Solutions to the problem of monotone likelihood in Cox and logistic regression

Theory and applications

Georg Heinze
Core Unit of Medical Statistics and Informatics
Medical University of Vienna
georg.heinze@meduniwien.ac.at

Outline

- Monotone likelihood: examples
- Options of analysis
- Penalized likelihood (PL) estimates
- Profile PL confidence intervals
- Application to Cox, unconditional and conditional logistic regression
- General comments about bias reduction and PL estimates

Example 1: Preterm infants

From: Berger et al, J Perinat Med, 2003

Group	**No CLD**	**CLD**
No amniotic cavity culture found	40	0
Ureaplasma urealyticum found	17	4

OR estimate: 4/0 / 40/17 = infinite

Example 2: Urinary tract infection

From: Foxman et al, Epidemiology, 1997

Group	No infection	Infection
No diaphragm use	109	123
Diaphragm use	0	7

Other variables in the model:

Age, use of condoms, use of lubricated condoms, use of spermicides, oral contraceptives

Example 2: Urinary tract infection

- OR estimates of diaphragm use obtained by glm of SPLUS:
- Convergence criterion is change in deviance

Criterion	**Estimate**	**Lower**	**Upper**	**P-value**
0.001	220	0.8	6.5e5	0.06
0.0001	1726	4e-4	6.8e9	0.34
0.00001	34774	<1e-4	1.1e34	0.76

Example 3: Breast cancer

- From Lösch et al, Brit J Cancer, 1998
- Survival of 100 patients, 74 censored
- 4 risk factors (pT, N, G, CD)
- Analysis via Cox regression:

Faktor	**RR (95% c.i.)**	**P-value**
pT	3.6 (1.3 – 9.6)	0.01
N	2.6 (1.1 – 5.9)	0.03
G	248054 (0 – 2 x 10^{188})	0.95
CD	1.5 (0.6 – 3.6)	0.37

Common to examples 1-3:

- Degenate variation of outcome in one subgoup
  - Ex 1: no CLD+ for no ureaplasms found
  - Ex 2: only „infections“ for diaphragm users
  - Ex 3: no deaths for G=0

- Parameter estimates $\hat{\beta}$ are infinite
- Standard errors infinite
- Confidence interval [-∞, +∞] uninformative
- $\hat{\beta}$/se -> 0

First occurrance in literature

- In Cox regression:

  „*Monotone likelihood*“

  (Bryson and Johnson, Technometrics, 1981)

- In logistic regression:

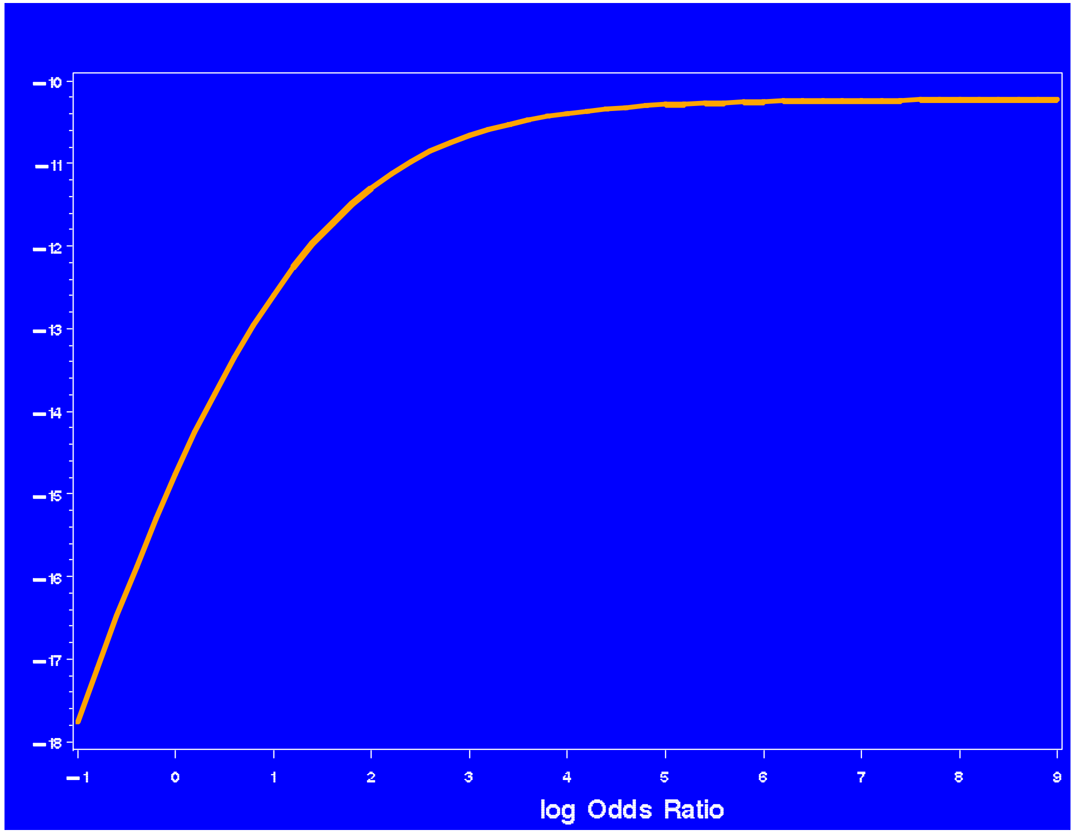
  „*(Quasi-)complete Separation*“

  (Day and Kerridge, Biometrics 1967)

Monotone likelihood

- •Likelihood is monotone
- •no finite maximizer

- •Likelihood is flat
- •second derivative is 0
- •variance is infinite



Incidence of monotone likelihood

- High incidence if:
- Small N / heavy censoring
- Unbalanced covariates
- Large underlying effects
- Strong correlation among covariates

Options of analysis

- Omit covariate X that is causing monotone likelihood

- Stratify analysis by covariate X

- Choose different type of model
  - Transform X (e.g. use quasi-metric scaling instead of dummies)
  - Use additive risk model instead of multiplicative risk model

- Ad hoc data adjustment

○ Haldane:	add ½ to each cell (for rx2 tables)
○ Laplace:	add 1 to each cell (for rx2 tables)
○ Clogg et al, JASA 1991:	generalized data adjustment

- Exact logistic regression (LogXact, Cytel Software Corp.)

Why exact logistic regression?

- Should read
  „*Exact conditional logistic regression*“

- Implementations:
  - LogXact
  - SAS/PROC LOGISTIC (from V8.1 on)

- *Exact*: inference based on exact distribution of sufficient statistic under null hypothesis

- *Conditional*: eliminate nuisance parameters by conditioning on sufficient statistics

- Point estimate is not „exact“, rather conditional

Exact conditional log regression

- Maximum conditional likelihood estimate (MCLE):
  - $\Pr(T=T_{obs}|\beta)$=max!
  - Infinite if T_{obs} is largest (smallest) possible value of T
- Median unbiased estimate (MUE):
  - $\Pr(T \geq T_{obs} \mid \beta) \geq 0.5$, $\Pr(T \leq T_{obs} \mid \beta) \geq 0.5$
  - MUE is defined even if MCLE is infinite
- LogXact: MUE replaces MCLE if MCLE is infinite

Exact conditional log regression

- Can even be (ab)used for estimating a Cox model:
  - Each risk set contributes a nuisance parameter that is eliminated by conditioning
  - Conditioning on risk sets improves on asymptotic Cox model, but still violates nominal significance level because of interdependence of risk sets
  - shown for exact logrank test in Heinze et al (2003)

- Problems if exact null distribution is (nearly) degenerate:
  - Conditioning on continuous covariates
  - Too many covariates, too many different levels of covariates
- If exact null distribution is degenerate:

  no estimation/inference possible

Example 4: Lung cancer

- Case-control study, 18 matched sets, 1:m matching
- Factors smoking (S), radiation (R), RxS

- Options of analysis:
  - CML (conditional maximum likelihood)
    - conditions on matched sets
    - but estimates effects S, R, RxS simultanously
  - CXL (conditional exact maximum likelihood),
    - conditions on matched sets
    - and eliminates other effects by conditioning

Example 4: Lung cancer, OR (95% ci)

Method	Radiation	Smoking	RxS
CML	1.2 (0.17, 8.5)	21 (2.6, 167)	∞ (0.0, ∞)
CXL	1.2 (0.14, 20)	20 (2.6, 859)	2.5 (0.06, ∞)

•CXL estimate for RxS is a MUE,
based on a distribution consisting of 2 possible values only
(overconditioning)

•Had the other of these two values been observed:
CML estimate = - ∞,
but CXL estimate = still 2.5

•Therefore, CXL estimate and c.i. are very conservative!

A solution through bias reduction

- Firth, Biometrika 1993:
- Eliminate O(1/n)-bias from maximum likelihood parameter estimates

- Bias reduction is applied while estimating the parameter estimates:

  bias preventing, not *bias correcting*

- Maximize a penalized likelihood:

  logL* = logL + ½ log det (I)

  with I denoting Fisher information matrix

A solution through bias reduction (2)

- Firth‘s paper remained undiscovered for 8 years
- Application to log reg and Cox reg possible (Heinze and Schemper, 2001, 2002)
- We showed that parameter estimates are always finite
- Small-sample bias is greatly reduced

Penalized maximum likelihood estimation

- Penalisation by ½ log det(I) is like adding pseudo-observations with total weight *k* to the data

- Log Reg: each observation (x_i, y_i) is splitted into two new observations:
  - A: Outcome y_i, weight $1+h_i/2$
  - B: Outcome $1-y_i$, weight $h_i/2$
  - h_i are diagonal elements of hat matrix H (leverages)
  - $\Sigma h_i = k$; $0 \leq h \leq 1$
  - Balance of pseudo-observations guarantees finite estimates

- Weighting of pseudo-observations is done iteratively during estimation

Inference

- Standard errors deduced from second derivative of log L (rather than log L*)
- Although penalized likelihood has maximum, its shape is very asymmetric in case of monotone likelihood
- Normal approximation unsuitable
- Better: Profile penalized likelihood confidence intervals, penalized likelihood ratio tests

Profile penalized likelihood

- Likelihood ratio statistic

$$LR(\hat{\gamma},\gamma_0)=2\left\{\log L(\hat{\gamma},\hat{\delta})^{*}-\log L(\gamma_0,\hat{\delta}_{\gamma_0})^{*}\right\}$$

- Under H_0: $\gamma = \gamma_0$, LR ~ χ^2

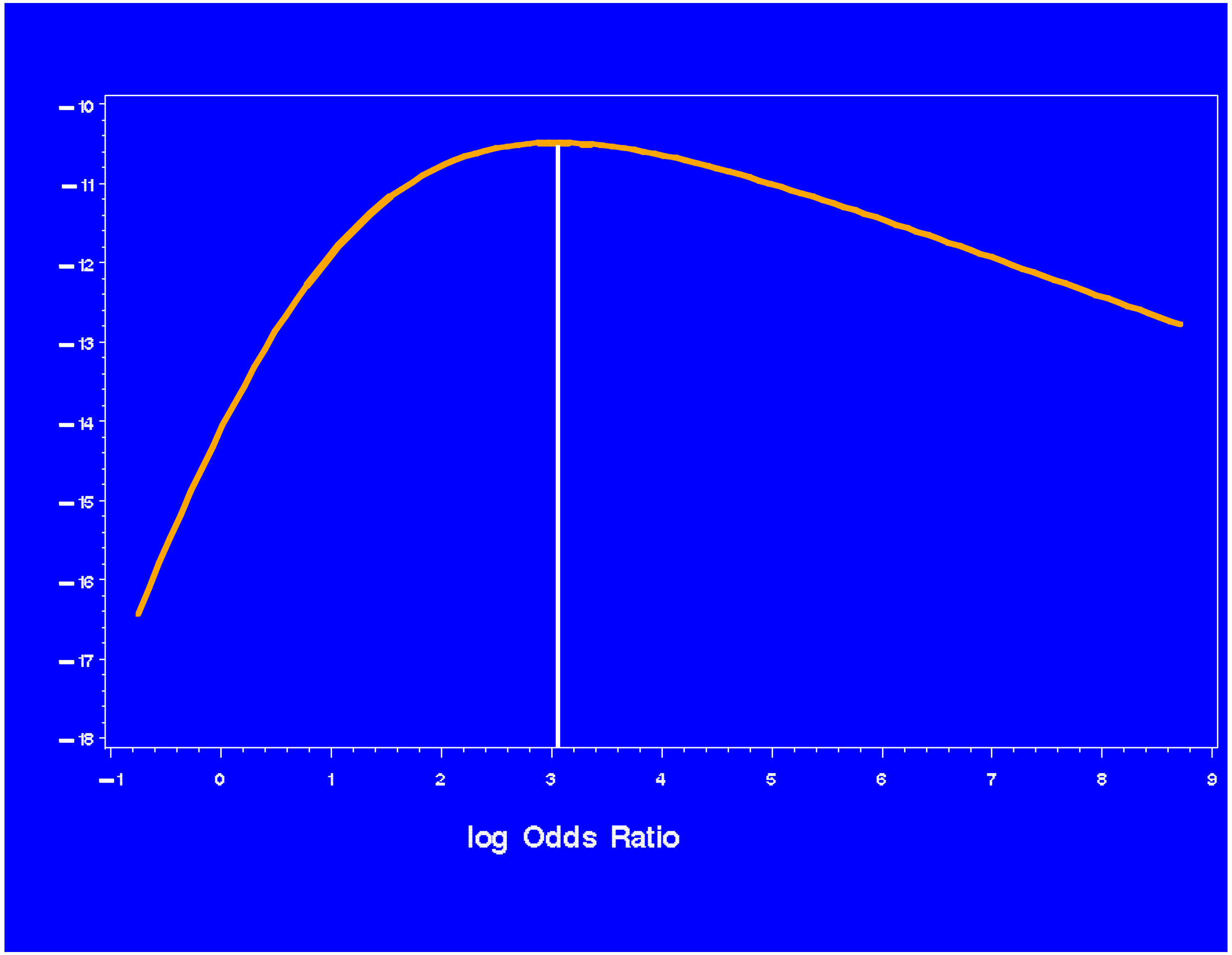
- 95% confidence interval:

  set of values γ_0 for which LR< $\chi^2_{1;0.95}$

Example 1: Preterm infants

OR (95% CI):

20.8
(2.1 – 2017)
p=0.007



Example 2: Urinary tract infection

	OR estimate	
Factor	**ML**	**PML**
Age	3.2	3.0
Oral contraceptives	0.9	0.9
Condom use	11.1	9.7
Lubricated condom	0.1	0.1
Spermicide use	0.4	0.5
Diaphragm use	INF	22.1

Example 3: Breast cancer

	ML RR (95% CI)	**P**	**PML RR (95% CI)**	**P**
pT	3.6 (1.3 – 9.6)	0.01	3.4 (1.4 – 9.5)	0.01
N	2.6 (1.1 – 5.9)	0.03	2.5 (1.1 – 5.8)	0.03
G	248054 (0 – 2 x 10^{188})	0.95	*11.3 (1.5 – 1452)*	*0.01*
CD	1.5 (0.6 – 3.6)	0.37	1.5 (0.6 – 3.5)	0.36

Example 4: Lung cancer, OR (95% ci)

Method	Radiation (main effect)	Smoking (main effect)	RxS
CML	1.2 (0.17, 8.5)	21 (2.6, 167)	∞ (0.0, ∞)
CXL	1.2 (0.14, 20)	20 (2.6, 859)	2.5 (0.06, ∞)
CPML	*0.99 (0.2, 3.1)*	*14 (3.1, 128)*	*11 (0.4, 1800)*

- Our approach can easily be adopted to CML log reg

Example 5: Childhood leukemia matched-pairs study

- Ebi et al, Epidemiology 1999; Greenland, 2000
- Case: house with an index case of leukemia
- Control: reflection of case house across the street
- Exposure: backyard power line (3-phase, secondary, none)
- Standard analysis via conditional maximum likelihood (CML) log reg

Example 5: Childhood leukemia matched-pairs study

Case exposure	Control exposure		
	Three-phase	Secondary	None
Three-phase	15	24	11
Secondary	11	107	9
None	0	1	81

Example 5: Childhood leukemia matched-pairs study

- No separation, but sparse data: CML expected to be biased away from 0

	OR (95% CI)	
Method	Three-phase vs none	Secondary vs none
CML	32 (4.0, 0.253)	14 (1.8, 107)
CXL	30 (4.5, 1328)	14 (2.1, 507)
CPML	21 (3.7, 124)	9.6 (2.4, 87)
Haldane (add ½ to cells)	16 (3.5, 78)	7.4 (1.6, 34)
Laplace (add 1 to cells)	11 (2.9, 43)	5.2 (1.4, 19)

CML: conditional ML; CXL: conditional exact ML;
CPML: conditional penalized ML

Penalized likelihood/bias reduction

- log L* = log L + c log det (I)
- Jeffreys prior: c = ½, removes O(1/n) bias
  - shrinks parameter estimates toward the point of minimum variance
  - shrinkage not equal for each parameter
- c > ½: reduces bias on exp(β) scale, but introduces negative bias on estimate of β (Greenland, 2000)
  - c=1: generalization of Laplace ad-hoc estimator

Comments on bias reduction

- Firth: $O(n^{-1})$-bias reduction
  - Optimal to reduce bias and MSE

- Should we try to obtain higher-order bias reduction?
  - MSE=bias^2 + variance
  - bias smaller, but variance greater => MSE worse

Bias reduction if n/k is small

- Since estimates exist in each and every situation, we are seduced to analyze samples with very small n and very large k
- Watch out! Firth‘s bias reduction overcorrects bias if n/k is small
- This means, a negative bias (bias towards zero) is introduced
- Overcorrection becomes severe (effects are halvened) if y is unbalanced AND n/k is small

Why the overcorrection?

- Bias reduction implicitely estimates bias b(β) replacing β by its estimate
- If estimate is inconsistent, approximation fails (Leung and Wang, 1998)
- Common to all bias correcting approaches that need to estimate the bias
- Pseudo-observations obtain to much weight (k compared to n)

Bias reduction if n/k is small (2)

- Smaller amount of correction necessary
- Maximize Log L* = log L + c log det (I) with $c < ½$
- Optimization of c via simulation (?)
- Or switch to ridge regression (leCessie and Van Houwelingen, 1992)

Penalized likelihood: ridge regression

- leCessie and Van Houwelingen, JRSS C 1992:
- $\text{Log L * = log L} - \lambda \,||\beta||^2$
  - $||\beta||^2 = (\Sigma\ \beta_j^{\,2})^{1/2}$
  - λ optimized to yield small prediction error
  - Shrinks parameter estimates towards 0
  - Can be used to apply restrictions on estimation, e.g.
    - Smooth transitions of parameter estimates corresponding to neighbouring categories
    - Smooth transitions of hazards in piecewise exponential models

  - Purpose: primarily for prediction, not for estimation of parameter estimates
  - In case of separation, produces finite estimates

Model comparison with penalized maximum likelihood

- Comparison of models is difficult:
  - Hierarchical models: use penalization term of larger model
    - Model 1: y=A+B+C
    - Model 2: y=A+B

    - Log L1* := log L(A,B,C) + ½ log det I(A,B,C)
    - Log L2* := log L(A,B,C) + ½ log det I(A,B,C) with β_C=0

    - Please note that in this comparison,
      Log L2* :≠ log L(A,B) + ½ log det I(A,B) !!!

    - That‘s how inference is performed in our programs

Model comparison (2)

- Non-hierarchical model comparison
  - Model 1: y=A+B+C
  - Model 2: y=A+B+D

- Penalized likelihoods cannot be compared, because the structure of penalization term

  ½ log det (I) is not comparable

- Comparison via some information criterion (DIC? BIC?)

- Still an open issue

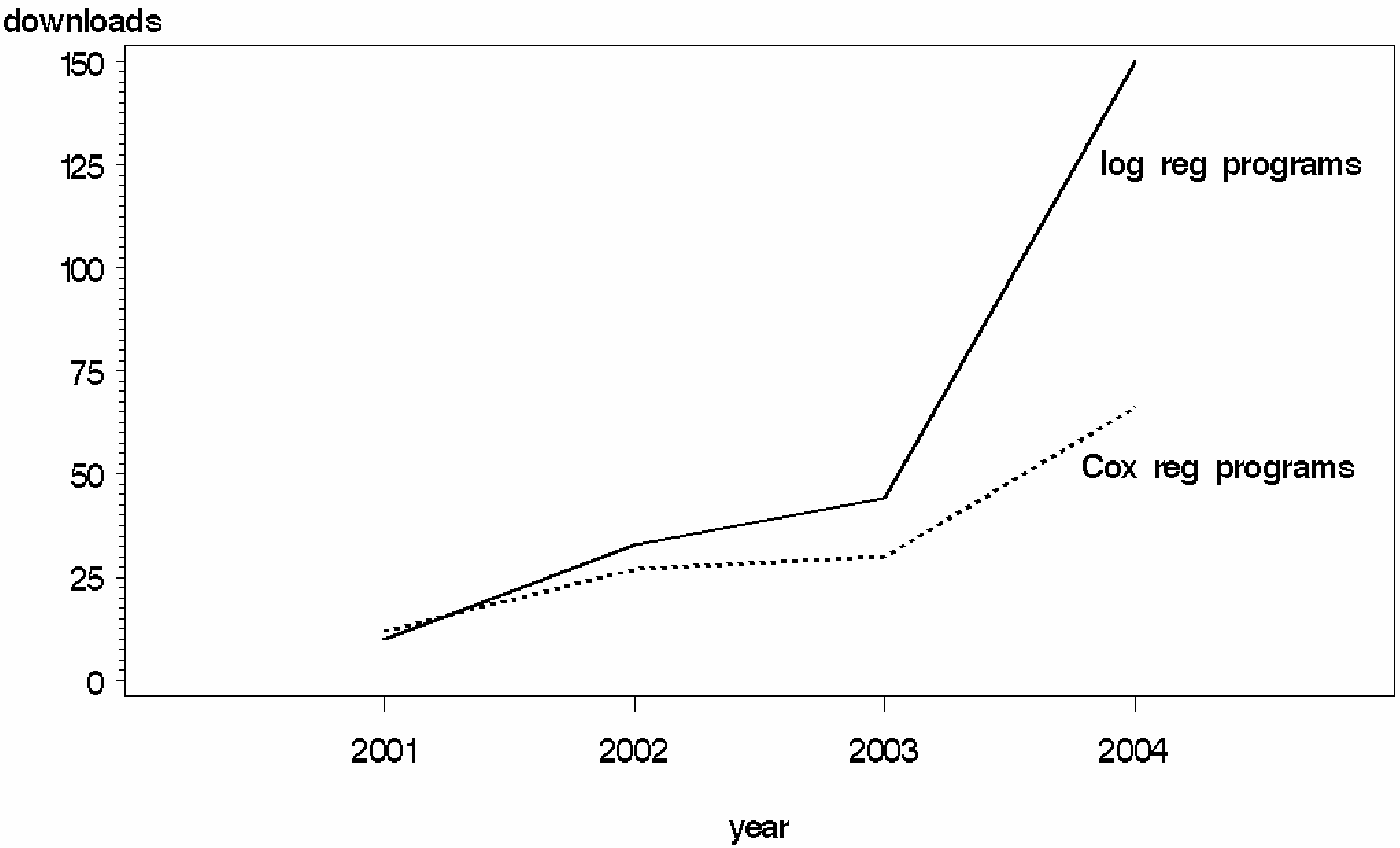
Software

- Cox regression
  - SAS macro **FC** (Heinze and Ploner, 2002)
  - SPLUS function **coxphf** (Heinze and Ploner, 2002)
  - New SAS macro **FGCSS** (time-dependent variables/effects, counting process style, stratified analysis)
  - All based on FORTRAN

- Logistic regression
  - SAS macro **FL** (Heinze and Ploner, 2003)
  - SPLUS function **logistf** (Heinze and Ploner, 2003)
  - R package **logistf** (Heinze and Ploner, 2004)
  - R package **brlr** (by D. Firth)

  - Conditional logistic regression:
    - 1:1 matching: FL/logistf, suppress estimation of intercept
    - 1:m matching: FGCSS

Software: registered downloads



Conclusions

- Penalized likelihood approach removes the problem of reporting infinite odds/risk ratios
- PPL confidence intervals account properly for assymetry of likelihood
- PML estimates have smaller bias than ML estimates
- PPL confidence intervals have better coverage than PL or Wald ci
- Approach works, better than others, for all normal problems
- Software is available, have a look at
  - www.muw.ac.at/msi/biometrie/programme/fl
  - www.muw.ac.at/msi/biometrie/programme/fc

References

- Day and Kerridge (Biometrics 1967). A general maximum likelihood discriminant.
- Bryson and Johnson (Technometrics 1981). The incidence of monotone likelihood in Cox regression.
- Clogg, Rubin, Schenker, Schultz and Weidman (J Am Stat Assoc 1991). Multiple imputation of industry and occupation codes in census public-use samples using Bayesian logistic regression.
- Le Cessie and van Houwelingen (Applied Statistics 1992). Ridge estimators in logistic regression.
- Firth (Biometrika 1993). Bias reduction of maximum likelihood estimates.
- Leung and Wang (Austr New Zeal Journal of Statistics 1998). Bias reduction using stochastic approximation.
- Greenland (Biostatistics 2000). Small-sample bias and corrections for conditional maximum-likelihood odds-ratio estimators.
- **Heinze and Schemper (Biometrics 2001). A solution to the problem of monotone likelihood in Cox regression.**
- **Heinze and Schemper (Statistics in Medicine 2002). A solution to the problem of separation in logistic regression.**
- Heinze and Ploner (Comp Meth Prog Biomed 2002). SAS and SPLUS programs to perform Cox regression without convergence problems.
- Heinze, Gnant and Schemper (Biometrics 2003). Exact logrank tests for unequal follow-up.
- Heinze and Ploner (Comp Meth Prog Biomed 2003). Fixing the nonconvergence bug in logistic regression using SPLUS and SAS.
- Heinze and Ploner (TechRep, 2004). A SAS macro, SPLUS library and R package to perform logistic regression without convergence problems. http://www.meduniwien.ac.at/msi/biometrie/programme/fl/tr2_2004.pdf