

Monte Carlo Conditional Analysis for Log-Linear and Logistic Models

Brian S. Caffo

Department of Biostatistics

Johns Hopkins University School of Public Health

bcaffo@jhsph.edu

References

- Caffo and Booth (1999). *An MCMC Algorithm for Approximating Exact Conditional Probabilities*. JCGS.
- Caffo and Booth (To Appear). *Monte Carlo and Markov Chain Monte Carlo Inference for Log-linear and Logistic Models*. SMMR.
- Booth and Butler (1998). *An importance sampling algorithm for exact conditional tests in log-linear models*. Biometrika.
- <http://www.biostat.jhsph.edu/~bcaffo/>

Fisher's exact test for 2×2 tables

Treatment	Successes	Failures	
A	X	$N_A - X$	$\text{Bin}(p_A, N_A)$
B	Y	$N_B - Y$	$\text{Bin}(p_B, N_B)$

- $\text{logit}(p_A) = \beta$

- $\text{logit}(p_B) = \beta + \lambda$

λ is the *parameter of interest*

β is the *nuisance parameter*

$$H_o : \lambda = 0 \text{ vs } H_a : \lambda \neq 0$$

- h test statistic with observed value h_{obs}

$$P(h \geq h_{obs}; \beta)$$

General Frequentist Solutions to Nuisance Parameters

- Plug in $\hat{\beta}$
- For many statistics the large sample behavior of h does not depend on β
- Choose h to be an exact pivot for β (for example t test or F test)
- Unconditional tests

$$\sup_{\beta \in \mathbb{R}} P(h \geq h_{obs}; \beta) \geq P(h \geq h_{obs}; \beta)$$

- Eliminate β by conditioning on its sufficient statistic
- Approximate conditioning

Benefits of Conditional Inference

- Eliminates β
- Preserve the nominal type I error rate unconditionally
- Many standard tests are conditional tests
- Many conditional tests are UMPU
- Induce correlation
- Avoid Neyman/Scott

Criticisms of Conditional Inference

- In the cases we consider today, the sufficient statistic for β is not ancillary for λ .

For example, in the two sample binomial example $X + Y$ is not ancillary for λ

- For categorical data
 - The conditional distributions can be very discrete or even degenerate (logistic regression)
 - Can lead to overly conservative tests
 - Often hard/impossible to do because of the complexity or size of the support of the conditional distribution

Conditional analysis for log-linear models

- $\mathbf{y} = (y_1, \dots, y_n)^t \sim \text{Poisson}(\boldsymbol{\mu})$
- Alternative model: $\log \boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\lambda}$
- Null model: $\log \boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\lambda}_0$
- Sufficient statistics for $\boldsymbol{\beta}$ under H_0 are
 $\mathbf{s}_{\boldsymbol{\beta}} = \mathbf{X}^t \mathbf{y}$
- We can eliminate $\boldsymbol{\beta}$ from null distribution by conditioning on $\mathbf{s}_{\boldsymbol{\beta}}$

This setting includes conditional logistic regression and multinomial log-linear models as special cases

Conditional Distribution Under H_0

$$f(\mathbf{y}; \boldsymbol{\beta}) = \frac{1}{\prod_{i=1}^n y_i!} \exp(\boldsymbol{\beta}' \mathbf{s}_\beta + \boldsymbol{\lambda}_0' \mathbf{Z}' \mathbf{y} - \mu_+)$$

$$f(\mathbf{s}_\beta; \boldsymbol{\beta}) = c \exp(\boldsymbol{\beta}' \mathbf{s}_\beta - \mu_+) \sum_{\mathbf{v} \in \Gamma} \exp(\boldsymbol{\lambda}_0' \mathbf{Z}' \mathbf{v})$$

$$f(\mathbf{y} | \mathbf{s}_\beta) = \frac{1}{c \prod_{i=1}^n y_i!} \times \frac{\exp(\boldsymbol{\lambda}_0' \mathbf{Z}' \mathbf{y})}{\sum_{\mathbf{v} \in \Gamma} \exp(\boldsymbol{\lambda}_0' \mathbf{Z}' \mathbf{v})}$$

For $\mathbf{y} \in \Gamma = \{\mathbf{v} | \mathbf{X}' \mathbf{v} = \mathbf{s}_\beta\}$

Γ is referred to as *the reference set*

Exact Conditional P-value

$$\sum_{\mathbf{y} \in \Gamma} \frac{I(h(\mathbf{y}) \geq h_{obs})}{\prod_{i=1}^n y_i!} \left(\sum_{\mathbf{y} \in \Gamma} \frac{1}{\prod_{i=1}^n y_i!} \right)^{-1}$$

Example

Pathologist A	Pathologist B				
	1	2	3	4	5
1	22	2	2	0	0
2	5	7	14	0	0
3	0	2	36	0	0
4	0	1	14	7	0
5	0	0	3	0	3

- Uniform Association model (Agresti 1990)

$$\log(\mu_{ij}) = \beta_0 + \beta_i^X + \beta_j^Y + \gamma_{ij}$$

- Sufficient statistics for the β 's are the margins of the table
- Sufficient statistic for γ is $\sum_{ij} y_{ij} i j$
- 12 billion tables with the same margins
- Only 34,000 tables with the same margins and $\sum_{ij} y_{ij} i j$

Monte Carlo and Markov chain Monte Carlo alternatives

- Simulate (independent or Markovian) samples from $f(\mathbf{y}|\mathbf{s}_\beta)$
- Use the SLLN or ergodic theorem to estimate conditional expectations
- Few cases where exact no-waste i.i.d. simulation is available
- Simulate and reject algorithms are often unrealistically inefficient

Caffo and Booth Algorithm

- Modify Booth and Butler's normal candidate to perform "local updates" to increase the number of data sets and models that can be analyzed
- Create a Markov chain that updates a portion of the table while leaving the remainder fixed
- Provide this chain is irreducible and aperiodic, we can use the Metropolis/Hastings/Green algorithm to guarantee the correct invariant distribution

		Wife's Rating				Tot
		N	F	V	A	
Husband's Rating	N	7	7	2	3	19
	F	2	8	3	7	20
	V	1	5	4	9	19
	A	2	8	9	14	33
		12	28	18	33	91

		Wife's Rating				Tot
		N	F	V	A	
Husband's Rating	N	7	7	2	3	19
	F	2	8	3	7	20
	V	1	5	4	9	19
	A	2	8	9	14	33
		12	28	18	33	91

		Wife's Rating				Tot
		N	F	V	A	
Husband's Rating	N	?	7	2	?	19
	F	2	8	3	?	20
	V	?	5	?	?	19
	A	?	?	?	?	33
		12	28	18	33	91

		Wife's Rating				Tot
		N	F	V	A	
Husband's Rating	N	7	?	2	?	19
	F	2	8	3	?	20
	V	?	5	4	?	19
	A	?	?	?	14	33
		12	28	18	33	91

\mathbf{y} is Poisson with mean vector $\boldsymbol{\mu}$

$\mathbf{s}_\beta = \mathbf{X}'\mathbf{y}$ Sufficient statistic (blue area)

\mathbf{y}_1 Free area given \mathbf{s}_β (red area)

$\mathbf{X}' = [\mathbf{X}'_1 \ \mathbf{X}'_2]$ where \mathbf{X}_2 is invertible

Constructing a candidate

- \mathbf{y} is approximately multivariate normal with mean $\boldsymbol{\mu}$ and covariance matrix $D(\boldsymbol{\mu})$
($D = \text{diagonal}$)

- Then it follows that

$$\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{s}_\beta \end{bmatrix} = \begin{bmatrix} I & 0 \\ \mathbf{X}'_1 & \mathbf{X}'_2 \end{bmatrix} \mathbf{y}$$

is approximately multivariate normal

- $\mathbf{y}_1 | \mathbf{s}_\beta$ is then also approximately multivariate normal

- $P\mathbf{y}_1 | \mathbf{s}_\beta$ for permutation matrix P is also multivariate normal
- Then $P_1\mathbf{y}_1 | \mathbf{s}_\beta, P_2\mathbf{y}_1$ where $P' = [P'_1 \ P'_2]$ is also multivariate normal
- Generate from this approximation *sequentially*
“round as you go”
Easy to specify the probability of the candidate given the current state
- Solve for the **orange cells** (linear system)
- Automatically reject tables with negative entries

- Choose the number of cells in the **magenta area** as a binomial random with success probability that is treated as a tuning parameter.
- If we update few cells on average
 - The more valid tables with non-negative entries we obtain
 - The slower the mixing of the chain
- The probability generating the current table given the candidate can be calculated along with sequential means
- Use a t distribution instead of a normal
- Computing tricks ...

Benefits

- Allows for an arbitrary number of table entries to be updated at each iteration in a random order.
- t candidate attempts to model covariance structure of $\mathbf{y}|\mathbf{s}_\beta$.
- Algorithm (theoretically) works for any log linear model.
- Algorithm guaranteed to produce an aperiodic irreducible chain. Asymptotic normality guaranteed.

Example	DF	$-2\log\lambda$	P_{χ^2}	P_{MCMC}	P
I	9	15.49	.078	.118	.114
QI	5	13.55	.019	.022	.023*
QS	3	2.98	.394	.390	.393*
UA	15	16.21	.368	.044	.044
LL	40	50.27	.128	.202	
CR	55	57.29	.366	.368	

An * indicates P is a Monte Carlo approximation from another algorithm.

Example: stratified binomial counts

Let y_{ij} be independent binomial counts with sample sizes N_{ij} and success probabilities π_{ij} satisfying

$$\text{logit}(\pi_{ij}) = \beta_i + z_{ij}\lambda$$

- Monte Carlo exact Cochran/Armitage test for $\lambda = 0$. Statistic $s_\lambda = \sum_{ij} y_{ij} z_{ij}$
- Monte Carlo exact confidence interval for λ by inverting C/A test
- Monte Carlo conditional likelihood λ
- Conditional ML estimate of λ

Stratum	z_{ij}	Successes	Failures	Total
1	15	1	0	1
1	7	0	1	1
1	6	0	1	1
1	5	0	1	1
1	3	0	2	2
1	2	0	3	3
1	0	0	1	1
2	2	1	0	1
2	0	0	1	1
3	9	1	0	1
3	2	0	1	1
3	1	0	1	1
4	2	1	0	1
4	0	0	4	4
5	6	0	1	1
5	3	1	0	1
5	0	1	0	1
6	3	0	1	1
6	0	1	3	4
7	6	1	0	1
7	2	0	1	1

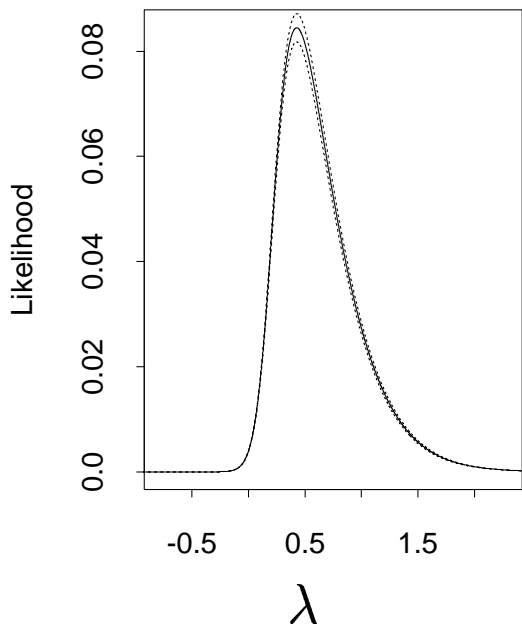
- Importance sampling the likelihood

$$\frac{\sum_{l=1}^M I(\sum_{ij} y_{ijl} z_{ij} = \mathbf{s}_\lambda) \exp(\mathbf{s}_\lambda \boldsymbol{\lambda})}{\sum_{l=1}^M \exp(\sum_{ij} y_{ijl} z_{ij} \boldsymbol{\lambda})},$$

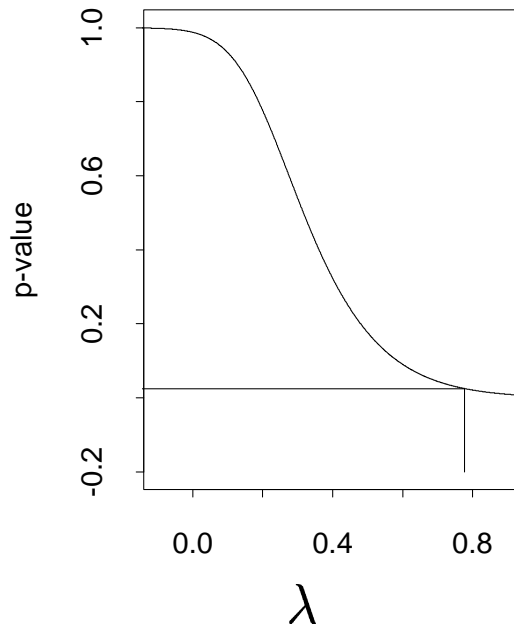
- Importance sampling conditional cumulative probabilities

$$\frac{\sum_{l=1}^M I(\sum_{ij} y_{ijl} z_{ij} \geq \mathbf{s}_\lambda) \exp(\sum_{ij} y_{ijl} z_{ij} \boldsymbol{\lambda})}{\sum_{l=1}^M \exp(\sum_{ij} y_{ijl} z_{ij} \boldsymbol{\lambda})}.$$

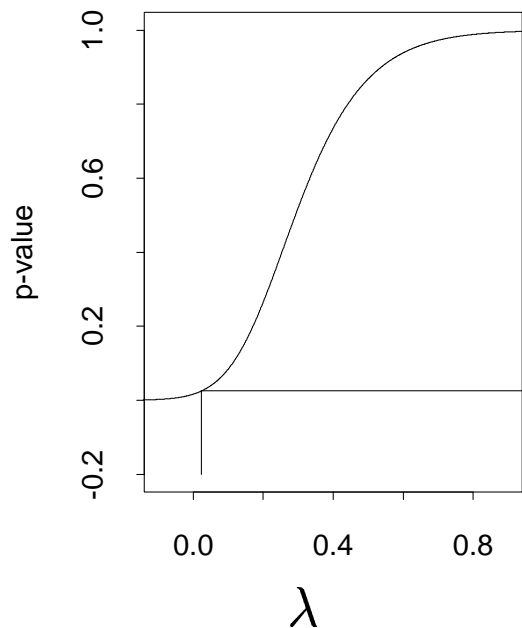
Likelihood for lambda



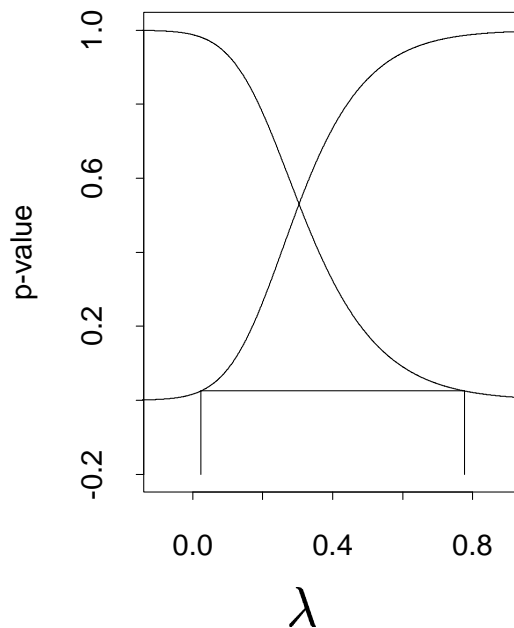
Upper 97.5% confidence limit for lambda



Lower 97.5% confidence limit for lambda



95% confidence interval for lambda



Future Work

- Monte Carlo standard errors, currently using batching
- General R library, `mcexact`

```
> library(mcexact)
> data(pathologist)
> pv <- mcexact(pathologist$y,
                 pathologist$x,
                 nosim = 10 ^ 3,
                 p = .5,
                 batchsize = 100,
                 method = "cab")

> pv
```

	V1	V2
observed.stat	16.214350396	14.72916547
pvalue	0.012000000	0.061000000
mcse	0.005966574	0.02099762

```
> pv <- update(pv)
```