Counting and Locating Multiple Solutions of Estimating Equations

Speaker: Donald Richards (Penn State University)

This talk is based on joint work with:

Despina Stasi (Penn State University)

Elizabeth Gross (NC State University)

Sonja Petrović (Illinois Institute of Technology)

Logistic regression

θ_i: The probability that individual i in a random sample of n individuals will develop a particular characteristic during a follow-up period.

Y_i: Bernoulli random variable which indicates whether or not individual i develops the characteristic.

$Y_1, \ldots, Y_n$ are assumed independent, so they have joint p.d.f.

$$f(y_1, \ldots, y_n; \theta_1, \ldots, \theta_n) = \prod_{i=1}^{n} \theta_i^{y_i}(1-\theta_i)^{1-y_i}, \quad y_i = 0 \text{ or } 1$$

List the individuals so that the first m are those who have the characteristic; so, $y_i = 1$, $i \leq m$, and $y_i = 0$, $i > m$.

Likelihood function:

$$L(\theta_1, \ldots, \theta_n) = \prod_{i=1}^{m} \theta_i \cdot \prod_{i=m+1}^{n} (1 - \theta_i)$$

Predictor variables: $x_1, x_2, \ldots, x_k$ (and $x_0 \equiv 1$)

Data: x_{ij}, the observed value of x_j for the ith individual.

$\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_k)$: A vector of unknown parameters to be estimated by the method of maximum likelihood.

Model θ_i through a *logistic relationship*:

$$\theta_i = \frac{1}{1 + e^{-\sum_{j=0}^{k} \beta_j x_{ij}}}$$

The likelihood function:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{m} \frac{1}{1 + e^{-\sum_{j=0}^{k} \beta_j x_{ij}}} \cdot \prod_{i=m+1}^{n} \frac{1}{1 + e^{\sum_{j=0}^{k} \beta_j x_{ij}}}$$

The derivatives of $\log L(\boldsymbol{\beta})$ w.r.t. β_r, $r = 0, \ldots, k$:

$$\frac{\partial}{\partial \beta_r} \log L(\boldsymbol{\beta}) = \sum_{i=1}^{m} x_{ir} \frac{e^{-\sum_{j=0}^{k} \beta_j x_{ij}}}{1 + e^{-\sum_{j=0}^{k} \beta_j x_{ij}}} - \sum_{i=m+1}^{n} x_{ir} \frac{e^{\sum_{j=0}^{k} \beta_j x_{ij}}}{1 + e^{\sum_{j=0}^{k} \beta_j x_{ij}}}$$

The system of $k+1$ likelihood equations:

$$\sum_{i=1}^{m} \frac{1}{1+e^{\sum_{j=0}^{k} \beta_j x_{ij}}} \begin{pmatrix} x_{i0} \\ x_{i1} \\ \vdots \\ x_{ik} \end{pmatrix} = \sum_{i=m+1}^{n} \frac{e^{\sum_{j=0}^{k} \beta_j x_{ij}}}{1+e^{\sum_{j=0}^{k} \beta_j x_{ij}}} \begin{pmatrix} x_{i0} \\ x_{i1} \\ \vdots \\ x_{ik} \end{pmatrix}$$

Change of variables:

$$\gamma_j \equiv e^{\beta_j} \;, \quad j = 0, \ldots, k$$

The likelihood equations: For $\gamma_0, \ldots, \gamma_k > 0$,

$$\sum_{i=1}^{m} \frac{1}{1+\gamma_0^{x_{i0}} \cdots \gamma_k^{x_{ik}}} \begin{pmatrix} x_{i0} \\ x_{i1} \\ \vdots \\ x_{ik} \end{pmatrix} = \sum_{i=m+1}^{n} \frac{\gamma_0^{x_{i0}} \cdots \gamma_k^{x_{ik}}}{1+\gamma_0^{x_{i0}} \cdots \gamma_k^{x_{ik}}} \begin{pmatrix} x_{i0} \\ x_{i1} \\ \vdots \\ x_{ik} \end{pmatrix}$$

Problems:

1. Count the number of solutions of this system of equations?
2. Can we calculate *all* solutions?

The Donner party data

Row 1: Age

Row 2: Sex (1=male, 0=female)

Survived vs. Died

40	40	28	22	23	28	15	20	18	25	20	32	32	24	30
0	1	1	0	0	1	0	0	1	1	1	1	0	0	1
21	46	32	23	25	23	30	28	40	45	62	65	45	25	28
0	1	0	1	0	1	1	1	1	0	1	1	0	0	1
23	47	57	25	60	15	50	25	30	25	25	25	30	35	24
1	0	1	1	1	1	0	1	1	1	1	1	1	1	1

Suppose we were given the data on individuals 8, 10, 29, and 43 only, then the system of likelihood equations is:

$$\begin{pmatrix} 1 & 1 & 1 & 1 \\ 20 & 25 & 25 & 30 \\ 0 & 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} a \\ b \\ c \\ d \end{pmatrix} = 0,$$

where $\gamma_0, \gamma_1, \gamma_2 > 0$ and

$$a = \frac{1}{1+\gamma_0\gamma_1^{20}\gamma_2^{0}}\,, \qquad b = \frac{1}{1+\gamma_0\gamma_1^{25}\gamma_2^{1}}\,,$$

$$c = -\frac{\gamma_0\gamma_1^{25}\gamma_2^{0}}{1+\gamma_0\gamma_1^{25}\gamma_2^{0}}\,, \quad d = -\frac{\gamma_0\gamma_1^{30}\gamma_2^{1}}{1+\gamma_0\gamma_1^{30}\gamma_2^{1}}\,.$$

Row-reduction leads to: $a = -b = -c = d$, so $ab < 0$, $cd < 0$.

Conclusion: The likelihood equations have no real solutions.

Suppose we were given the data on individuals 2, 20, 24, and 29 only. Then the likelihood equations are

$$\begin{pmatrix} 1 & 1 & 1 & 1 \\ 40 & 25 & 40 & 25 \\ 1 & 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} a \\ b \\ c \\ d \end{pmatrix} = 0$$

where $\gamma_0, \gamma_1, \gamma_2 > 0$ and

$$a = \frac{1}{1 + \gamma_0 \gamma_1^{40} \gamma_2^{1}}\,, \qquad b = \frac{1}{1 + \gamma_0 \gamma_1^{25} \gamma_2^{0}}\,,$$

$$c = -\frac{\gamma_0 \gamma_1^{40} \gamma_2^{1}}{1 + \gamma_0 \gamma_1^{40} \gamma_2^{1}}\,, \quad d = -\frac{\gamma_0 \gamma_1^{25} \gamma_2^{0}}{1 + \gamma_0 \gamma_1^{25} \gamma_2^{0}}\,.$$

Row-reduction leads to two equations in four variables:

$a + c = 0$ and $b + d = 0$

There are *infinitely many real solutions* to this system:

$$\gamma_0 = \gamma_1^{-25}, \quad \gamma_2 = \gamma_1^{-15}, \quad \gamma_1 > 0$$

This is not surprising, for we were given uninformative data:

$$\begin{array}{cccc} 40 & 25 & 40 & 25 \\ 1 & 0 & 1 & 0 \end{array}$$

A rigorous estimation method should not be able to provide unique estimates from such data.

Is it possible to maximize $L(\gamma_1^{-25}, \gamma_1, \gamma_1^{-15})$ w.r.t. γ_1 and describe the root surface corresponding to each γ_1?

If we were given the data on individuals 16-20 and 31-35 only, then the likelihood equations are

$$\begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 21 & 46 & 32 & 23 & 25 & 23 & 47 & 57 & 25 & 60 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} a_1 \\ \vdots \\ a_{10} \end{pmatrix} = 0$$

where

$$a_1 = \frac{1}{1 + \gamma_0 \gamma_1^{21} \gamma_2^{0}}, \ldots, \; a_{10} = -\frac{\gamma_0 \gamma_1^{60} \gamma_2^{1}}{1 + \gamma_0 \gamma_1^{60} \gamma_2^{1}}$$

Load the data into *Macaulay2*, a software package for numerical algebraic geometry

Let a laptop computer run for hours

Macaulay2 finds *all* 1,346 complex solutions

Only 3 of the 1,346 solutions are real

Only 1 of the 3 real solutions has all components positive:

$$(87982.8, 0.751485, 0.0197566)$$

Conclusion: $(87982.8, 0.751485, 0.0197566)$ is the unique MLE.

Macaulay2 has therefore proved that the MLE exists and is unique.

The General Case

Suppose that the x_{ij} are integers (e.g., the Donner data) or rational numbers.

The ML equations reduce to a system of polynomial equations.

The Fundamental Theorem of Algebra: Every non-zero, one-variable polynomial of degree n, with complex coefficients, has exactly n complex roots (counted with multiplicity).

Rothe (1608), Euler (1749), Lagrange (1772), Laplace (1795), Gauss (1799), Argand (1806), Ostrowski (1920), . . .

How does the Fundamental Theorem of Algebra generalize to several variables?

1841: F. Minding generalizes the FTA to two variables.

1975: D. Bernstein generalizes the FTA to arbitrary number of variables.

Bernstein's proof motivated numerical algorithms for sweeping through the values of the polynomial system to find *all* complex *isolated* roots.

Polynomial Homotopy Continuation algorithms

J. Verschelde, Univ. Illinois at Chicago: Extensive PHC website with software, examples, manuals, free downloads.

Garcia-Puente, Gross, Kahle, Petrović, Stasi, Sommese: People who know how to apply the software

Buot and Richards (2006). Counting and locating the solutions of polynomial systems of maximum likelihood equations, I. *J. Symbolic Computation*

Buot, Hoşten, and Richards (2007). Counting and locating . . ., II: The Behrens-Fisher problem. *Statistica Sinica*

Cox, Little, and O'Shea (1998). *Using Algebraic Geometry*, Springer

Gross, Drton, and Petrović (2012). The maximum likelihood degree of variance component models. *Electron. J. Statist.*

Sturmfels (1998). Polynomial equations and convex polytopes. *Amer. Math. Monthly*

As $n \to \infty$, the number of roots of ML equations does not always converge to 1

Problem: Estimate the *correlation* matrix of a multivariate normal distribution

Social scientists wish to estimate tetrachoric and polychoric correlations.

Constrained estimation problems; more difficult than estimating the covariance matrix.

This problem cannot be solved by estimating each bivariate correlation separately.

We must parametrize the set of correlation matrices carefully.

$N_3(0, R)$, a trivariate normal distribution with mean 0 and correlation matrix R

Collect a random sample and write down the likelihood function.

We solve the likelihood equations using *Bertini*, a software package for numerical algebraic geometry.

The likelihood equations seem to *always* have 35 complex solutions.

The number of statistically relevant solutions varies from 5 to 9.

Even with $n = 10^7$, we found cases with 9 statistically relevant solutions.

Conclusions

Statisticians often have complicated estimating equations with:

- Small sample sizes
- Large numbers of parameters
- Multiple roots

We recommend the use of *numerical algebraic geometry*

- 21st-century mathematical methods
- Powerful algorithms for solving estimating equations
- These algorithms compute *all* solutions of the equations