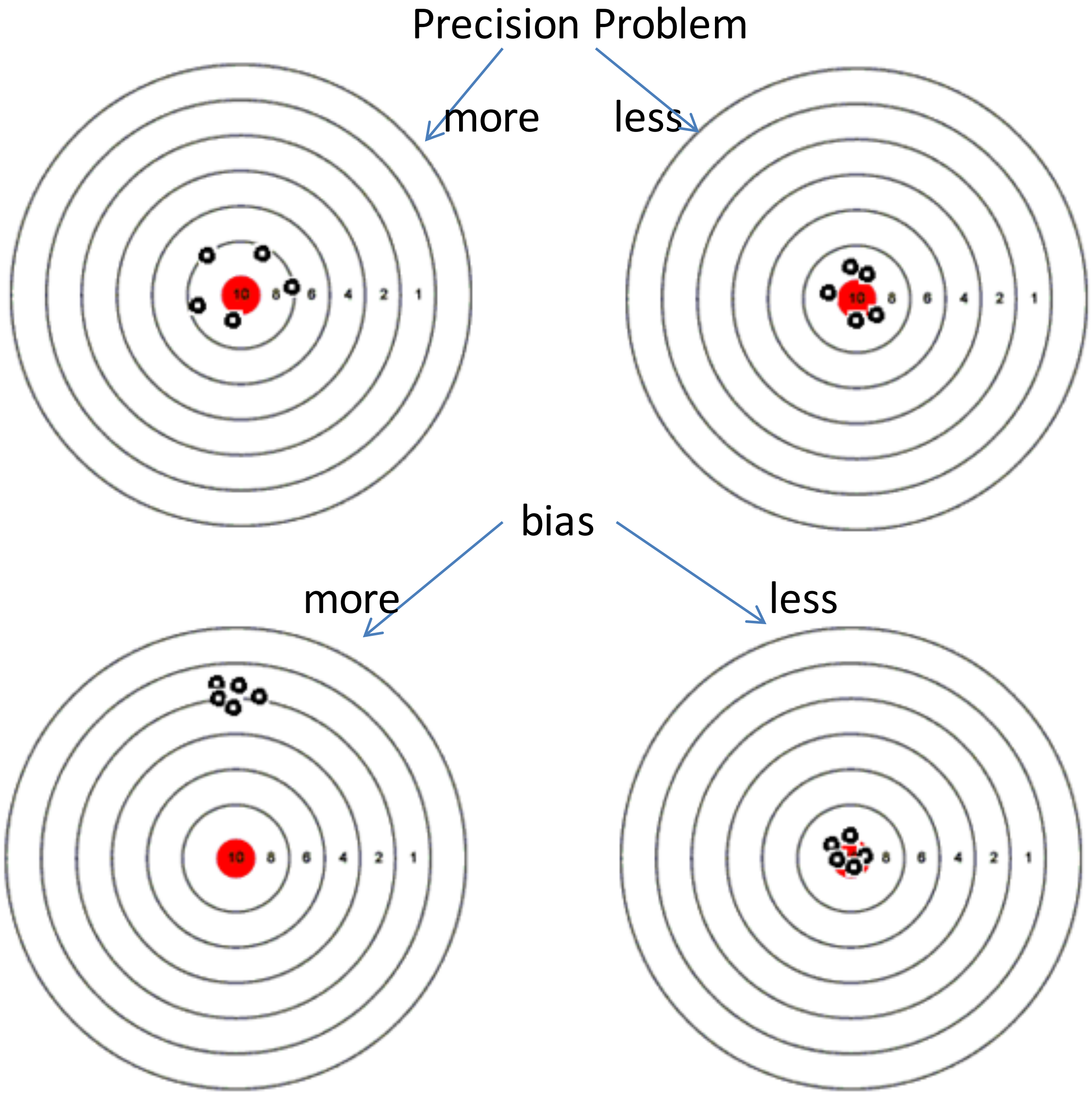
How to reduce bias in the estimates of count data regression?

Ashwini Joshi
Sumit Singh

PhUSE 2015, Vienna


Precision Problem
more
less
bias
more
less

Agenda

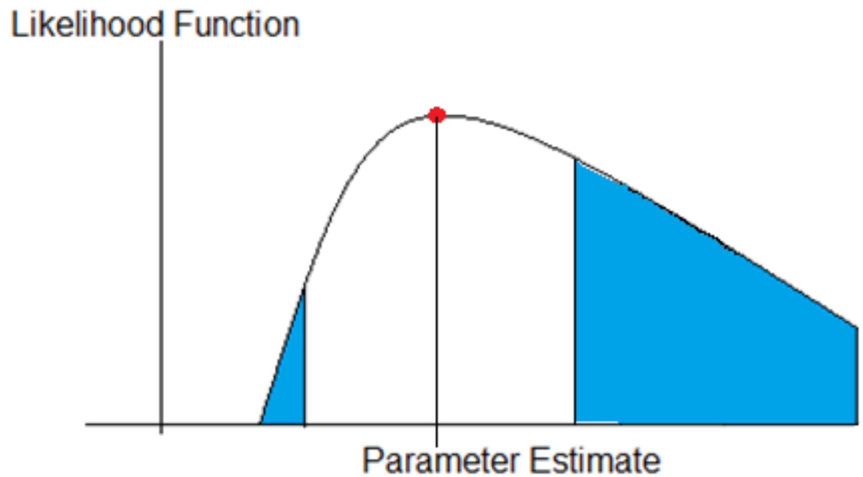
- Count Data
- Poisson Regression
- Maximum Likelihood Estimate and Bias Reduction Method
- Profile Likelihood Based Confidence Interval
- Negative Binomial Regression

Count Data in Clinical Trials

- Number of adverse events occurring during a follow up period
- Number of lesions or number of relapses in multiple sclerosis patients
- Number of hospitalizations
- Number of seizures in epileptics
- And many more…

Poisson Regression

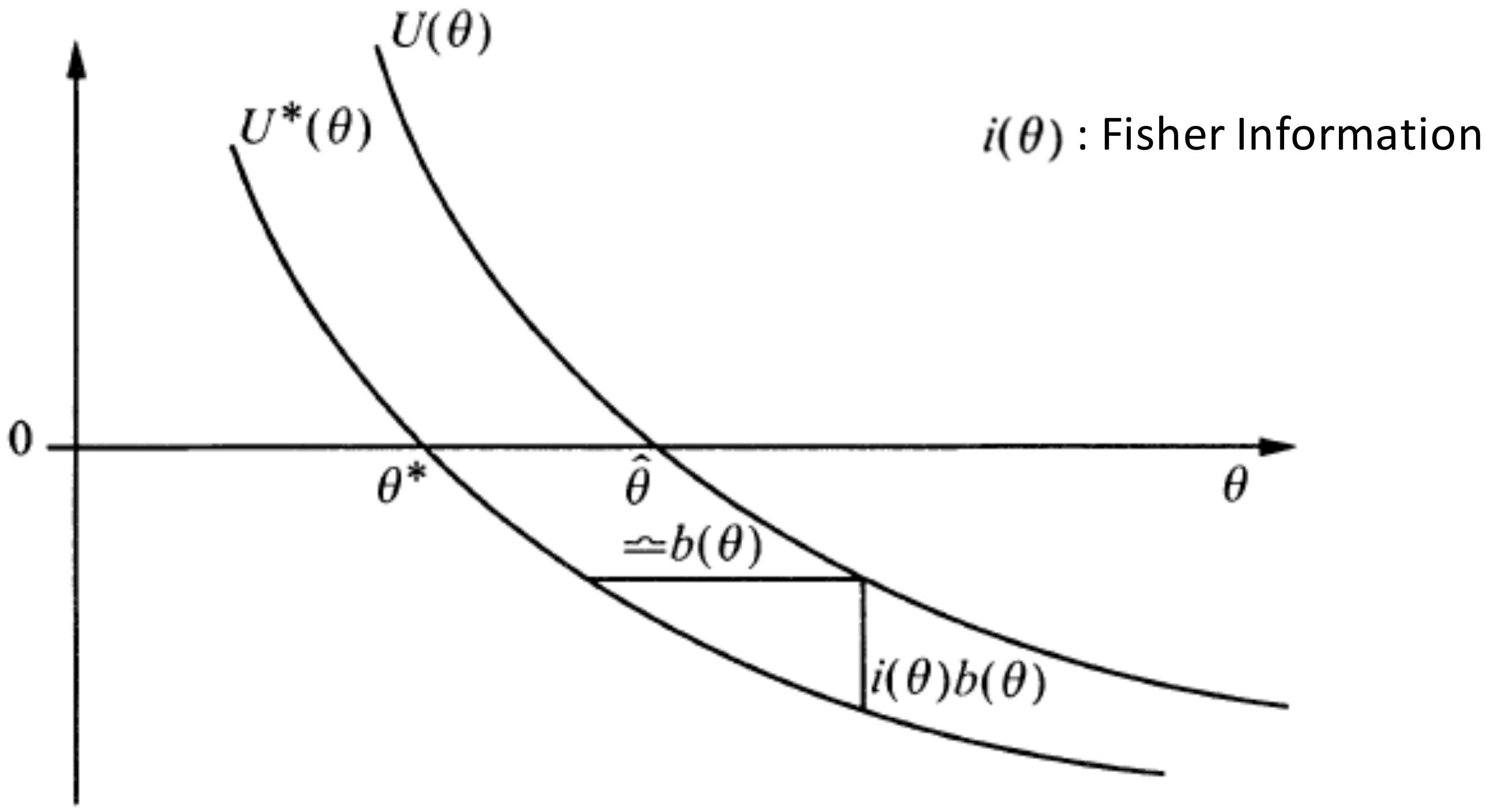
- Parameter Estimation
- Maximum Likelihood Estimate (MLE)
  - Bias in the case of small sample data
  - Asymmetric Wald Confidence Interval



Bias correction and Reduction Methods

- Bias Correction: MLE corrected for bias
- Bias Reduction: Likelihood function or score equations are modified
- Firth (1993): penalization of the likelihood or score function in opposite direction of bias

Firth Method



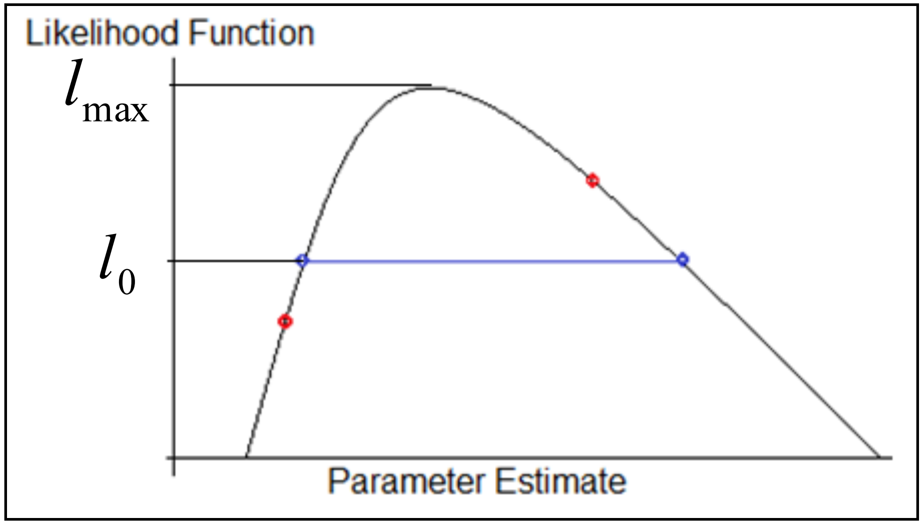
Modification of the unbiased score function.

Profile likelihood based Confidence Interval

- For Single Covariate

  Find CI limits where likelihood l_0 satisfy

$$l_0 = l_{\max} - 0.5\chi^2_{1-\alpha,1}$$



- In case of multiple covariates
  - Fix parameter for desired covariate
  - maximize the likelihood function over all other parameters
  - Check the above condition
  - Keep changing parameter value till above condition is satisfied

Scenarios

- Aspects of bias reduction method for small sample data
- Existence in the case of Quasi-separation
- Performance in case of large data

Example 1: Multiple Sclerosis

- Commonly used End Points:
  - Number of Relapses
  - Annualized relapse rate (ARR)

- Relapse is the appearance of a new neurological abnormality

Fit Poisson Model:

number of relapses ~ Baseline EDSS

EDSS: Expanded Disability Status Score

Example 1 (cont.):Proc LogXact

```
data MultScler;
input SubID EDSS Age SexCode RaceCode Arm Count Time;
cards;
1   1   19   2   1   1   0   758
2   1   54   1   2   2   0   57
.
.
.
31  2   53   2   1   1   0   751
32  2   46   1   1   0   1   449
;
PROC LOGXACT DATA=MultScler;
MODEL Count=  EDSS / link = Poisson;
ES/AS EDSS;
RATE Time;
RUN;

 PROC LOGXACT DATA=MultScler;
MODEL Count=  EDSS / link = Poisson;
ES/AS PMLE EDSS;
RATE Time;
RUN;
```

Example 1 (cont.): Output

```
The SAS System        16:15 Saturday, August 1, 2015   4

Output from LogXact (r) (v11.0) PROCs _SAS9_1
Copyright (c) 1997-2015 Cytel Inc., Cambridge, MA, USA.

------------------------------------------------------------
C O U N T   R E G R E S S I O N
------------------------------------------------------------
BASIC INFORMATION:
Data file:              MULTSCLER
Model:                  Count=Intercept+EDSS
Link type:              Poisson
Rate Multiplier:        Time
Stratum variable:       <Unstratified>
Analysis type:          Estimate :: Asymptotic
Number of terms in model: 2
Number of term(s) dropped: 0
Sum of Rate Multiplier:   20654
Number of records rejected: 0
Number of groups:       32
------------------------------------------------------------
SUMMARY STATISTICS:
------------------------------------------------------------
Statistic                  Value       DF      P-value
Deviance                   32.8859     30      0.3275
------------------------------------------------------------
PARAMETER ESTIMATE:
------------------------------------------------------------
           Point Estimate          Confidence Interval and P Value for Beta
Model Term  Type  Beta       SE        Type       95.0%     C.I.      P-Value
                                        Lower      Upper     2*1-sided
------------------------------------------------------------
Intercept   MLE  -5.2358    1.1002    Asymptotic  -7.3922   -3.0793    0.0000
EDSS        MLE  -1.5366    1.0260    Asymptotic  -3.5475    0.4743    0.1342
------------------------------------------------------------
Analysis Time = 00:00:00
```

Example 1 (cont.): Comparison

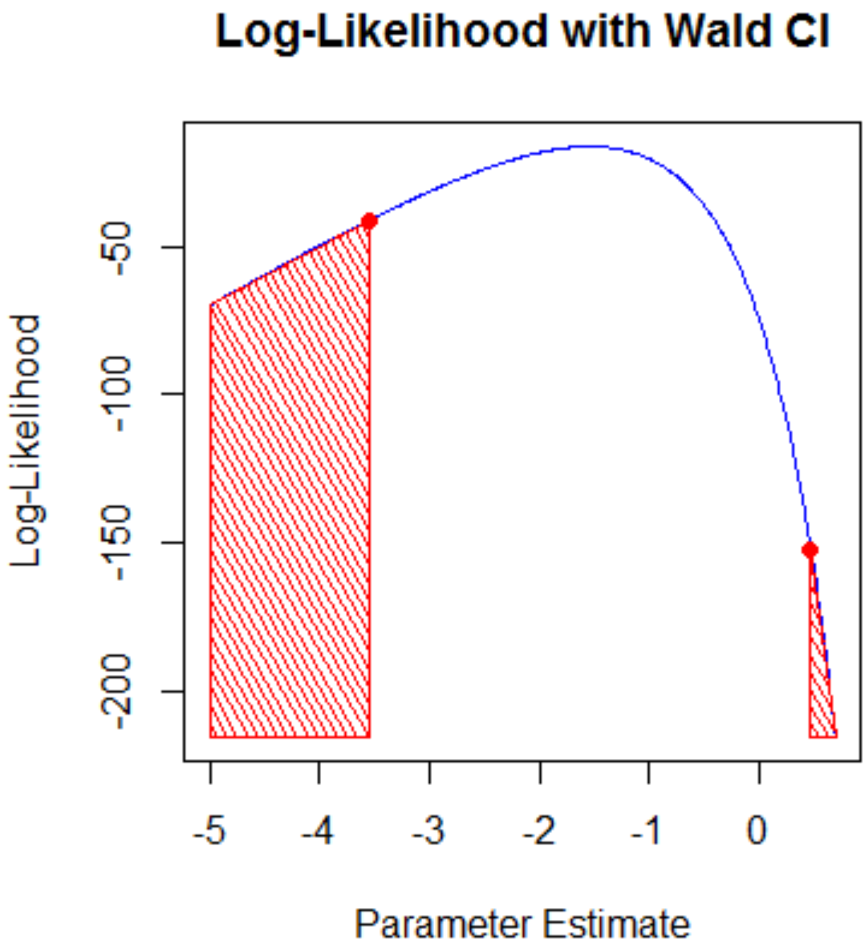
PARAMETER ESTIMATE:

Model Term	Point Estimate Type	Beta	SE	Type	Confidence Interval and P Value for Beta: 95.0% C.I. Lower	Upper	P-Value 2*1-sided
Intercept	MLE	-5.23	1.10	Asymptotic	-7.3922	-3.0793	0.0000
EDSS	MLE	-1.54	1.03	Asymptotic	-3.5475	0.4743	0.1342

PARAMETER ESTIMATE:

Model Term	Type	Beta	Point Estimate SE	Type	Confidence Interval and P Value for Beta: 95.0% C.I. Lower	Upper	P-Value 2*1-sided
Intercept	PMLE	-5.58	0.93	Asymptotic	-7.4193	-3.7593	2.706e-009
EDSS	PMLE	-1.16	0.85	Asymptotic	-2.8178	0.5036	0.1721

Example 1 (cont.)



Separation, Quasi-separation

- Problem of Separation (Perfect Fit) occurs when
  - Response value divides set of covariate or combinations of covariates
  - A covariate value predicts value of the response

Age	Gender	Events
old	Male	1
old	Male	1
old	Female	1
young	Female	0
young	Female	0
young	Female	0

For Age<60, Events =0:

Separation

For Gender: Quasi-separation

Example 2

Animal Data: Data from Heinze and Puhr (2010).

Study: To investigate effect of heparinized vs non-heparinized, vascular substitute in rats on aneurysm formation.

The **event** of interest is aneurysm formation.

The **covariate** is non-heparinized implant.

Stratum is follow up time.

- Model: Aneur ~ Nohep
- Case of Quasi-separation
- MLE can not be obtained

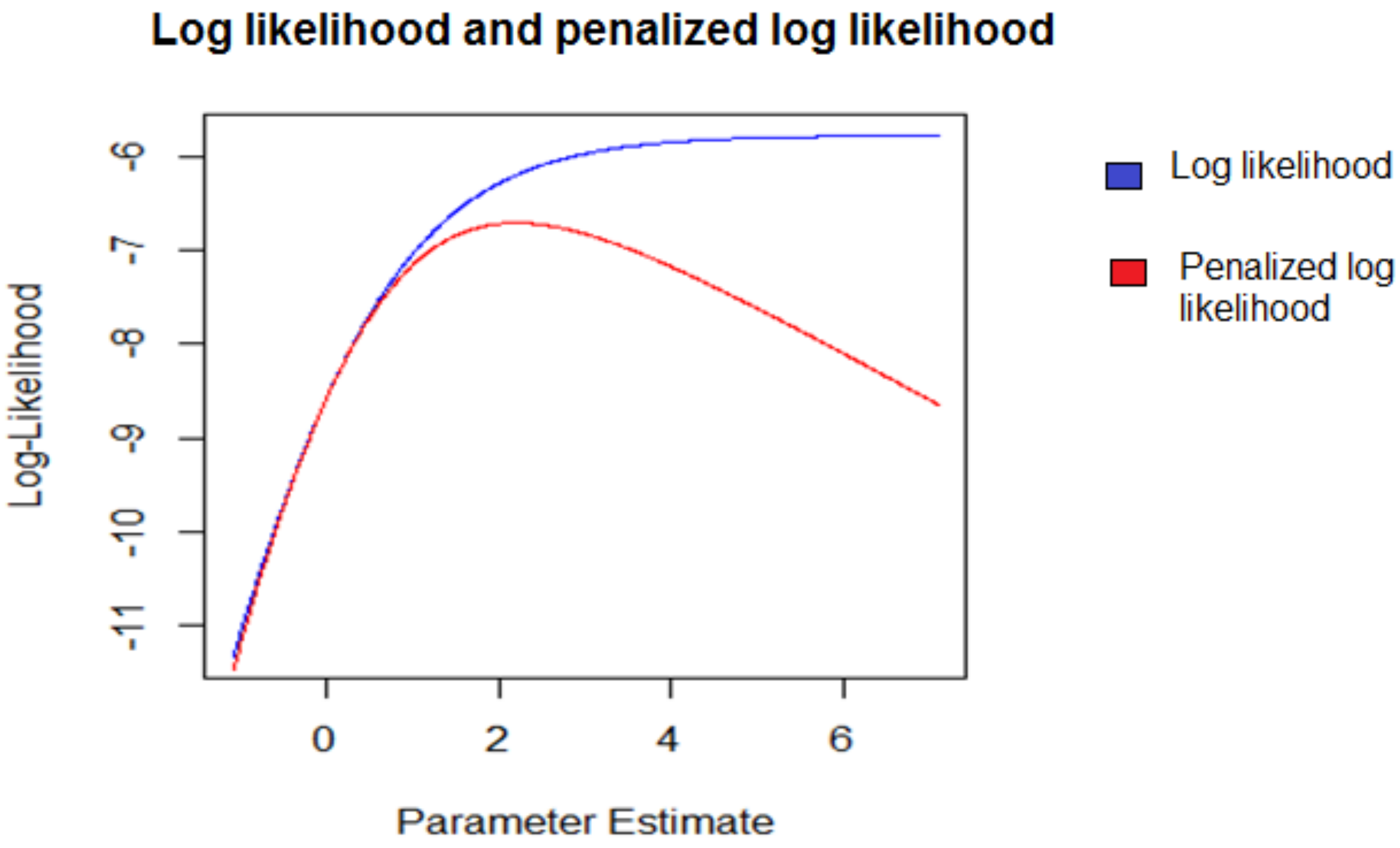
Example 2 (cont.)

PARAMETER ESTIMATE:

Model Term	Type	Point Estimate Beta	SE
NoHep	MLE	?	?
NoHep	PMLE	2.1970	1.6665

? : non-convergence

Example 2 (cont.)



Negative Binomial Regression

- Poisson: Variance = mean
- Over dispersion in Poisson: Variance > mean
- NB-2

  variance = $\mu + a.\ \mu^2$

Example 3 – Medpar Data

- **Data:** From the US national Medicare inpatient hospital database (Medpar) for the 1991 Medicare files for the state of Arizona (Hilbe 2011).
- **RESPONSE**

  los: length of stay in the hospital

- **PREDICTORS**

  hmo: Patient belongs to a Health Maintenance Organization(1), or private pay(0)

  white: Patient identifies themselves as primarily Caucasian(1) in comparison to non-white(0)

  type: A three-level factor predictor related to the type of admission. 1=elective(referent), 2=urgent, 3=emergency

- Fit **NB-2** model: **los ~ factor(type)**

Example 3 (cont.)

PARAMETER ESTIMATE:

		Point Estimate		Confidence Interval and P Value for Beta			
Model Term	Type	Beta	SE	Type	95.0% C.I. Lower	Upper	P-Value 2*1-sided
Intercept	MLE	2.1783	0.0222	Asymptotic	2.1347	2.2219	0.0000
type_2	MLE	0.2379	0.0502	Asymptotic	0.1394	0.3363	0.0000
type_3	MLE	0.7252	0.0757	Asymptotic	0.5768	0.8736	2.04e-019
Intercept	PMLE	2.1781	0.0222	Asymptotic	2.1345	2.2217	0.0000
type_2	PMLE	0.2368	0.0502	Asymptotic	0.1384	0.3352	0.0000
type_3	PMLE	0.7235	0.0756	Asymptotic	0.5752	0.8717	2.278e-019

Concluding remarks

- Properties of Bias reduced Estimates for small sample data:
  - Smaller standard errors
  - Shorter confidence intervals
  - Existence in the case of quasi-separation

- Profile Likelihood based CI
  - better than Wald CI for asymmetric likelihood function

References

- Firth D. (1993). “Bias reduction of maximum likelihood estimates”. Biometrika (1993), 80, 1, pp. 27-38
- Heinze G., Puhr R.(2010). “Bias-reduced and separation-proof conditional logistic regression with small or sparse data sets”. Statistics in Medicine
- Hilbe J.(2011). “ Negative Binomial Regression”. Cambridge University Press
- Kosmidis I. (2007). “Bias Reduction in Exponential Family Nonlinear Models”.
- Kosmidis I.(2010). “A generic algorithm for reducing bias in parametric estimation”. Electronic Journal of Statistics. Vol. 4 (2010) 1097–1112
- LogXact 11 Software and User Manual

Thank you

Ashwini.Joshi@cytel.com
Sumit.Singh@cytel.com