Partially Missing Data: Right Censoring

Censoring machanism

- Censoring is a common mechanism to result in missing data for time-to-event outcome.
- One of the most common censoring is called right-censoring, where the event is known to be larger than an observed time for whoever drops out, e.g., time-to-death (cancer survival), time-to-relapse.
- Censored data can be considered as partially missing data: the missing values are not completely unknown, but are known partially.
- Survival analysis is a well-developed research area for analyzing censored data.

Objectives in survival analysis

- Prediction: what is the chance that a cancer patient can survive more than 10 years?
- Comparison: does treatment A improve survival in patients with liver cancer than treatment B?
- Association/causaility: are patients who ever smoke more likely to develop lung cancer than non-smokers?

Example: prediction and comparison

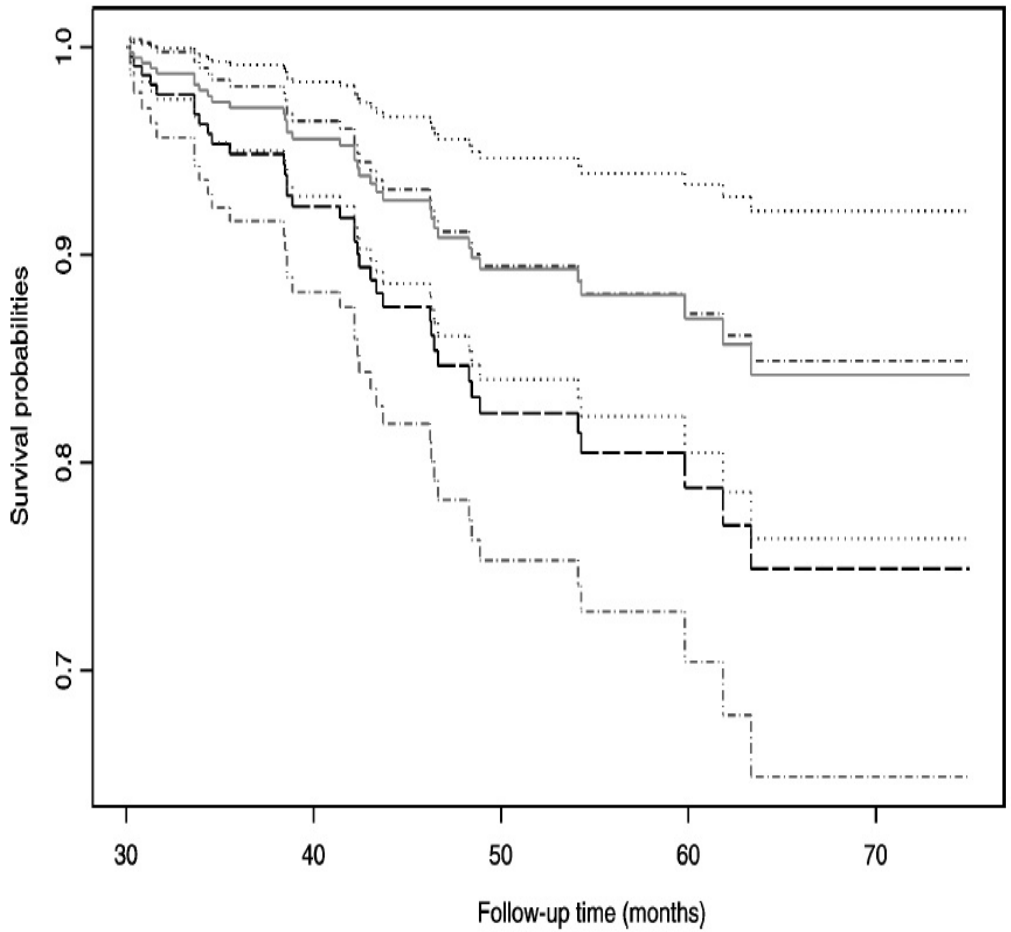

Figure 1. Estimated Conditional Survival Probabilities for the Diabetic Retinopathy Patients. The "——" curve represents the point estimate of the survival function for the adult-onset diabetics; the "·····" curves represent the corresponding 95% confidence limits; the "– – –" curve represents the point estimate of the survival function for the juvenile-onset diabetics; and the "– · – · –" curves represent the corresponding 95% confidence limits.

Why time-to-event outcome should be treated differently?

- It is positive.
- Its distribution is often skewed.
- However, the most unique is due to censoring!

What is censoring?

- Let's follow one patient till he dies (time-to-death)...
  - Case 1: patient dies at year 15;
  - Case 2: patient drops out of the study at year 10;
  - Case 3: patient dies between year 10 and 15 (I faked this).
  - Case 4: ...
- Questions: for Case 2 and 3, do we discard this patient's data from analysis?

More complex data in survival analysis

- Multivariate survival
  - the same patient can experience multiple types of time-to-events: cancer relapse and death
  - the same patient can experience the same type of time-to-event over multiple times: cancer recurrence
  - the patients in the same cluster can experience the same time-to-event similarly
- Different types of outcome present
  - longitudinal biomarkers
  - competing risks
  - informative censoring

Complex designs

- Different study designs
  - Clinical trial design
  - Case-cohort design
  - Nested case-control design
  - Multiple stage designs

A number of methods or tools developed for survival analysis

- Nonparametric methods: Kaplan-Meier estimator ...
- Semiparametric models: accelerated failure time model, proportional hazards model, proportional odds model, linear transformation models, gamma frailty model, and more.
- Machine learning methods: survival tree
- Advanced mathematical tools: martingales, empirical processes

Current research topics in survival analysis

- Joint analysis of survival outcomes and other non-survival outcomes
- High-dimensional analysis for survival outcomes
- Personalized medicine for survival outcomes

One-Sample Problem

- Let T denote the time-to-event outcome and C the potential censoring time.
- Full data consist of T can C.
- Observed data consist of (Y, Δ) where $Y = \min(T, C)$ and $\Delta = I(T \leq C)$. Thus, missing observations occur for those with $\Delta = 0$.
- For a one-sample problem, we are interested to estimate the survival function $S(t) = P(T \geq t)$ (the corresponding density is $f(t)$).

- Let $S_C(\cdot, t)$ be the censoring survival function given $T = t$. The observed likelihood function for $(Y = y, \Delta = \delta)$ is

$$\{f(y)S_C(y, y)\}^{\delta} \left\{ \int_y^{\infty} f(t) f_C(y; t) dt \right\}^{1-\delta}.$$

- Assume C and T are independent; then $S_C(\cdot, t) \equiv S_C(\cdot)$ is indepenent of t.
- The likelihood function becomes

$$f(y)^{\delta} S(y)^{1-\delta} S_C(y)^{\delta} f_C(y)^{1-\delta}.$$

- Therefore, we only need to consider maximizing the observed likelihood

$$\prod_{i=1}^{n} f(Y_i)^{\Delta_i} (S(Y_i))^{1-\Delta_i}$$

from n observations $(Y_i, \Delta_i), i = 1, ..., n$ to estimate $S(t)$.

NPMLE for survival function

- It assumes that the estimator for $S(t)$ (equivalently, $F(t) = 1 - S(t)$) has jumps at the observed times.
- Let p_i be the jump size of F at Y_i. The likelihood to be maximized is given by

$$\prod_{i=1}^{n} p_i^{\Delta_i} (1 - \sum_{Y_j \le Y_i} p_j)^{1-\Delta_i}.$$

- EM algorithm or iterative algorithms can be used to find p_i, resulting in the Kaplan-Meier estimator

$$\hat{S}(t) = \prod_{Y_i \le t} \left\{ 1 - \frac{\Delta_i}{\sum_{Y_j \ge Y_i} 1} \right\}.$$

NPMLE for hazards function

- Hazards rate function is commonly used to reparameterize the model:

$$\lambda(t) = f(t)/S(t) = \lim_{\epsilon \to 0+} P(T \in [t, t+\epsilon) | T \geq t)/\epsilon.$$

- Relationship between the hazards and the survival function:

$$S(t) = \exp\{-\Lambda(t)\}, \quad \Lambda(t) = \int_0^t \lambda(s) ds.$$

- The likelihood function in terms of the hazards rate:

$$\prod_{i=1}^{n} \lambda(Y_i)^{\Delta_i} \exp\{-\Lambda(Y_i)\}.$$

- NPMLE assumes Λ to jump at the observed times with jump size λ_i at Y_i.
- It gives the Breslow estimator for $S(t)$ as

Cox proportional hazards model

- This is the most common model to model the relationship between T and covariate X by assuming

$$\lambda(t|X) = \lambda(t) \exp\{X^T \beta\},$$

  where $\lambda(t|X)$ is the conditional hazards rate of T given X.
- In the model, $\lambda(t)$ is completely unknown, and β is the unknown parameter describing the log-hazards ratio of X on the outcome T.
- Under the conditional independent censoring assumption, i.e., T and C are independent given X, the observed likelihood function is

$$\prod_{i=1}^{n} \left\{ \lambda(Y_i) e^{X_i^T \beta} \right\}^{\Delta_i} \exp\left\{ -\Lambda(Y_i) e^{X_i^T \beta} \right\}.$$

Cox partial likelihood

- NPMLE assumes that the estimator for Λ has jumps at the observed time points.
- For given β, NPMLE for Λ has jump size at Y_i as

$$\frac{\Delta_i}{\sum_{Y_j \geq Y_i} e^{X_j^T \beta}}.$$

- Plugging it back into the likelihood function gives

$$\prod_{i=1}^{n} \left\{ \frac{e^{X_i^T \beta}}{\sum_{Y_j \geq Y_i} e^{X_j^T \beta}} \right\}^{\Delta_i},$$

which is the famous Cox partial likelihood function.

NPMLE for more general transformation models

- Event history data: survival data, recurrent event time data
- Counting process: $N^*(t)$—the number of events that have occurred by time t
- Classical models
  - the proportional hazards model for survival data
  - the Andersen-Gill intensity model for recurrent event time data
- A general form of hazard rate/intensity rate

$$\lambda_Z(t) = Y^*(t) \exp\{\beta^T Z(t)\} \lambda(t)$$

  - $Z(t)$ is a vector of possibly time-varying covariates;
  - $Y^*(t)$ is the at-risk process;

- Estimation based on the partial likelihood principle (Cox, 1972, 1975)
- Large-sample theory developed via the counting process martingale theory (Andersen & Gill, 1982)
- The proportional hazards model may be violated in certain applications; other alternatives include
  - the proportional odds model (Bennett, 1983; Pettitt, 1984)
  - the linear transformation models (Dabroska & Doksum, 1988, Cheng, Wei & Ying, 1995, 1997, Chen, Jin & Ying, 2002)

- We consider transformation models for general counting process

$$\Lambda_Z(t) = G\left\{\int_0^t Y^*(s)e^{\beta^T Z(s)}d\Lambda(s)\right\}.$$

  - $G(x)$ is a strictly increasing function; $G(0) = 0$ and $G(\infty) = \infty$.
  - $G(x) = x$ reduces to the Andersen-Gill intensity model.
  - For survival data, it becomes

$$\int_0^T e^{\beta^T Z(s)}d\Lambda(s) = G^{-1}(-\log\epsilon_0), \quad \epsilon_0 \sim \text{Unif}(0,1).$$

- Choices of transformation $G(x)$
  - the Box-Cox transformations

$$G(x) = \{(1+x)^{\rho} - 1\}/\rho$$

  - the logarithmic transformations

$$G(x) = \log(1+rx)/r.$$

  - plots of these transformations

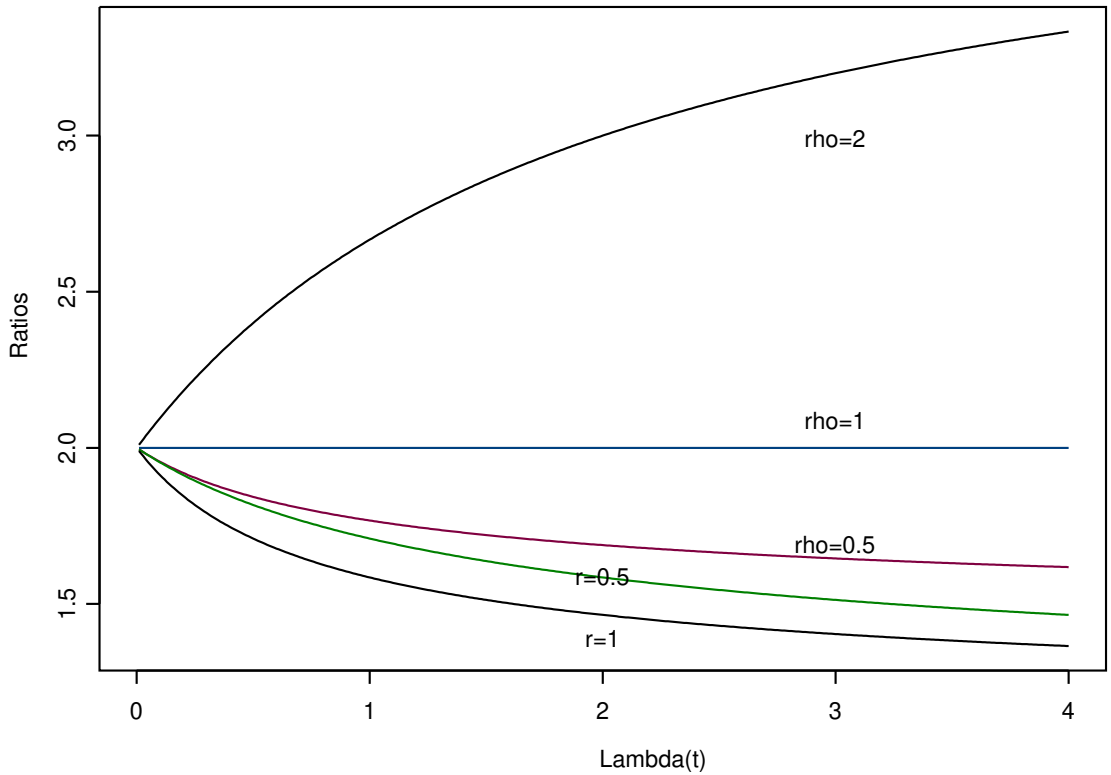

Figure: Plots of the ratios $\Lambda_z(t)/\Lambda_{z=0}(t)$ against $\Lambda(t)$ with $e^{\beta z} = 2$ under the Box-Cox and logarithmic transformations

Our work aims to

- propose inference procedure for general transformation models;
- give the asymptotic properties of the MLE;
- analyze two data sets using the proposed method.

Inference Procedure

- Observed data

$$\{N_i(t), Y_i(t), Z_i(t); t \in [0, \tau]\}$$

  - $N_i(t) = N_i^*(t \wedge C_i)$ where C_i is independent censoring time;
  - $Y_i(t) = I(C_i \geq t) Y_i^*(t)$ is the observed at-risk process;
  - τ is the duration of the study.

- Parameters of interest: β and $\Lambda(t)$
- The observed log-likelihood function

$$\sum_{i=1}^{n} \Big[\int_0^\tau \log \lambda(t) dN_i(t) + \int_0^\tau \log G' \left\{ \int_0^t Y_i(s) e^{\beta^T Z_i(s)} d\Lambda(s) \right\} dN_i(t)$$

$$+ \int_0^\tau \beta^T Z_i(t) dN_i(t) - G \left\{ \int_0^\tau Y_i(t) e^{\beta^T Z_i(t)} d\Lambda(t) \right\} \Big].$$

- Nonparametric maximum likelihood estimation (NPMLE)
  - the maximum is ∞;
  - we restrict Λ to be a right-continuous step-function with jumps at the observed events;
  - we maximize the following objective function

$$\sum_{i=1}^{n} \Big[\int_0^\tau \log \Lambda\{t\} dN_i(t) + \int_0^\tau \log G' \left\{ \int_0^t Y_i(s) e^{\beta^T Z_i(s)} d\Lambda(s) \right\} dN_i(t) + \int_0^\tau \beta^T Z_i(t) dN_i(t) - G \left\{ \int_0^\tau Y_i(t) e^{\beta^T Z_i(t)} d\Lambda(t) \right\} \Big],$$

  - $\Lambda\{t\} = \Lambda(t) - \Lambda(t-)$

- Computation algorithm
  - it is an optimum search based on the quasi-Newton method;
  - gradient and hessian of the objective function are provided;
  - an approximate quadratic function is maximized locally;
  - algorithm stops when either search step or the length of search direction is small.
- Optimization can be implemented using *fminunc* in MatLab.
- We recommend starting values $\beta = 0$ and $\Lambda\{X_{ij}\} = 1/n$.

- Variance estimation
  - We treat the jump sizes of Λ as classical parameters.
  - The inverse of the observed information matrix for β and these jump sizes approximate the asymptotic covariance in a "non-rigorous" way.
  - The Delta method can be used to estimate the asymptotic variance of $F(\hat{\Lambda}_n, \hat{\beta}_n)$ for any differentiable functional $F(\cdot, \cdot)$.
  - When $\hat{\Lambda}_n$ is not of interest, the profile likelihood function can be used to estimate the asymptotic variance of $\hat{\beta}_n$.

Asymptotic Properties

- Technical assumptions
  - $\gamma_0(t) + \gamma^T Z(t) = 0$ a.s implies $\gamma_0(t) = 0$ and $\gamma = 0$.
  - The probability of observing the whole event process in $[0, \tau]$ is positive.
  - The true β is in a compact set and $\Lambda_0'(x) > 0$.
  - $G(x)$ arises from either the Box-Cox transformations or the logarithmic transformations.

- Summary of results
  - Parameters (β_0, Λ_0) are identifiable.
  - With probability one,

$$|\hat{\beta}_n - \beta_0| + \sup_{t \in [0,\tau]} |\hat{\Lambda}_n(t) - \Lambda_0(t)| \to 0.$$

  - $\sqrt{n}(\hat{\beta}_n - \beta_0, \hat{\Lambda}_n - \Lambda_0)$ converges in distribution to a Gaussian process in $R^d \times l^\infty([0,\tau])$.
  - The variance estimators based on the observed information matrix or the profile likelihood functions are consistent.

Simulation Studies

- Simulation setting
  - $Z_1 \sim \text{Bernoulli}(0, 1, 0.5)$;
  - $Z_2 = Z_1 + \epsilon I(|\epsilon| \leq 3)$, $\epsilon \sim N(0, 1)$;
  - $\Lambda(t) = t$, $\beta_1 = -1$ and $\beta_2 = 0.2$;
  - $G(t)$ comes from either the Box-Cox transformations or the logarithmic transformations;
  - Sample sizes $n = 100$ or $n = 200$.

- Results from the Box-Cox transformations
  - The average number of events for $\rho = 0.5$ is 1.41.
  - The average number of events for $\rho = 2$ is 4.38.

		$n = 100$				$n = 200$			
Model	Parameter	Bias	SE	SEE	CP	Bias	SE	SEE	CP
$\rho = 0.5$	β_1	-0.015	0.245	0.250	0.955	0.007	0.173	0.175	0.952
	β_2	0.005	0.109	0.111	0.959	-0.007	0.075	0.077	0.950
	$\Lambda(\tau/4)$	-0.004	0.132	0.136	0.961	-0.005	0.095	0.095	0.958
	$\Lambda(\tau/2)$	-0.007	0.219	0.229	0.966	-0.004	0.161	0.162	0.949
	$\Lambda(\tau)$	0.000	0.407	0.423	0.954	-0.009	0.299	0.297	0.950
$\rho = 2.0$	β_1	-0.006	0.087	0.086	0.955	-0.002	0.060	0.060	0.955
	β_2	0.000	0.033	0.032	0.945	0.000	0.023	0.022	0.952
	$\Lambda(\tau/4)$	-0.005	0.073	0.071	0.943	-0.002	0.050	0.050	0.947
	$\Lambda(\tau/2)$	-0.003	0.084	0.084	0.956	-0.001	0.058	0.059	0.955
	$\Lambda(\tau)$	-0.003	0.113	0.110	0.946	0.000	0.079	0.077	0.945

- Results from the logarithmic transformations
  - The average number of events for $r = 0.5$ is 0.76.
  - The average number of events for $r = 1$ is 1.05.
  - The average number of events for $r = 2$ is 1.32.

		$n = 100$				$n = 200$			
Model	Parameter	Bias	SE	SEE	CP	Bias	SE	SEE	CP
$r = 0.5$	β_1	-0.007	0.268	0.277	0.961	-0.006	0.190	0.195	0.951
	β_2	0.007	0.123	0.125	0.961	-0.001	0.086	0.087	0.945
	$\Lambda(\tau/4)$	-0.009	0.141	0.143	0.961	-0.004	0.101	0.101	0.958
	$\Lambda(\tau/2)$	-0.001	0.249	0.253	0.955	-0.004	0.175	0.179	0.962
	$\Lambda(\tau)$	-0.014	0.471	0.497	0.966	0.010	0.350	0.351	0.950
$r = 1$	β_1	-0.008	0.355	0.355	0.955	0.003	0.253	0.249	0.943
	β_2	0.005	0.161	0.161	0.948	-0.002	0.112	0.112	0.955
	$\Lambda(\tau/4)$	-0.007	0.177	0.175	0.948	-0.001	0.125	0.124	0.947
	$\Lambda(\tau/2)$	-0.004	0.338	0.331	0.952	0.002	0.237	0.234	0.950
	$\Lambda(\tau)$	0.023	0.710	0.682	0.941	0.009	0.483	0.477	0.940
$r = 2$	β_1	0.006	0.467	0.479	0.961	0.003	0.325	0.335	0.949
	β_2	-0.001	0.227	0.217	0.952	-0.008	0.151	0.151	0.946
	$\Lambda(\tau/4)$	-0.011	0.230	0.229	0.952	-0.002	0.171	0.163	0.945
	$\Lambda(\tau/2)$	0.005	0.459	0.462	0.955	0.007	0.341	0.326	0.947
	$\Lambda(\tau)$	0.044	0.975	0.977	0.950	0.026	0.707	0.683	0.951

- Conclusions from simulation studies
  - Estimates are virtually unbiased;
  - The variance estimates reflect the true variations;
  - The confidence intervals achieve proper coverages.
  - It took 3 hours on an IBM BladeCenter HS20 machine to complete all the simulations.
  - No convergence problem was encountered in any of 10,000 simulated data sets.

- Efficiency gain over other approaches
  - Chen, Jin & Ying (2002) studies transformation models for survival data based on estimating equations.
  - A simulation study is conducted with $\Lambda(t) = 3t$, $\beta_1 = -1$ and $\beta_2 = 0.2$; sample size is 100.

			Proposed estimator				Chen et al. estimator			
C%	r	Par.	Bias	SE	SEE	CP	Bias	SE	SEE	CP
25%	0.5	β_1	-0.026	0.378	0.358	0.937	-0.035	0.393	0.366	0.947
		β_2	0.005	0.165	0.159	0.949	0.006	0.172	0.164	0.940
	1	β_1	-0.022	0.440	0.420	0.941	-0.032	0.482	0.446	0.941
		β_2	0.005	0.193	0.187	0.956	0.007	0.210	0.203	0.949
	2	β_1	-0.023	0.545	0.523	0.944	-0.029	0.655	0.602	0.949
		β_2	0.005	0.242	0.234	0.939	0.005	0.286	0.279	0.943
50%	0.5	β_1	-0.029	0.437	0.413	0.951	-0.051	0.444	0.410	0.945
		β_2	0.006	0.187	0.183	0.951	0.006	0.191	0.184	0.947
	1	β_1	-0.031	0.488	0.463	0.944	-0.054	0.512	0.469	0.940
		β_2	0.007	0.213	0.207	0.955	0.008	0.225	0.214	0.948
	2	β_1	-0.025	0.579	0.555	0.942	-0.045	0.644	0.588	0.938
		β_2	0.006	0.257	0.249	0.956	0.009	0.284	0.274	0.949

- Conclusions from efficiency comparison
  - The asymptotic approximation works well for both approaches.
  - Chen et al. estimators are less efficient, especially when r is large and censoring is low.
  - Our algorithm always converged.
  - Chen et al.'s failed to converge in about 2% of simulated data sets.

Data Applications

- Data set I
  - the Veteran'a Administration lung cancer trial
  - 97 patients without prior therapy were studied;
  - covariates included performance status and tumor types;
  - the data has been analyzed by Bennett (1983), Pettitt (1984), Cheng et al. (1995), Murphy et al. (1997) and Chen et al. (2002).
- We analyze data using transformation models.
- The log-likelihood function is used as the criterion of choosing the best fit.

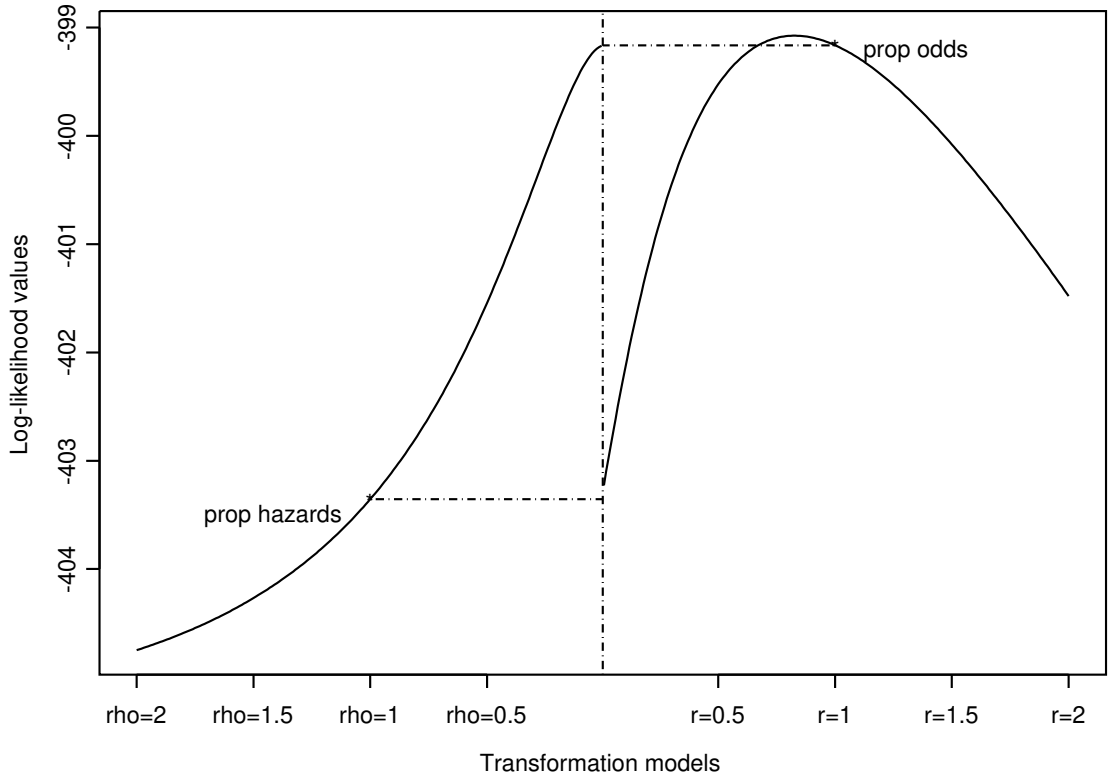

Figure: log-likelihood function in fitting VA lung cancer data

- Estimates in analyzing the VA lung cancer data

	$r = 0$	$r = 1$	$r = 1.5$	$r = 2$
Performance status	-0.024 (0.006)	-0.053 (0.010)	-0.063 (0.012)	-0.072 (0.014)
Adeno vs large tumor	0.851 (0.348)	1.314 (0.554)	1.497 (0.636)	1.679 (0.712)
Small vs large tumor	0.547 (0.321)	1.383 (0.524)	1.605 (0.596)	1.814 (0.661)
Squam vs large tumor	-0.215 (0.347)	-0.181 (0.588)	-0.075 (0.675)	0.045 (0.749)

- Comparison with other work
  - The results differ appreciably from those of Chen et al. (2002).
  - For $r = 0$, the numbers agree with the standard software output.
  - For $r = 1$, the results are similar to those of Murphy et al. (1997).

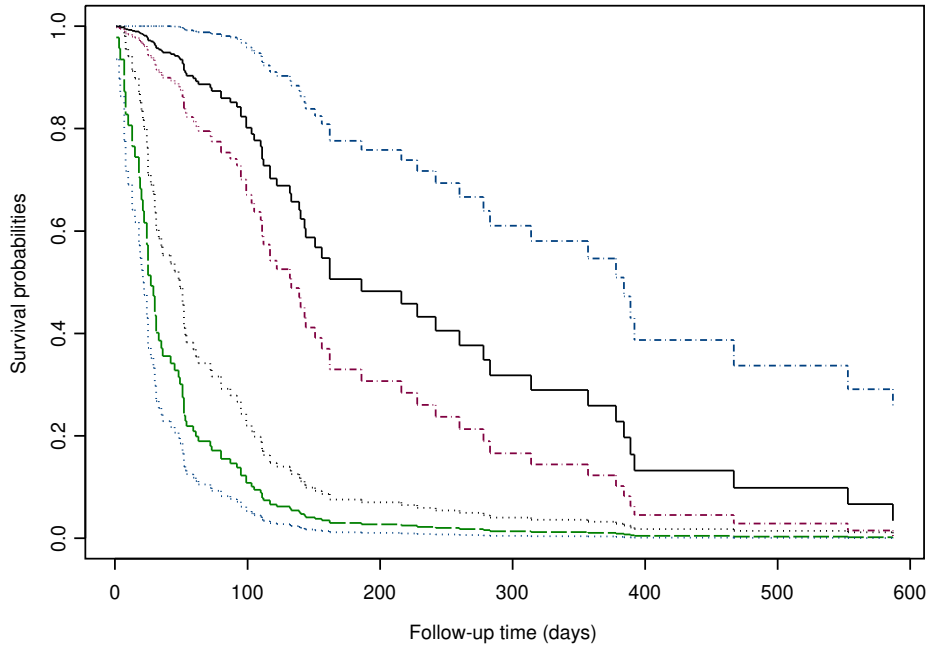

Figure: Estimated survival curves for the lung cancer patients: the upper three curves pertain to the point estimate and 95% confidence limits for a patient with large tumor and performance status of 80, and the lower three curves to those of a patient with small tumor and performance status of 40.

- Data set II
  - the recurrent bladder tumor data
  - 86 patients were on the placebo or thiotepa;
  - other covariates included number of tumors and tumor sizes;
  - the data has been analyzed by Wei, Lin & Weissfeld (1989) and Therneau & Grambsch (2000).
- We analyze data using transformation models.
- The log-likelihood function is used as the criterion of choosing the best fit.

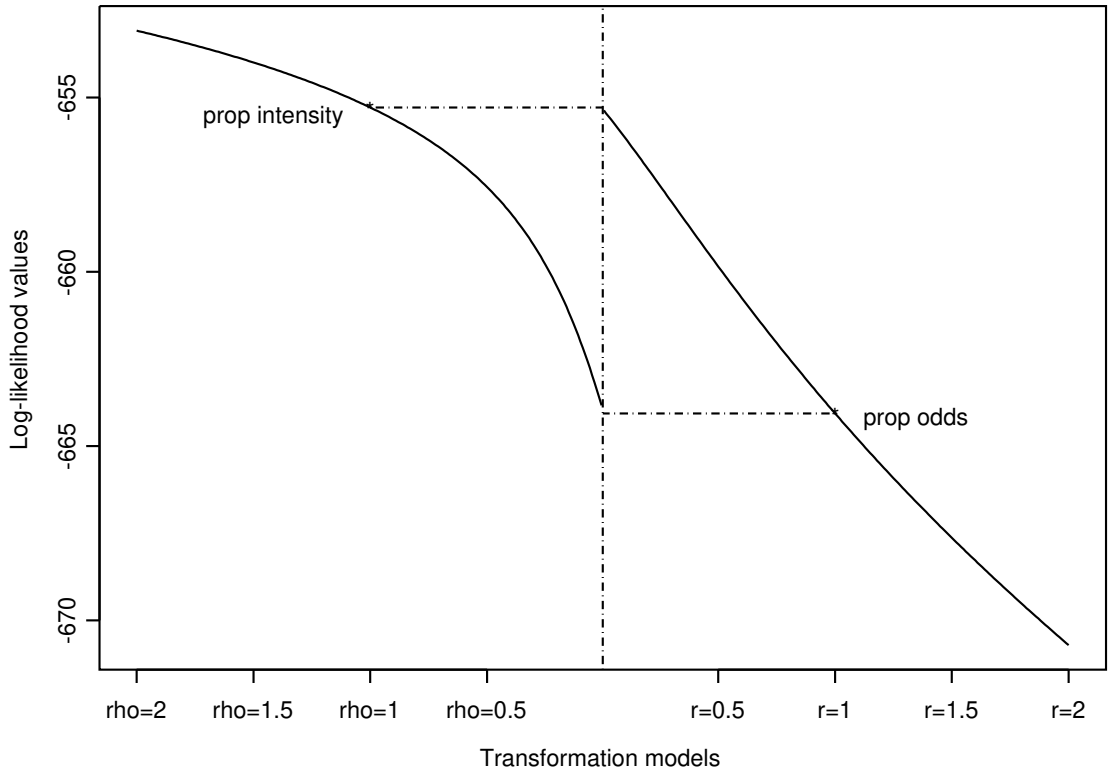

Figure: log-likelihood function in fitting bladder tumor data

- Estimates in analyzing the bladder tumor data

	$\rho = 2$	$\rho = 1$	$\rho = 0.5$	$r = 1$
Treatment	-0.369 (0.136)	-0.524 (0.187)	-0.701 (0.244)	-0.974 (0.358)
No. tumors	0.141 (0.030)	0.201 (0.044)	0.269 (0.061)	0.352 (0.101)
Tumor size	-0.035 (0.048)	-0.040 (0.065)	-0.041 (0.084)	-0.013 (0.123)

- Prediction of the survival function of X_2 given $X_1 = x_1$ for subjects with covariate values z:

$$\exp\left[G\left\{\hat{\Lambda}_n(x_1)e^{\hat{\beta}_n^T z}\right\} - G\left\{\hat{\Lambda}_n(t)e^{\hat{\beta}_n^T z}\right\}\right].$$

- The delta method is used to calculate the point-wise standard error.
- We plot the predicted curves for $x_1 = 20$ and $\rho = 2$.

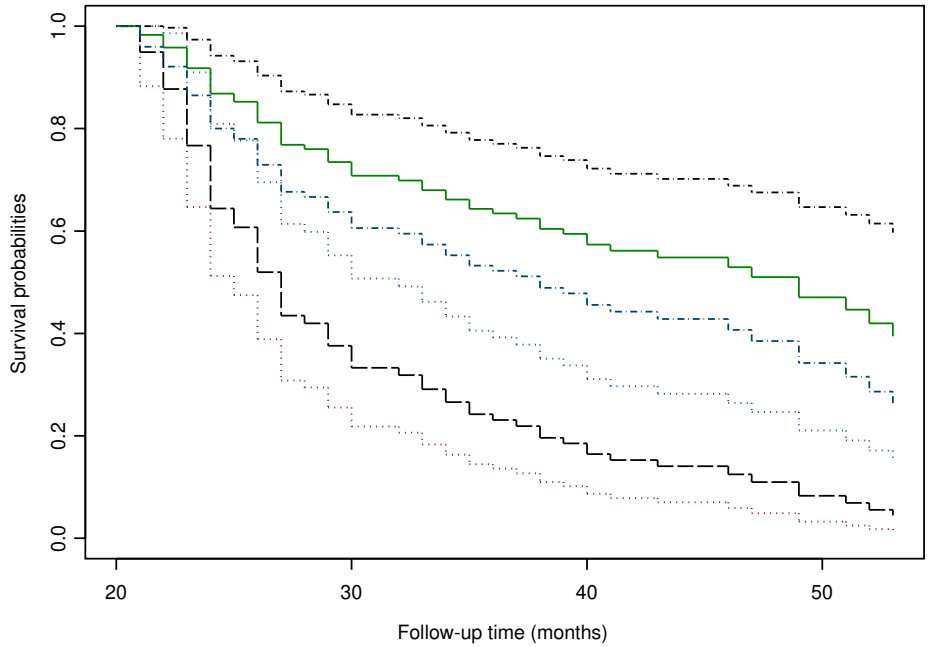

Figure: Estimated conditional survival curves for the bladder tumor patients: the upper three curves correspond to the point estimate and 95% confidence limits for a thiotepa patient with one initial tumor, and the lower three curves to those of a placebo patient with four initial tumors.

Concluding Remarks

- The transformation models provide a flexible choice for fitting complicated event time data.
- The NPMLEs have the advantage of being asymptotic efficient and good small-sample performance.
- The likelihood-based approach implies many model selection methods, such as the AIC and the likelihood-based cross validation.
- The observed information matrix yields reliable estimates of the variances in this semiparametric context.

- Other generalizations can be possible, including
  - recurrent event time data with subject-specific random effects,
  - clustered failure times,
  - data with truncation and other censoring patterns.