Regression Analysis of Competing Risks Data Using Pseudo-Observations.

Per Kragh Andersen

Department of Biostatistics, University of Copenhagen;

&

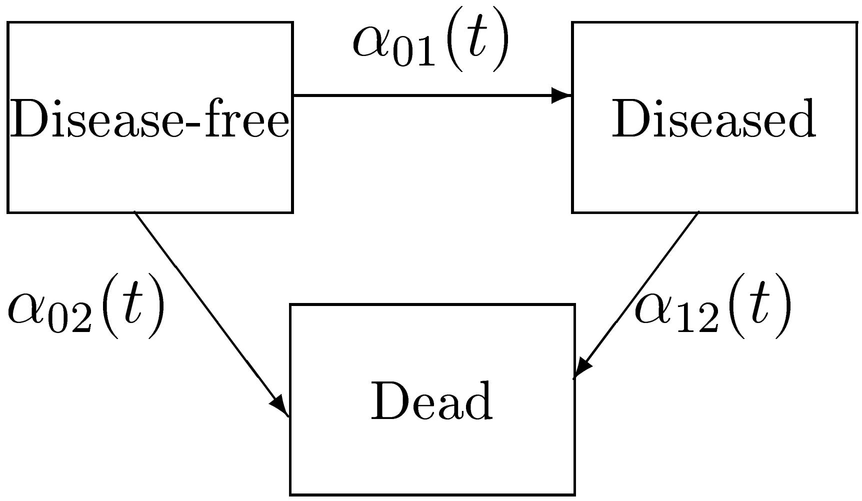
Danish Epidemiology Science Centre.

Joint work with:

John P. Klein, Milwaukee, USA,

(*Biometrika*, **90**, 2003, 15-27.)

Multi-state models.



Regression models may be formulated via the transition intensities, e.g.:

$$\alpha_{hj}(t \mid Z) = \alpha_{hj0}(t) \exp(\beta_{hj}^{\mathsf{T}} Z).$$

or

$$\alpha_{hj}(t \mid Z) = \alpha_{hj0}(t) + \beta_{hj}^{\mathsf{T}} Z.$$

A BMT example.

- 1441 patients bone marrow transplanted for acute leukemia
- 1066 AML ($Z = 0$), 375 ALL ($Z = 1$)
- Median age 27.7 ys., range 0.4-57
- 272 died/relapsed without GVHD
- 413 developed GVHD, 195 died/relapsed subsequently

Cox models for intensities:

	Age		Disease	
Transition	β	SD	β	SD
$0 \to 2$	0.012	0.0050	0.328	0.131
$0 \to 1$	0.016	0.0040	-0.050	0.116
$1 \to 2$	0.015	0.0066	0.386	0.162

Transition or state probabilities.

In Markov model:

$$P_{01}(0,t) = \int_0^t P_{00}(0,u-)\alpha_{01}(u)P_{11}(u+,t)\mathrm{d}u,$$

$$P_{00}(0,t) = \exp(-\int_0^t (\alpha_{01}(u) + \alpha_{02}(u))\mathrm{d}u),$$

$$P_{11}(0,t) = \exp(-\int_0^t \alpha_{12}(u)\mathrm{d}u).$$

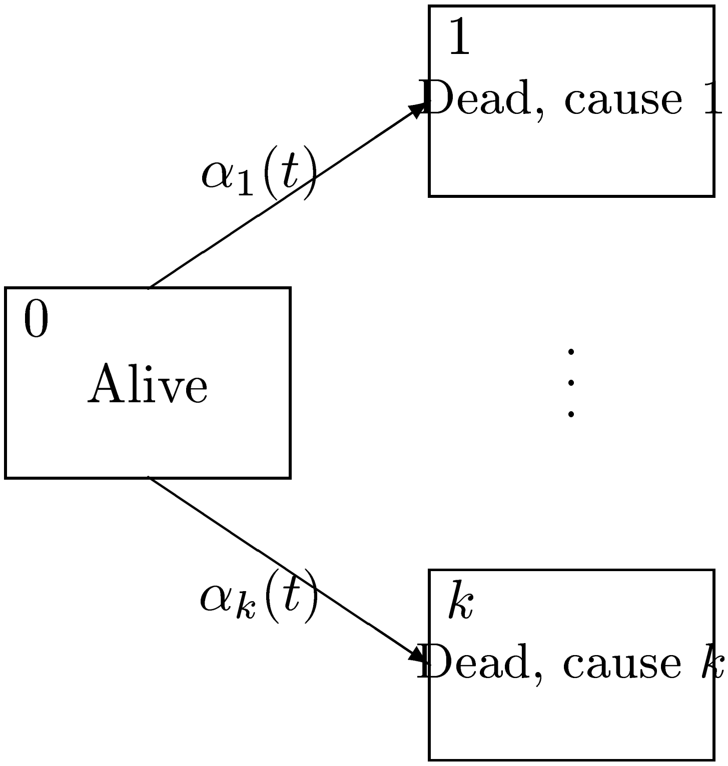
Intractable function of covariates.

<u>Aim:</u>

Direct (simple) regression models for $P_{hj}(0, t \mid Z)$, e.g.,

$$\mathrm{logit} P_{hj}(0, t \mid Z) = \alpha_t + \beta^{\mathsf{T}} Z.$$

Competing risks.



<u>Survival probability:</u>

$$S(t) = P_{00}(0,t) = \exp(-\int_0^t \sum_{h=1}^{k} \alpha_h(u)\mathrm{d}u).$$

<u>Cumulative incidences:</u>

$$P_{0h}(0,t) = \int_0^t S(u-)\alpha_h(u)\mathrm{d}u.$$

Example: Hoel and Walburg, 1972.

- 177 mice radiated at 5-6 weeks
- 95 in conventional, 82 in germ-free environment
- 3 causes of death: Thymic Lymphoma (TL), Reticulum Cell Sarcoma (RCS), Other Causes (OC)

Cox models for cause-specific hazards: conventional vs. germ-free

- TL $\widehat{\beta}$ =-0.323 (0.286)
- RCS $\widehat{\beta}$ =2.004 (0.345)
- OC $\widehat{\beta}$ =1.080 (0.304)

Effect of covariates on cumulative incidences?

General formulation.

- $X_1, \ldots, X_n$ i.i.d.
- $\theta = \mathrm{E} f(X_i)$ for some $f(\cdot)$
- $\widehat{\theta}$ unbiased estimator
- Examples:
  - Gaussian with common variance, $f(\cdot) = id(\cdot), \widehat{\theta} = \bar{X}$
  - Bernoulli, $f(\cdot) = id(\cdot), \widehat{\theta} = \bar{X}$
  - Censored life times, $f(X) = I(X > t), \widehat{\theta} =$ Kaplan-Meier estimator $\widehat{S}(t)$
  - Multi-state process (e.g. competing risks), $f(X) = I(X(t) = h)$, $\widehat{\theta} =$ Aalen-Johansen or Pepe estimator $\widehat{P}_{0h}(0, t)$

Inhomogeneous population due to covariates

$$Z_1, \ldots, Z_n \ \text{ i.i.d.} \approx Q.$$

We have

$$\mathrm{E}(f(X_i)) = \mathrm{E}(\mathrm{E}(f(X_i) \mid Z_i))$$

and $\widehat{\theta}$ estimates the marginal mean

$$\theta = \int_{\mathcal{Z}} \mathrm{E}(f(X_i) \mid Z_i) \mathrm{d}Q(Z_i).$$

Idea: Approximate Q by empirical distribution, let $\theta_i(Z_i) = \mathrm{E}(f(X_i) \mid Z_i)$.

Then $\widehat{\theta}(X)$ "estimates the same as"

$$\widetilde{\theta}(Z) = \frac{1}{n} \sum_j \theta_j(Z_j)$$

and $\widehat{\theta}_{-i}$ (based on $X_j, j \neq i$) "estimates the same as"

$$\frac{1}{n-1} \sum_{j \neq i} \theta_j(Z_j).$$

That is:

$$\widehat{\theta}_i := \widehat{\theta} + (n-1)(\widehat{\theta} - \widehat{\theta}_{-i})$$
$$= n\widehat{\theta} - (n-1)\widehat{\theta}_{-i}$$

"estimates the same as" $\theta_i(Z_i)$.

More precisely:

$$\mathrm{E}_P(\widehat{\theta}(X)) = \mathrm{E}_P(\widetilde{\theta}(Z))$$

and

$$\mathrm{E}_P(\widehat{\theta}_i) = \mathrm{E}_P(\theta_i(Z_i))$$

where P is the distribution of (X, Z).

The $\widehat{\theta}_i$ are pseudo-observations known from jackknife methods.

Model for $\theta_i(z_i) = \mathrm{E}(f(X_i) \mid Z_i = z_i)$:

Generalized linear model with link $g(\cdot)$:

$$g(\theta_i) = \beta^\mathsf{T} z_i.$$

$(z_{i0} = 1)$

Examples:

- Gaussian with common variance, $f(\cdot) = id(\cdot), \widehat{\theta} = \bar{X}, g(\cdot) = id(\cdot)$ ordinary linear regression, (in fact $\widehat{\theta}_i = X_i$)
- Bernoulli, $f(\cdot) = id(\cdot), \widehat{\theta} = \bar{X}, g(\cdot) =$ logit, logistic regression, (again $\widehat{\theta}_i = X_i$)
- Life times, $f(X) = I(X > t), \widehat{\theta} = \widehat{S}(t)$, $g(x) = \log(-\log(x))$ =cloglog(x), Cox regression for fixed t
- Multi-state process, $f(X) = I(X(t) = h), \widehat{\theta} = \widehat{P}_{0h}(0, t)$, $g(\cdot)$ =logit, "logistic regression" of state h probability for fixed t, or $g(\cdot)$ =cloglog.
- etc.

Multivariate problems.

Cox example: consider time points $t_1, \ldots, t_k$:

$\text{cloglog}(S_i(t_l)) = \log A_0(t_l) + \beta^\mathsf{T} z_i = \beta^\mathsf{T} z_{il}$.

Multi-state example: consider time points $t_1, \ldots, t_k$:

$\text{logit}(P_{0h,i}(t_l)) = a_l + \beta^\mathsf{T} z_i = \beta^\mathsf{T} z_{il}$.

Define pseudo-observations:

$$\widehat{\theta}_{il} = \widehat{\theta}(t_l) + (n-1)(\widehat{\theta}(t_l) - \widehat{\theta}_{-i}(t_l)),$$

$i = 1, \ldots, n, l = 1, \ldots, k$.

Estimation.

Regression of $(\widehat{\theta}_{il})$ on (z_{il}) with link function $g(\cdot)$, i.e. solve unbiased estimating equations:

$$\sum_i \left(\frac{\partial}{\partial \beta} g^{-1}(\beta^\mathsf{T} z_i)\right)^\mathsf{T} V_i^{-1}(\widehat{\theta}_i - g^{-1}(\beta^\mathsf{T} z_i))$$

$$= \sum_i U_i(\beta) = U(\beta) = 0.$$

(Use, e.g. SAS PROC GENMOD, $V_i =$ working covariance matrix for $\widehat{\theta}_i$.)

Variance estimates:

let $I(\beta) = \sum_i (\frac{\partial g^{-1}(\cdot)}{\partial \beta})^\mathsf{T} V_i^{-1} (\frac{\partial g^{-1}(\cdot)}{\partial \beta})$ then

$$\widehat{\mathrm{Var}}(\widehat{\beta}) \approx I(\widehat{\beta})^{-1} \widehat{\mathrm{Var}}(U(\beta))(I(\widehat{\beta})^{-1})^\mathsf{T},$$

$$\widehat{\mathrm{Var}}(U(\beta)) = \sum_i U_i(\widehat{\beta})^\mathsf{T} U_i(\widehat{\beta}),$$

i.e., standard GEE sandwich estimator.

Alternative: non-parametric bootstrap - resample over $i = 1, \ldots, n$.

Large sample properties.

Solution $\widehat{\beta}$ as. normal if:

- (1) $\mathrm{E}U(\beta_0) = 0$
- (2) $U_i(\beta); i = 1, \ldots, n$ independent
- (3) $\widehat{\mathrm{Var}}(\widehat{\beta}) \to_P$ ...

Liang and Zeger (1986)

- (1) above
- (2) old result (Tukey, 1958)
- (3) standard assumption

<u>Choice of V_i:</u>

$V_i \approx \mathrm{var}(\widehat{\theta}_i)$

<u>Choice of time points?</u>

Simulations.

Monte Carlo results available for:

- Logistic regression
- Cox regression model, 10 time points
- 3-state model
- competing risks, 5, 10, 15, 20 time points, various working correlations
- mean life time
- restricted mean life time

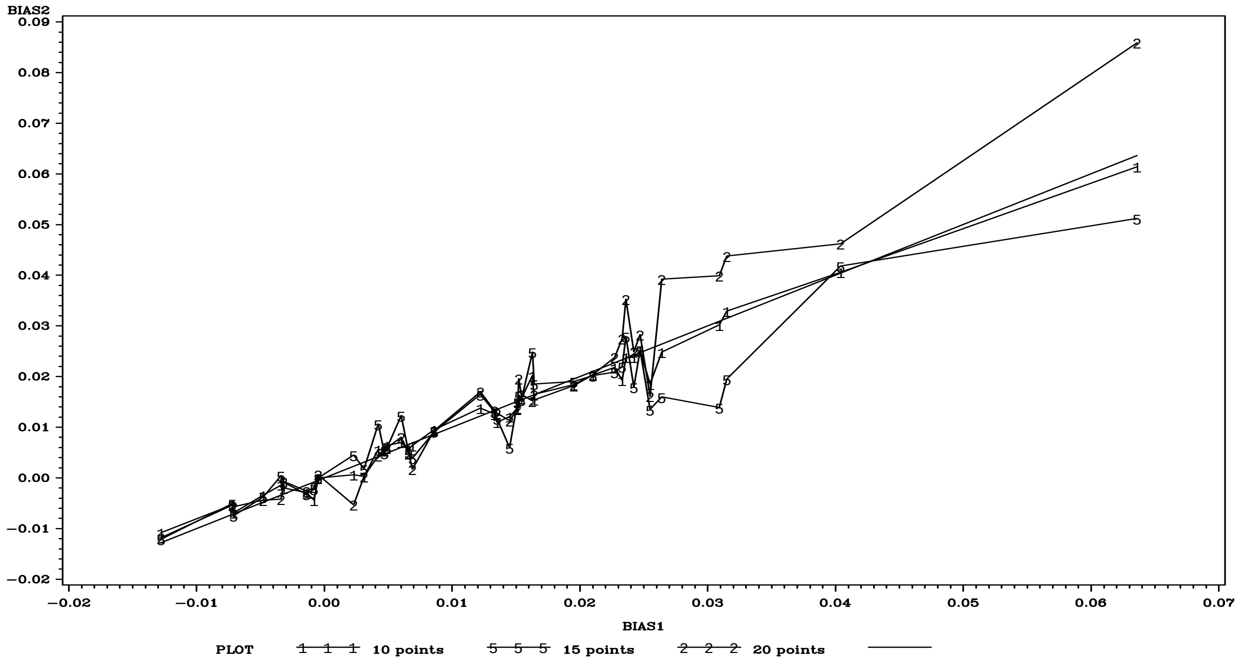
Here:

- competing risks

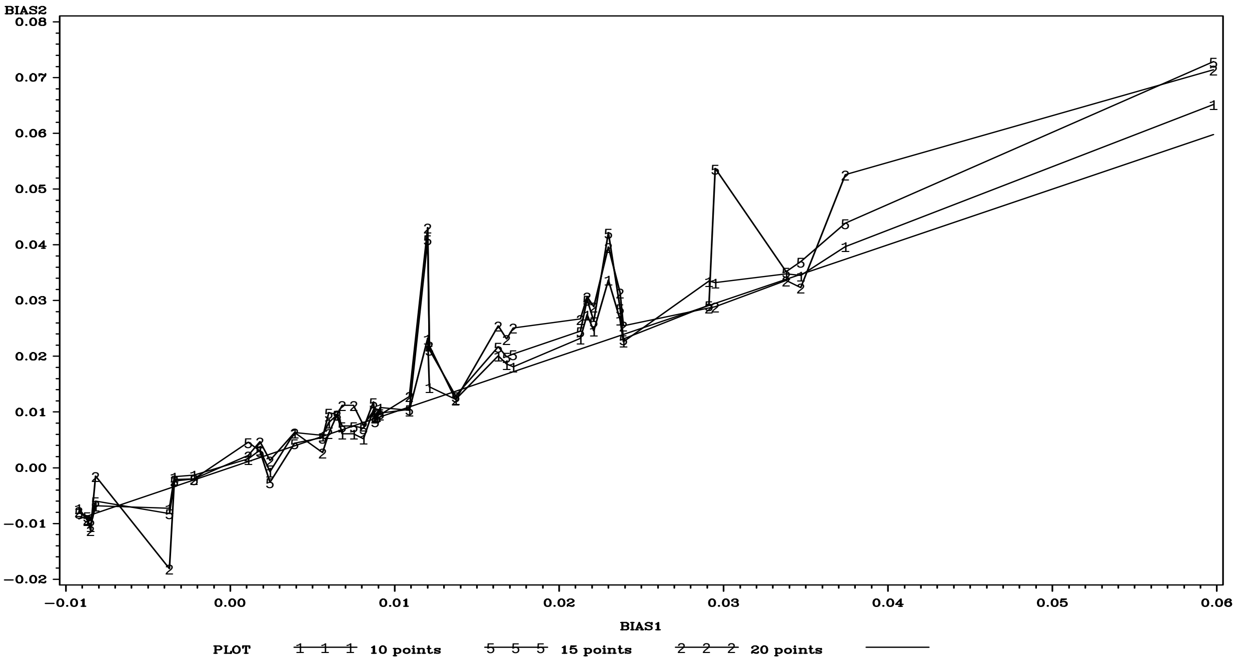
Simulations of competing risks model.

- $k = 2, n = 200, 1000$ repetitions
- logistic link:
  - $P_1(t) = \frac{p(1-e^{-t})e^{\beta Z}}{1-p+p(1-e^{-t})e^{\beta Z}}$,
  - $P_2(t) = \frac{(1-p)(1-e^{-t})}{1-p+pe^{\beta Z}}$
- clogclog link:
  - $P_1(t) = 1 - (1 - p(1 - e^{-t}))^{\exp(\beta Z)}$,
  - $P_2(t) = (1 - p)^{\exp(\beta Z)} \exp(-t \exp(\beta Z))$
- 25% or 50% (uniform) censoring, $p = 0.25, 0.50$, $\beta = 0, 0.4, 0.7, 1$, NO. of time points: 5, 10, 15 or 20
- 3 working correlations:
  - independence,
  - $\mathrm{cov}(\hat{\theta}_{il}, \hat{\theta}_{im}) = P(t_l \wedge t_m)(1 - P(t_l \vee t_m))$,
  - "empirical"

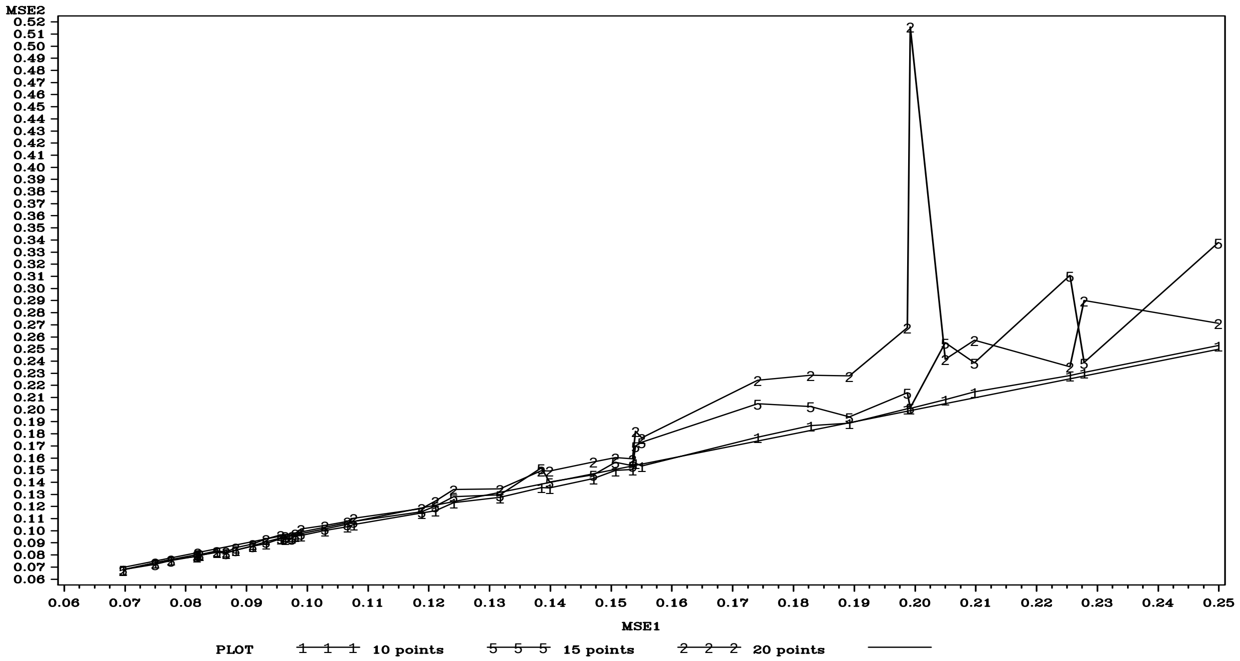
Bias for 10, 15, 20 pseudo−values plotted against 5: logit



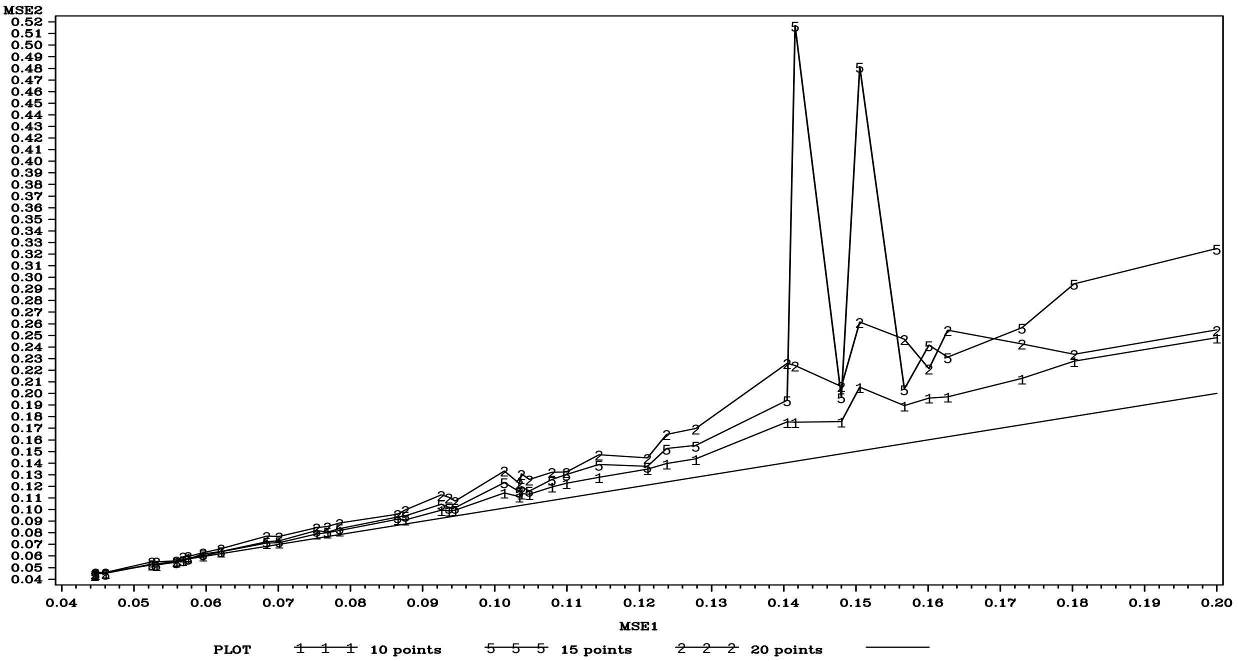
Bias for 10, 15, 20 pseudo−values plotted against 5: cloglog

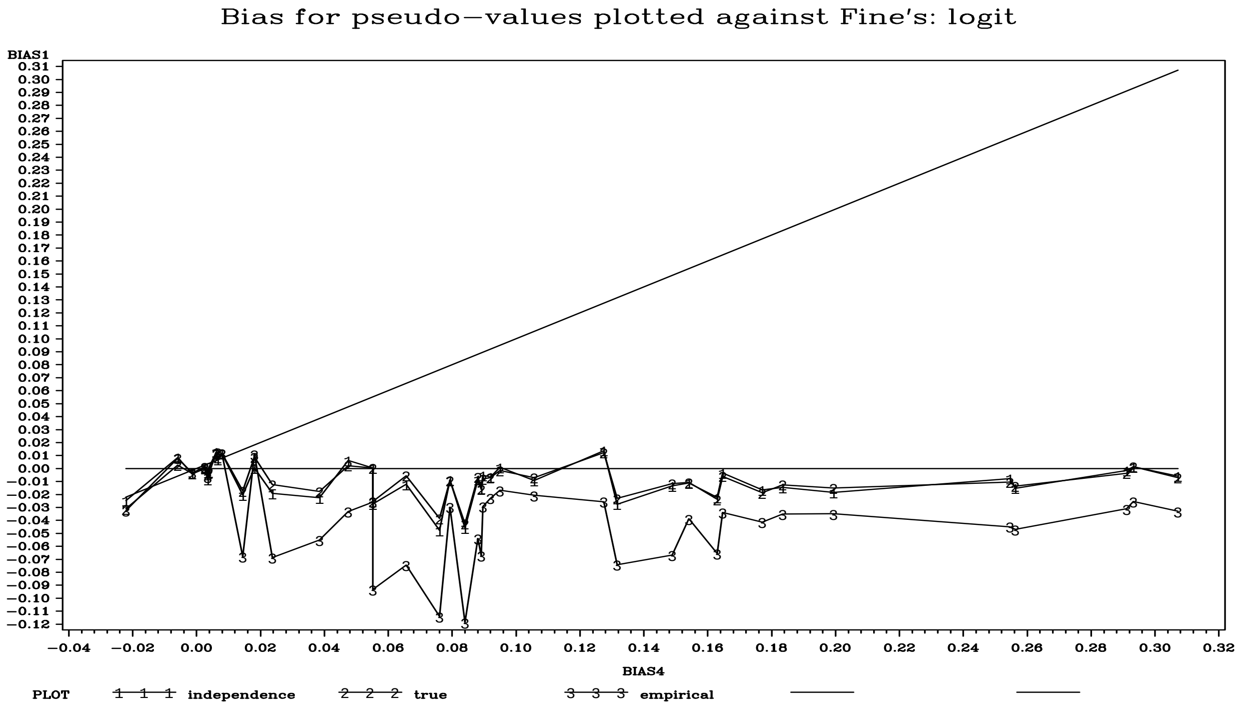


MSE for 10, 15, 20 pseudo−values plotted against 5: logit

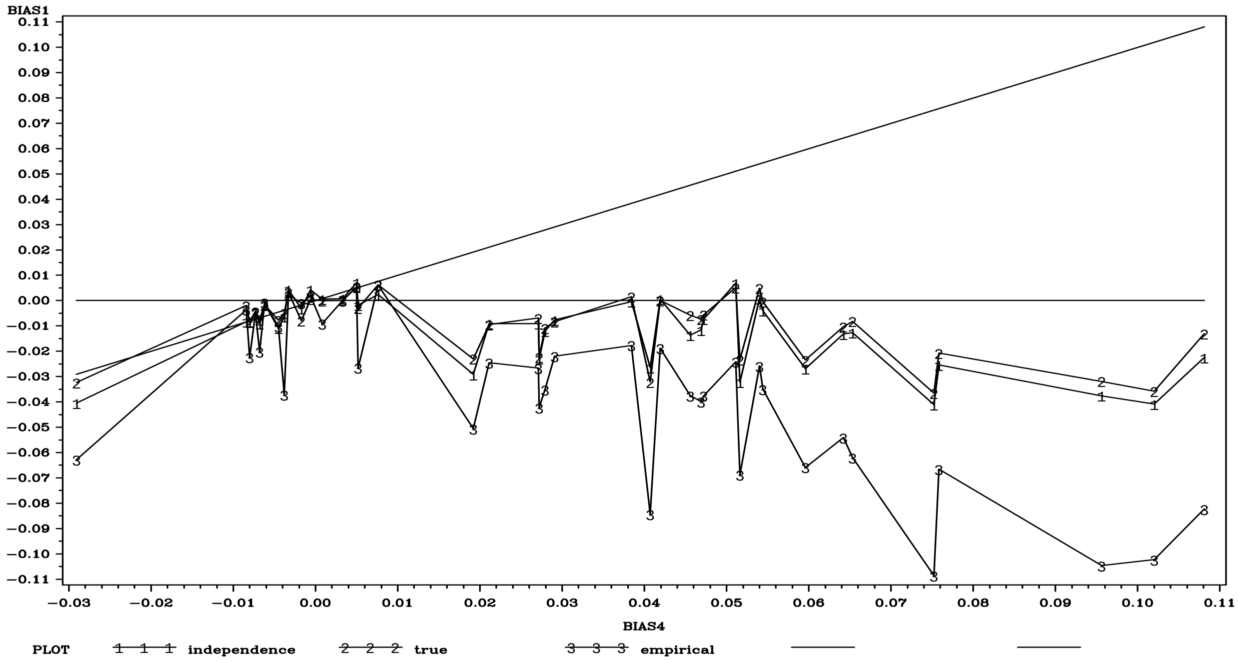


MSE for 10, 15, 20 pseudo−values plotted against 5: cloglog

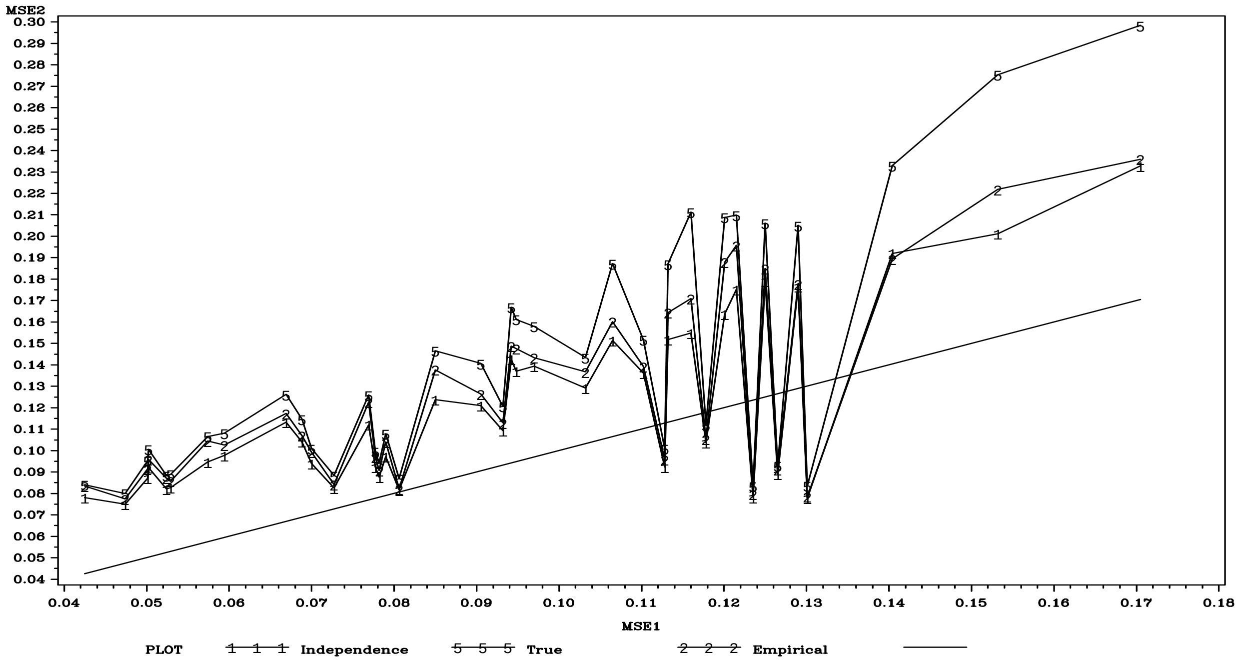



Bias for pseudo−values plotted against Fine's: logit
BIAS1
BIAS4
PLOT
independence
true
empirical

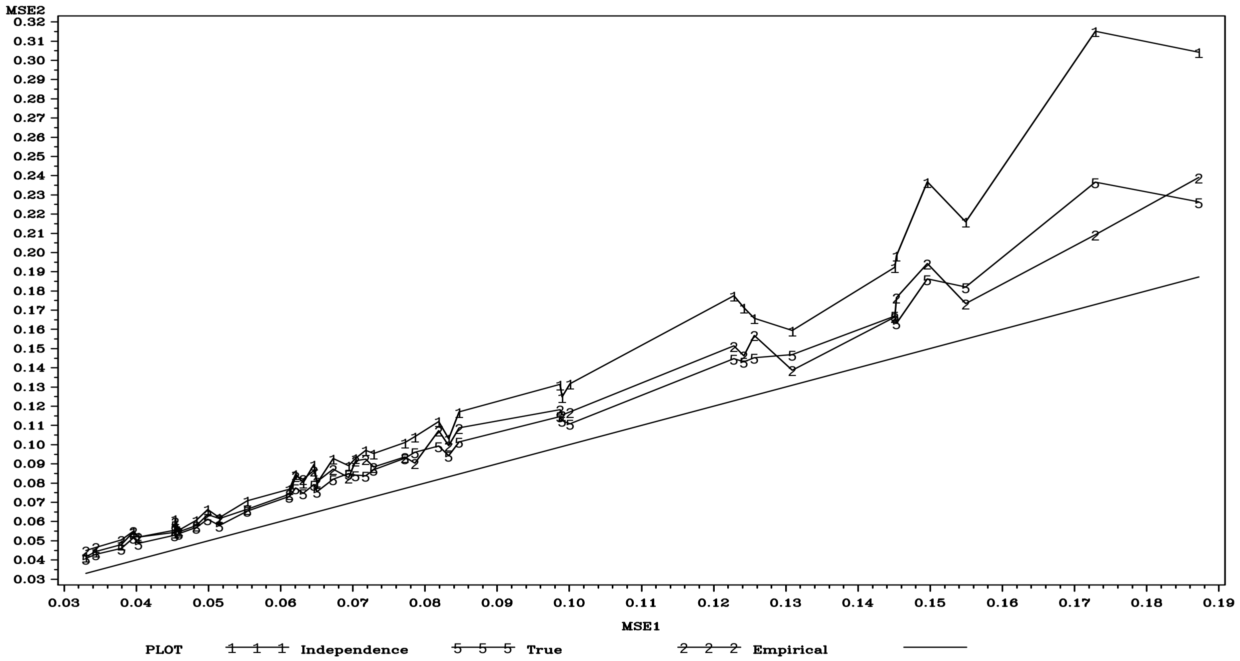
Bias for pseudo−values plotted against Fine's: cloglog



MSE for pseudo−values plotted against Fine's: logit



MSE for pseudo−values plotted against Fine's: cloglog



BMT example: results.

	5 Time Points		10 Time Points	
	β	SD	β	SD
	Q_0			
Disease	-0.174	0.127	-0.160	0.115
Age	-0.019	0.005	-0.020	0.004
	Q_1			
Disease	-0.255	0.152	-0.226	0.144
Age	0.009	0.005	0.010	0.005
	Q_2			
Disease	0.345	0.154	0.417	0.152
Age	0.007	0.006	0.008	0.006
	Q_3			
Disease	0.297	0.179	0.320	0.179
Age	0.027	0.007	0.028	0.007

Hoel and Walburg.

Regression analysis of "cumulative incidence functions" (Fine and Gray, JASA, 1999): cloglog link

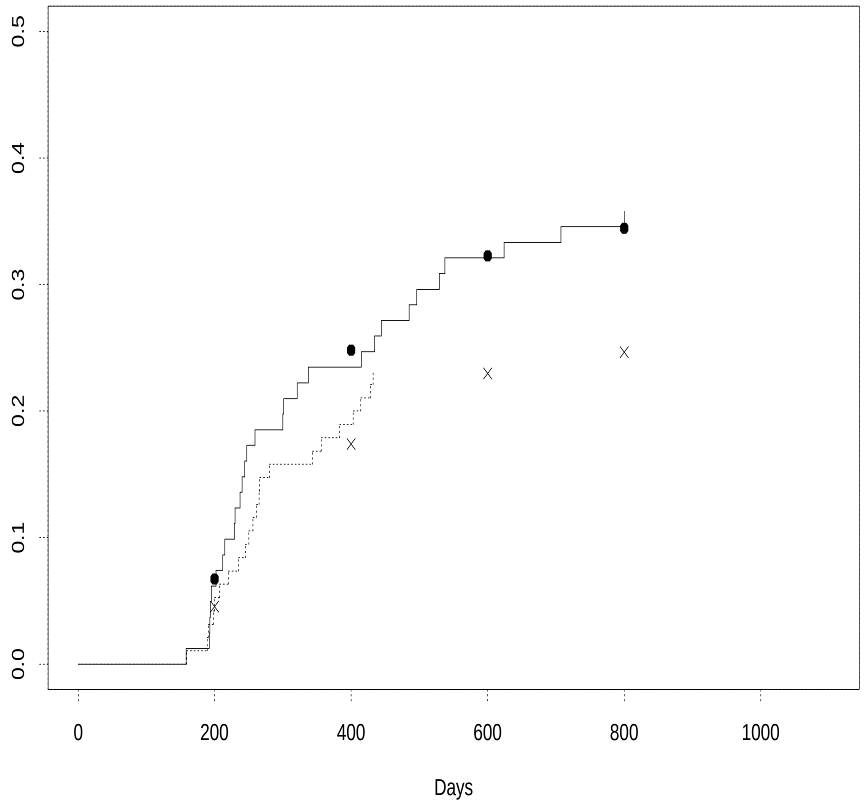
Pseudo-observations computed at $t = 200, 400, 600, 800$ days

Cause of death	TL	RCS	OC
Fine and Gray $\hat{\beta}$	-0.487	0.975	-0.090
Fine and Gray (SE)	(0.283)	(0.305)	(0.236)
Pseudo-obs. $\hat{\beta}$	-0.401	1.151	0.659
Pseudo-obs. (SE)	(0.286)	(0.321)	(0.276)

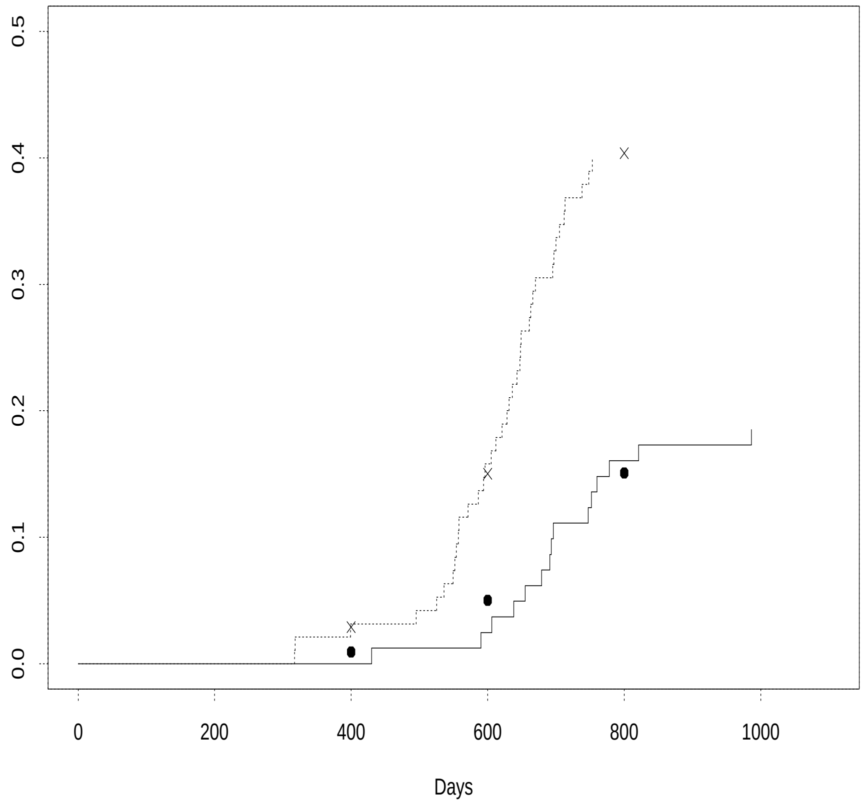
From pseudo-obs. we also get correlations:

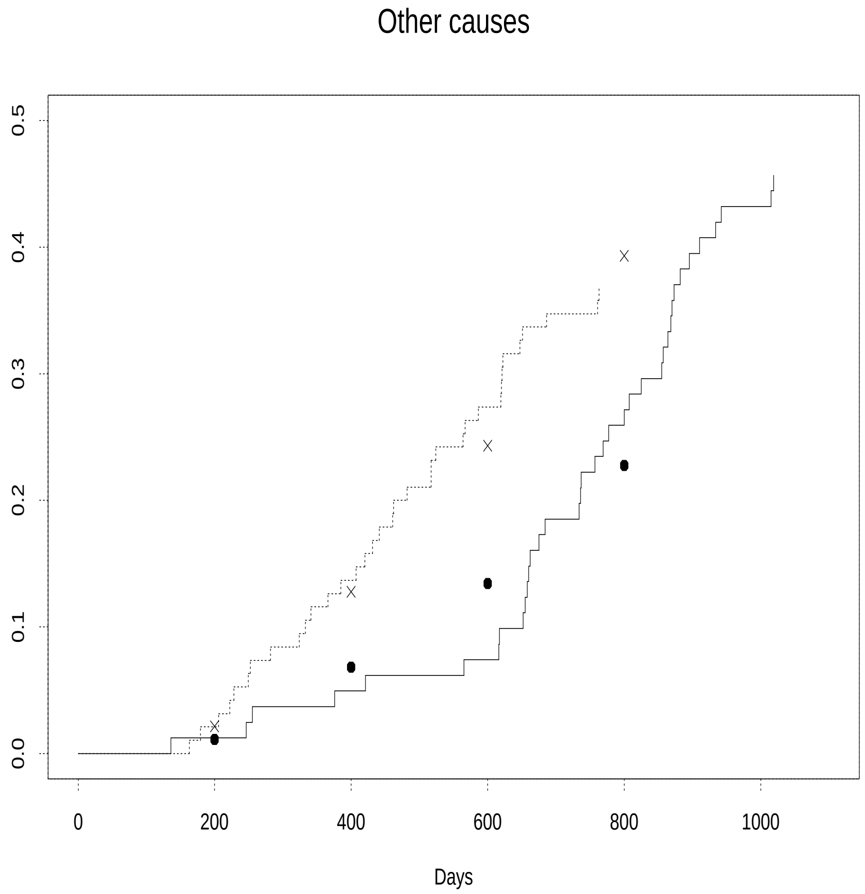
- (TL,RCS): -0.34
- (TL,OC) -0.40
- (RCS,OC) -0.35

Thymic Lymphoma



Recticulum cell sarcoma





Other causes
0.0
0.1
0.2
0.3
0.4
0.5
0
200
400
600
800
1000
Days

Conclusions.

- Pseudo-observations provide a general approach to a number of non-standard regression problems - including competing risks
- Future studies:
  - Regression of interval-censored data
  - Regression of variances and correlations
  - Goodness of fit examinations (link function), pseudo-residuals