Penalized likelihood estimation via data augmentation

Andrea Discacciati   Nicola Orsini

Unit of Biostatistics and Unit of Nutritional Epidemiology
Institute of Environmental Medicine
Karolinska Institutet
http://www.imm.ki.se/biostatistics/

2013 Nordic and Baltic Stata Users Group meeting

September 27, 2013

Introduction

- Bayesian analyses are rarely carried out in epidemiological research
- Partly because of the absence of Bayesian methods from most basic courses in statistics...
- ...but also because of the misconception that they are computationally difficult and require specialized software (e.g.: Stan, WinBugs)
- Yet, Bayesian methods can be a valuable tool for the analysis of epidemiological data

Aim

- Show that adequate Bayesian analyses can be carried out using standard software for frequentist analyses (e.g.: Stata)
- This can be done through penalized likelihood estimation via data augmentation

Priors

- A prior for a parameter β is a probability distribution that reflects one's uncertainty about β before the data under analysis is taken into account
- Focus on normal priors for $\log(RR) = \beta \sim N(\beta_{prior}, v_{prior})$
- These priors are symmetric: mean=median=mode=β_{prior}
- Equivalently, these are log-normal priors for $\exp\{\beta\} = RR$
- Prior specification can be done in terms of prior limits for RR rather than in terms of mean and variance for β
- 95% prior limits: $\Pr(RR_{lower} < RR < RR_{upper}) = 0.95$ if one disregarded the analysis data
- β_{prior} and v_{prior} are back-calculated from RR_{lower} and RR_{upper}

How to fit a Bayesian model

A partial list:

- Inverse-variance weighting (information-weighted averaging)
- Posterior sampling (e.g.: Markov chain Monte Carlo (MCMC))
- Penalized likelihood

Penalized likelihood (PL)

- A PLL is just the log-likelihood with a penalty subtracted from it
- The penalty will pull or shrink the final estimates away from the Maximum Likelihood estimates, toward β_{prior}
- Penalty: squared L_2 norm of $(\beta - \beta_{prior})$

Penalized log-likelihood

$$\tilde{\ell}(\beta; x) = \log\left[\mathcal{L}(\beta; x)\right] - \frac{r}{2} \| (\beta - \beta_{prior}) \|_2^2$$

- Where $r = 1/v_{prior}$ is the precision (weight) of the parameter β in the prior distribution

Penalized likelihood (PL)

- Parameter vector $\mathbf{b} = (\beta_1, \ldots, \beta_j) = (\log(RR_1), \ldots, \log(RR_j))$
- $\mathbf{b} \sim MVN(\mathbf{b}_{prior}, \mathbf{V}_{prior})$
- $\mathbf{b}_{prior} = \left(\beta_{prior_1}, \ldots, \beta_{prior_j}\right)$
- $\mathbf{V}_{prior} = diag\left(v_{prior_1}, \ldots, v_{prior_j}\right)$

Penalized log-likelihood

$$\tilde{\ell}(\mathbf{b}; \mathbf{x}) = \log[\mathcal{L}(\mathbf{b}; \mathbf{x})] - (\mathbf{b} - \mathbf{b}_{prior})^T \mathbf{V}_{prior}^{-1} (\mathbf{b} - \mathbf{b}_{prior})/2$$

Penalized likelihood (PL)

Link between PL and Bayesian models

From a Bayesian perspective, quadratic log-likelihood penalization corresponds to having independent normal priors on $\mathbf{b}$

- PL estimation allows semi-Bayesian analyses, i.e. where some but not all model parameters are given an explicit prior

Data-augmentation priors (DAPs)

- An equivalent way of maximizing the PLL is utilizing DAPs
- Prior distributions on the parameters are represented by prior data records created ad hoc
- Prior data records generate a quadratic penalty function that imposes the desired priors on the model parameters
- Estimation carried out using standard ML machinery on the augmented dataset (i.e. original and DAP records)

Advantage of PL via DAPs

This method allows one to carry out Bayesian analyses with any statistical software, exploiting commands that are readily available (e.g.: `glm` command in Stata)

Data-augmentation priors (DAPs)

- DAPs are not only a tool to fit Bayesian models

Advantage of PL via DAPs

DAPs are one way of understanding the logical strength of a prior distribution

- What hypothetical experiment would convey the same information as the proposed 95% prior limits for RR?
- After translating the prior to equivalent data, one might see that the original prior was, for example, overconfident

Example: Logistic regression

- Case-control study on the relation of maternal antibiotic use during pregnancy (X = 1) to sudden infant death syndrome (Y = 1)

	Antibiotic use		
	$X = 1$	$X = 0$	Total
Cases ($Y = 1$)	173	602	775
Controls ($Y = 0$)	134	663	797
Total	307	$1,265$	$1,572$

- Odds Ratio = 1.42 (95% Wald C.I.: 1.11, 1.83)

Example: Logistic regression

- Dataset for the analysis

```
clear

input x   y     n
      0   602   1265
      1   173   307
end
```

- Suppose that strong associations are unlikely
- A plausible prior for $\log(OR) = \beta \sim N(0, 0.5)$
- 95% Wald prior limits for $OR : \exp\{0 \pm 1.96\sqrt{0.5}\} \approx (0.25, 4.00)$

Example: Logistic regression

- $\tilde{\ell}(\beta_0, \beta_1; x) = \sum_i \{\log[\text{expit}(\beta_0 + \beta_1 x_i)]\, y_i + \log[1 - \text{expit}(\beta_0 + \beta_1 x_i)]\,(n_i - y_i)\} - \|\beta_1\|_2^2$

PLL maximized using `mlexp` in Stata 13

```
mlexp (log(invlogit({b0}+{xb:x}))*y + ///
log(1-(invlogit({b0}+{xb:})))*(n-y) - ({xb_x}^2)/2)
```

```
lincom [xb_x]_cons, eform
------------------------------------------------------------------------------
             |     exp(b)   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   1.406055   .1771661     2.70   0.007     1.098371    1.799931
------------------------------------------------------------------------------
```

- OR_{post} (95% Wald posterior limits) $= 1.41\ (1.10, 1.80)$
- Semi-Bayesian analysis because we do not impose a prior on β_0

Example: Logistic regression

- Estimation using DAPs
- The prior $N(0, 0.5)$ roughly corresponds to an hypothetical (and unethical) RCT with 4 cases in each arm

	Antibiotic use	
	$X = 1$	$X = 0$
Cases ($Y = 1$)	4	4
Controls ($Y = 0$)	$100,000$	$100,000$

- OR_{prior} (95% Wald prior limits) $\approx 1.00\ (0.25, 4.00)$

Example: Logistic regression

- Augmented dataset

```
clear

input  x  y  n  cons
 0   602  1265  1
 1   173   307  1
 1     4     8  0
end
```

- Check that prior data gives back the desired prior

PL via DAPs using glm

```
glm y x cons, family(binomial n) eform nocons
```

y	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
x	1.406201	.1772654	2.70	0.007	1.098361	1.80032
cons	.9099392	.0510718	-1.68	0.093	.8151497	1.015751

- OR_{post} (95% Wald posterior limits) $= 1.41\ (1.10, 1.80)$

Example: Logistic regression

- We developed a Stata command that takes care of generating the DAPs and fitting the penalized logistic model

PL via DAPs using plogit

```
plogit y x, prior(x 0.25 4) binomial(n) or s(1)
```

```
Penalized logistic regression                          No. of obs   =          2
Prior _b[x]: Normal(0.000, 0.500)
------------------------------------------------------------------------------
           y | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |    1.40621   .1772681     2.70   0.007     1.098365    1.800335
       _cons |   .9099382   .0510718    -1.68   0.093     .8151486     1.01575
------------------------------------------------------------------------------
```

- OR_{post} (95% Wald posterior limits) $= 1.41\ (1.10, 1.80)$

Example: Logistic regression

- Check the compatibility between the data and the prior
- $c = \left(\left(\beta_{observed} - \beta_{prior} \right) / \left(v_{observed} + v_{prior} \right)^{\frac{1}{2}} \right)^2 \sim \chi_1$
- In Stata

```
scalar c = ((.3408918 - 0) / sqrt(.1260598^2 + 0.5))^2

scalar p = chi2tail(1, scalar(c))

di %5.3f scalar(p)
0.635
```

- No evidence of incompatibility between the frequentist results and the prior ($p = 0.635$)

Example: Logistic regression

- Comparison with Markov chain Monte Carlo

MCMC using Stan (from R)

```
fitl <- stan(model_code = binomial, data = sids,
        iter = 10000, chains = 4, seed = 1983)
```

- Plus other $\approx$ 20 lines of code: not very user-friendly for a 2x2 table
- Results from the three analyses are similar

	OR_{post}	95% posterior limits
Direct PLE (`mlexp`)	1.406	$(1.098, 1.799)$
PLE via DAPs (`plogit`)	1.406	$(1.098, 1.800)$
MCMC (`stan`)	1.408	$(1.115, 1.778)$

Example: Poisson regression

- Cohort study on smoking and overall mortality among male British doctors (Doll and Peto, 1976)
- Baseline information on:
- Smoking habits (yes, no) (exposure)
- Age category (35-44, 45-54, 55-64, 65-74, 75-84) (potential confounder)
- 731 deaths (630 among smokers, 101 among non smokers)

```
webuse dollhill3, clear

describe
[... output omitted ...]
-------------------------------------------------------------------------
              storage   display    value
variable name   type    format     label      variable label
-------------------------------------------------------------------------
agecat          byte    %9.0g      agelbl     age category
smokes          byte    %9.0g                 whether person smokes
deaths          int     %9.0g                 number of deaths
pyears          float   %9.0fc                person-years
-------------------------------------------------------------------------
```

Example: Poisson regression

- Frequentist analysis (no explicit prior on β_{smokes})
- This corresponds to an implicit prior $N(0, +\infty)$
- This prior gives equal odds on IRR$=10^{-10}$, IRR$=1$ or IRR$=10^{10}$

```
xi: poisson deaths smokes i.agecat, exposure(pyears) irr

------------------------------------------------------------------------------
      deaths |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      smokes |   1.425519   .1530638     3.30   0.001     1.154984    1.759421
  _Iagecat_2 |   4.410584   .8605197     7.61   0.000     3.009011    6.464997
[... output omitted ...]
       _cons |   .0003636   .0000697   -41.30   0.000     .0002497    .0005296
  ln(pyears) |          1  (exposure)
------------------------------------------------------------------------------
```

- IRR (95% Wald C.I.) = 1.42 (1.15, 1.76)

Example: Poisson regression

- We specify the prior for $\log(IRR_{smokes})$ in terms of 95% prior interval
- 95% Wald prior limits for $\text{IRR}_{smokes} = (1.50,\ 2.50)$
- This corresponds to a prior for $\log(IRR_{smokes}) \sim N(\log(1.94), 0.017)$
- Hypothetical RCT with 118 deaths in each arm

	Smoking	
	$X = 1$	$X = 0$
Deaths	118	118
Person-years	$100,000$	$194,000$

- IRR_{prior} (95% Wald prior limits) $\approx 1.94\ (1.50, 2.50)$

Example: Poisson regression

- $\tilde{\ell}(\mathbf{b}; \mathbf{x}) = \sum_i \{deaths_i \left(\mathbf{x}_i^T \mathbf{b} + \log(pyears_i)\right) - \exp\{\mathbf{x}_i^T \mathbf{b} + \log(pyears_i)\}\} - \frac{1}{2} 0.017^{-1} \|\beta_{smokes} - \log(1.94)\|_2^2$

PLL maximized using `mlexp` in Stata 13

```
mlexp (deaths*({b0}+{xb:smokes _Iagecat_?} + ///
 log(pyears))-exp({b0}+{xb:}+log(pyears)) - ///
  .5*0.017^(-1)*({xb_smokes}-log(1.94))^2/10)
```

```
lincom [xb_smokes]_cons, eform
------------------------------------------------------------------------------
             |     exp(b)   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   1.620877   .1379238     5.68   0.000     1.371891    1.915052
------------------------------------------------------------------------------
```

- IRR_{post} (95% Wald posterior limits) $= 1.62\ (1.37, 1.91)$

Example: Poisson regression

- We developed a command for penalized Poisson regression via DAPs

PL via DAPs using ppoisson

```
xi:  ppoisson deaths smokes i.agecat, exposure(pyears) ///
               prior(smokes 1.50 2.50) irr
```

```
Penalized poisson regression                          No. of obs   =         10
Prior _b[smokes]: Normal(0.661, 0.017)
------------------------------------------------------------------------------
      deaths |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      smokes |   1.618651   .1380122     5.65   0.000     1.369546    1.913066
   _Iagecat_2 |    4.38198   .8547662     7.57   0.000     2.989728    6.422574
[... output omitted ...]
       _cons |   .0003281   .0000608   -43.33   0.000     .0002283    .0004717
------------------------------------------------------------------------------
```

- IRR_{post} (95% Wald posterior limits) $= 1.62\ (1.37, 1.91)$

Example: Poisson regression

- Comparison with Markov chain Monte Carlo

MCMC using Stan (from R)

```
fitp <- stan(model_code = poisson, data = dollhill3,
        iter = 10000, chains = 4, seed = 1492)
```

- Results from the three analyses are, again, similar

	IRR_{post}	95% posterior limits
Direct PLE (mlexp)	1.621	(1.372, 1.915)
PLE via DAPs (ppoisson)	1.619	(1.370, 1.913)
MCMC (stan)	1.623	(1.375, 1.916)

Sparse data

- Bayesian approach can be useful to address the sparse-data problem
- Data with few or no subjects at crucial combinations of variables (e.g.: few exposed cases)
- Prior pulls the parameter towards its prior expected value (β_{prior}) and the degree of adjustment is determined by v_{prior}
- Frequentist perspective: prior (penalty) as a smoothing device (ridge regression)
- Profile-likelihood limits are generally preferable with sparse data

Example: Sparse data

- Data from a study of obstetric care and neonatal death (Y = 1). The exposure is hydramnios during pregnancy (X = 1). (Neutra et al., 1978; Sullivan and Greenland, 2013)

	Hydramnios		
	$X = 1$	$X = 0$	Total
Deaths ($Y = 1$)	1	16	17
Survivals ($Y = 0$)	9	$2,966$	$2,975$
Total	10	$2,982$	$2,992$

- OR = 20.59 (95% profile-likelihood C.I.: 1.08, 119.57)
- OR is about an order of magnitude above clinical expectation

Example: Sparse data

- 95% Wald prior limits for $OR_{hydram} = (1, 16)$, corresponding to a "probably strong" association (centered around 4)

Profile-posterior limits using plogit

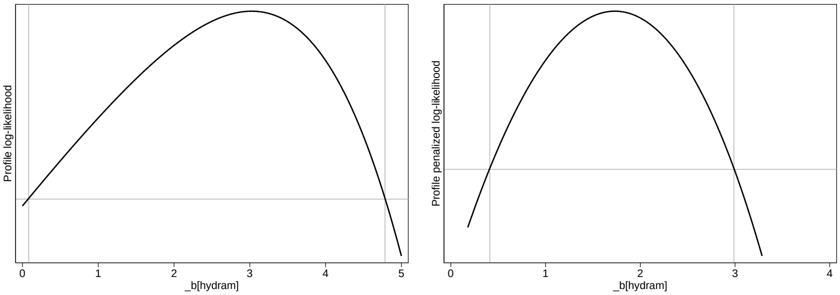
```
plogit deaths hydram, bin(n) p(hydram 1 16) pl(hydram) or
```

```
Penalized logistic regression                         No. of obs   =          2
Prior _b[hydram]: Normal(1.386, 0.500)
------------------------------------------------------------------------------
      deaths | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      hydram |   5.653545   3.733989     2.62   0.009     1.549277    20.63064
       _cons |    .005629    .001371   -21.27   0.000     .0034923    .0090728
------------------------------------------------------------------------------
-----------------------------------------
      deaths |  [95% PLL Conf. Interval]
-------------+---------------------------
      hydram |       1.509143    19.84804
-----------------------------------------
```

- IRR_{post} (95% profile-posterior limits) $= 5.65\ (1.51, 19.85)$

Example: Sparse data

- Bayesian results appear clinically more reasonable ($OR \approx 6$)
- The effect of the prior (penalty) on the asymmetry of the profile log-likelihood for β_{hydram} is evident



Conclusions

Strengths of PLE via data augmentation priors

- Can be used to conduct Bayesian and semi-Bayesian analyses
- DAPs provide a critical perspective on the proposed priors
- Useful tool to address sparse-data artefacts (with the advantage of incorporating prior information)
- Computationally easier than simulation methods (e.g.: MCMC)
- Easily implemented in Stata (`glm`, `plogit`, `ppoisson`)

Caveats

- Approximate posterior mode (β_{post}) and 95% posterior limits (but adequate in the context of observational epidemiology)
- Uses same large-sample approximations as ML (but more stable)
- Profile-posterior limits if the posterior distribution is non-normal

References

- Greenland, S. (2006). Bayesian perspectives for epidemiologic research. I. Foundations and basic methods. International Journal of Epidemiology, 35, 765-778.
- Greenland, S. (2007). Bayesian perspectives for epidemiologic research. II. Regression analysis. International Journal of Epidemiology, 36, 195-202.
- Greenland, S. (2007). Prior data for non-normal priors. Statistics in Medicine, 26, 3578-3590.
- Rothman K.J., Greenland S. and Lash T.L. (2008). Introduction to Bayesian statistic (ch. 18), in Modern epidemiology. Philadelphia, PA: Lippincott, Williams & Wilkins.
- Sullivan, S., and Greenland, S. (2013). Bayesian regression in SAS software. International Journal of Epidemiology, 42, 308-317.