in R_k} P(J_k = j \mid \mathcal{H}_{k-1}) \left\{ x_j - \frac{\sum_{j \in R_k} x_j \gamma_j}{\sum_{j \in R_k} \gamma_j} \right\} = 0.$$

Notice that this is the usual unbiasedness property of scores!

The Score is a Martingale

Let N be the number of failures observed (which is typically random and smaller than the sample size because of censoring). The partial score is then M_N. It is a fact that $M_1, \ldots, M_N$ is a martingale which means

$$\mathrm{E}(M_k \mid \mathcal{H}_{k-1}) = M_{k-1}.$$

Here is how we prove that:

$$\begin{aligned} \mathrm{E}(M_k \mid \mathcal{H}_k) &= \mathrm{E}(\sum_{i=1}^{k} D_i \mid \mathcal{H}_k) \\ &= \sum_{i=1}^{k-1} D_i + \mathrm{E}D_k \mid \mathcal{H}_k) \\ &= M_{k-1} + 0 = M_{k-1}. \end{aligned}$$

Notice that for $i < k$ the value of D_i is part of the history $\mathcal{H}_{k-1}$. That is why it just comes out of the conditional expectation as you move to the second line.

Definition: A sequence $M_0, M_1, M_2, \ldots$ is a martingale relative to a sequence $\mathcal{H}_k$ (of σ fields, technically) if

$$\mathcal{H}_k \subset \mathcal{H}_{k+1}$$

and

$$\mathrm{E}(M_{k+1} \mid \mathcal{H}_k) = M_k.$$

There are many important theorems concerning martingales: optional sampling, martingale convergence, and martingale central limit theorems. Let $D_k = M_k - M_{k-1}$ be the *martingale differences*. Notice that

$$\begin{aligned}\mathrm{Var}(M_{k+1}) &= \mathrm{Var}\left\{\mathrm{E}(M_{k+1} \mid \mathcal{H}_k)\right\} + \mathrm{E}\left\{\mathrm{Var}(M_{k+1} \mid \mathcal{H}_k\right\} \\ &= \mathrm{Var}(M_k) + \mathrm{E}\left\{\mathrm{Var}(D_{k+1} \mid \mathcal{H}_k)\right\}\end{aligned}$$

So as with sums of independent observations the variances of the differences add up. This can be turned into a central limit theorem for M_N. For details you could look at the book by Peter Hall and Chris Heyde.

Theorem 1 *Suppose that for each N there is a martingale $M_{k,n}, k = 0, \ldots, N$ with associated* filtration

$$\mathcal{H}_{k,n}, k = 0, \ldots, N-1$$

Suppose that

$$\sum_{k=0}^{N-1} \mathrm{Var}(D_{k+1} \mid \mathcal{H}_k) \to 1$$

and that for every $\epsilon > 0$:

$$\sum_{k=0}^{N-1} \mathrm{E}\left\{D_{k+1}^2 1(|D_{k+1}| > \epsilon) \mid \mathcal{H}_k\right\} \to 1.$$

Then

$$M_{N,N} \text{ converges in distribution to } N(0,1).$$

The last condition is the martingale version of *Lindeberg's condition*.