REGRESSION METHODS FOR MARKED OUTCOMES

Brent A. Johnson

Department of Biostatistics and Bioinformatics
Emory University

13 July 2013

Motivating Applications

- In some clinical studies, there may be an interest in an auxiliary endpoint that is measured at the point of failure, not merely the time to failure.
- If every participant in the study experiences a failure, then the observed data includes the failure time and the auxiliary endpoint for the entire study sample.
- On the other hand, if some participants do not experience the failure event by the end of the follow-up period, both the time to failure and the auxiliary endpoint are unobserved.
- In this case, we say the failure time is right-censored and auxiliary endpoint is dependently censored.

Vaccine Efficacy Trials

From 1998 - 2003, VaxGen Inc. conducted the first HIV vaccine efficacy trial.

- 5403 high-risk, uninfected subjects randomized to the AIDSVAX vaccine ($n_1 = 3598$) or placebo ($n_2 = 1805$)
- 368 subjects became infected with the virus
- Naturally, VaxGen hypothesized that vaccine efficacy would be higher against HIV strains with amino acid sequences similar to those strains used to make the vaccine (MN and GNE8)
- **Question**: What is the efficacy of the AIDSVAX vaccine?
- **Note**: Endpoint is only available on patients that acquire HIV

Therapeutic AIDS Trials

- ACTG 5095 randomized, multi-center clinical trial of HIV+ ART-naive patients
- Study designed for 120 weeks follow-up
- 1147 study participants randomized to one of three treatment groups: (i) ABC+3TC+ZDV, (ii) 3TC+ZDV+EFV, or (iii) ABC+3TC+ZDV+EFV
- **Questions of Interest**:
  - What is the (un)-adjusted difference in resistance endpoints among ARTs at the time of virologic failure?
  - What is the (un)-adjusted difference in immunological endpoints (CD4+, CD8+, etc.) among ARTs at the time of virologic failure?
- **Note**: Resistance endpoint is only available at virologic failure

Lifetime Censored Medical Cost

- SWOG 9509: randomized clinical trial of untreated patients with advanced nonsmall cell lung cancer
- 408 study participants randomized to two treatment groups: (i) Paclitaxel plus Carboplatin, and (ii) Vinorelbine plus Cisplatin
- **Hypothesis**: What is the (un)-adjusted difference in lifetime medical costs between treatments?
- **Notes**:
  - Lifetime cost accrues over time
  - Lifetime medical cost observed for only those patients whose entire lifetime is observed

Notation

Consider the linear regression model

$$Y_i = X_i'\beta + \varepsilon_i,\ i = 1, \ldots, n,$$

- $Y_i :=$ auxiliary endpoint of interest, e.g. lifetime medical cost
- $T_i :=$ time at which Y_i is measured (event time)
- $X_i :=$ d-vector of risk factors
- $\beta := (\beta_1, \ldots, \beta_d)'$ regression coefficients
- ε_i are i.i.d. errors

Observed Data

The true data are:

$$(Y_i, T_i, X_i),\ i = 1, \ldots, n.$$

However, the observed data are:

$$(W_i, U_i, \Delta_i, X_i),\ i = 1, \ldots, n,$$

- $W_i := Y_i \cdot I(T_i \leq C_i)$
- $U_i := T_i \wedge C_i$
- $\Delta_i := I(T_i \leq C_i)$

Objective: Estimate regression coefficients β.

Identifiability

Because of dependent censoring, there is concern whether the marginal mean of lifetime medical cost is identifiable from the observed data.

Define

- $\tau_T :=$ max support for T
- $\tau_C :=$ max support for C

Then

- If $\tau_T \leq \tau_C$, then all is well
- If $\tau_C < \tau_T$, the joint distribution of (Y, T) is not observed on $(-\infty, \infty) \times (\tau_C, \infty)$ and, hence, the marginal distribution of Y can be nowhere identifiable

Modeling the Time-Restricted Marginal Mean

One approach to address identifiability problem is to model the time-restricted lifetime medical cost, instead of lifetime medical cost.

Suppose we want to model the $L = 5$-year restricted medical cost, Y^L, defined as the cumulative medical cost to T or $L = 5$ years, whichever comes first. Define the time-restricted observables:

- $W_i^L := Y_i \cdot I(T_i \wedge L \leq C_i)$
- $U_i^L := T_i \wedge L \wedge C_i$
- $\Delta_i^L := I(T_i \wedge L \leq C_i)$

Then, we can estimate $E(Y^L)$ or model $E(Y^L|X)$.

IPW Estimation

To estimate mean $E(Y^L)$, for example, Zhao and Tsiatis (1997) proposed the statistic,

$$n^{-1}\sum_{i=1}^{n}\frac{\Delta_i^L W_i^L}{\text{pr}\{C > \min(T_i, L)\}}.$$

To parametrically model $E(Y^L|X)$, a regression estimator may be constructed by replacing Y_i^L above with a score, say ψ,

$$n^{-1}\sum_{i=1}^{n}\frac{\Delta_i^L \psi(W_i^L, X, \beta)}{\text{pr}\{C > \min(T_i, L)|X_i\}}.$$

Using arguments from survival analysis, one can show these estimators are CAN.

Notes on Time-Restricted Marginal Mean

This technique has been studied extensively with contributions by Zhao and Tsiatis (1997), Lin et al. (1997), Bang and Tsiatis (2000), Zhao and Tian (2001), Strawderman (2002), Zhao et al. (2007).

Potential caveats:

- Clearly, Y and Y^L are not the same quantity, although both may be of interest;
- Recall, $Y = Y^L$ if $\min(T, L) = T$ but can be very different otherwise;
- **Example**: Approximately 90% of patients living with ALS die within 7 years of the date of diagnosis. So, one expects that modeling $E(Y|X)$ or $E(Y^L|X)$ with $L = 7$ is similar.

Model the Joint Distribution

Huang and Louis (1998) showed that one may obviate the identifiability concern in the marginal mean by estimating the mark & time joint distribution. Huang and Lovato (2002) extended these ideas to two-sample tests and Huang (2002) proposed a regression coefficient estimator. Suppose

$$\begin{pmatrix} Y_i \\ T_i \end{pmatrix} = X_i' \begin{pmatrix} \beta \\ \gamma \end{pmatrix} + \varepsilon_i, \ (i = 1, \ldots, n),$$

- $\gamma := (\gamma_1, \ldots, \gamma_d)'$ time-scale coefficients
- ε_i are i.i.d. according to a unspecified bivariate distribution function

Note: WLOG, assume T_i is measured on the log-scale

Coefficient Estimation

Define the counting process $N_i(t,\gamma) = I(U_i - X_i'\gamma \leq t, \Delta_i = 1)$ and the marked process $N_i^\dagger(t,\theta) = (W_i - X_i'\beta)N_i(t,\gamma)$. Then, define the pair of estimating functions:

$$\begin{aligned}
\mathbb{S}_\beta(\beta,\gamma;O) &= \sum_{i=1}^{n} \int_{-\infty}^{\infty} \phi(t,\gamma)\left\{X_i - \bar{X}(t,\gamma)\right\} dN_i^\dagger(t,\gamma), \\
\mathbb{S}_\gamma(\gamma;O) &= \sum_{i=1}^{n} \int_{-\infty}^{\infty} \phi(t,\gamma)\left\{X_i - \bar{X}(t,\gamma)\right\} dN_i(t,\gamma),
\end{aligned}$$

where $\bar{X}(t,\gamma) = \sum_i X_i R_i(t,\gamma)/\sum_i R_i(t,\gamma)$, the at-risk indicator $R_i(t,\gamma) = I(U_i - X_i'\gamma \geq t)$, $\phi(t,\gamma)$ is a weight function satisfying regularity conditions and O is the observed data.

Comments

- $N_i(t,\gamma)$ is an ordinary counting process that increments by $+1$ on the time-scale support points whereas the marked process $N_i^{\dagger}(t,\theta)$ takes random jump size of $(W_i - X_i'\beta)$;
- $\mathbb{S}_\gamma(\gamma; O)$ is the weighted log-rank estimating function (Tsiatis, 1990; Wei et al., 1990; Louis, 1981);
- Solving for $0 = (\mathbb{S}_\beta', \mathbb{S}_\gamma')'$ leads to a strongly consistent estimator for $\theta = (\beta', \gamma')'$.
- $n^{1/2}(\widehat{\theta} - \theta_0)$ is asymptotically normal

Closed-form Solution

Note that $\mathbb{S}_\gamma(\gamma; O)$ does not depend on β so one can solve for the mark-scale coefficients in two-stages.

Let $\widehat{\gamma}_\phi$ be the weighted log-rank estimator for γ and solve for β in $0 = \mathbb{S}_\beta(\beta, \widehat{\gamma}_\phi; O)$:

$$\widetilde{\beta}_\phi = \left[\sum_{i=1}^{n} \int_{-\infty}^{\infty} \phi(t, \widehat{\gamma}_\phi) \left\{ X_i - \bar{X}(t, \widehat{\gamma}_\phi) \right\} X_i' \, dN_i(t, \widehat{\gamma}_\phi) \right]^{-1} \times \left[\sum_{i=1}^{n} \int_{-\infty}^{\infty} \phi(t, \widehat{\gamma}_\phi) \left\{ X_i - \bar{X}(t, \widehat{\gamma}_\phi) \right\} W_i \, dN_i(t, \widehat{\gamma}_\phi) \right],$$

Limitations of Huang's Estimator

- Despite its robustness on the time-scale, Huang's estimator can be quite sensitive on the mark- or cost-scale
- Given the numerous articles describing the unusual architecture of cost data, it would seem that robustness on the mark-scale is as or more important as on time-scale
- Huang (2002) noted that one could construct a robust marked process through $N_i^{\dagger}(t,\theta) = \psi(W_i - X_i'\beta)N_i(t,\gamma)$ but offered no details on such extension, $\psi(\cdot)$ is Huber's ψ
- I contend that such robust extension would be non-trivial and not follow the standard tricks from robust M-estimators

Sufficient Condition for Consistency

Let

- $e_{Y_i}(\beta) = W_i - X_i'\beta$
- $f(t)$ be a monotone function in t
- Define $N_i^\dagger(t,\theta) = f\{e_{Y_i}(\beta)\}\, dN_i(t,\gamma)$

Define

$$\begin{aligned}\mathbb{S}_\beta(\beta,\gamma;O) &= \sum_{i=1}^{n}\int_{-\infty}^{\infty}\phi(t,\gamma)\left\{X_i-\bar{X}(t,\gamma)\right\}\, dN_i^\dagger(t,\theta),\\ &= \sum_{i=1}^{n}\int_{-\infty}^{\infty}\phi(t,\gamma)\left\{X_i-\bar{X}(t,\gamma)\right\}\, f\{e_{Y_i}(\beta)\}\, dN_i(t,\gamma).\end{aligned}$$

Solving $0 = \mathbb{S}_\beta(\beta,\gamma_0;O)$ leads to consistent estimator for β.

Rank-based Estimator

Let

- $R\{e_{Y_i}(\beta)\}$ to be the rank of $e_{Y_i}(\beta)$ among the uncensored residuals

Define new estimator $\widehat{\theta}$ as solution to $0 = (\mathbb{S}'_\beta, \mathbb{S}'_\gamma)'$ with

$$\mathbb{S}_\beta(\beta, \gamma; O) = \sum_{i=1}^{n} \int_{-\infty}^{\infty} \phi(t, \gamma) \left\{ X_i - \bar{X}(t, \gamma) \right\} R\{e_{Y_i}(\beta)\} \, dN_i(t, \gamma).$$

Under the same conditions in Huang (2002), one can show that $\widehat{\theta}$ is strongly consistent and $n^{1/2}(\widehat{\theta} - \theta_0)$ is asymptotically normal under somewhat different arguments than Huang (2002).

MARTINGALE-BASED ARGUMENTS

For Huang's (2002) estimator,

$$\mathbb{S}_\beta(\beta,\gamma;O)=\sum_{i=1}^{n}\int_{-\infty}^{\infty}\underbrace{\phi(t,\gamma)\left\{X_i-\bar{X}(t,\gamma)\right\}\psi\{e_{Y_i}(\beta)\}}_{\mathcal{F}_t\text{-measureable}}\,dN_i(t,\gamma).$$

where

$$\begin{aligned}\mathcal{F}_t \;=\; \sigma\big\{&N_i(u,\gamma),I(U_i-X_i'\gamma\le u,\Delta_i=0),\\ &Y_i\cdot I(U-X_i'\gamma\le u,\Delta_i=1),X_i,\ u\le t,\ i=1,\ldots,n\big\}.\end{aligned}$$

Evidently, this is not true for the rank process $R\{e_{Y_i}(\beta)\}$.

Rewrite Score as Function of Empirical Measure

WLOG, write $R\{e_{Y_i}(\beta)\} = \sum_j I\{e_{Y_j}(\beta) \leq e_{Y_i}(\beta), \Delta_j = 1\}$

$$
\begin{aligned}
&\mathbb{S}_\beta(\beta, \gamma; O) = \\
&= \sum_{i=1}^{n} \int_{-\infty}^{\infty} \phi(t, \gamma) \left\{ X_i - \bar{X}(t, \gamma) \right\} R\{e_{Y_i}(\beta)\} \, dN_i(t, \gamma), \\
&= \sum_{i=1}^{n} \Delta_i \phi\{e_{T_i}(\gamma), \gamma\} \left[X_i - \bar{X}\{e_{T_i}(\gamma), \gamma\} \right] R\{e_{Y_i}(\beta)\} \\
&= \sum_{i=1}^{n} \sum_{j=1}^{n} \Delta_i \Delta_j \phi\{e_{T_i}(\gamma), \gamma\} \left[X_i - \bar{X}\{e_{T_i}(\gamma), \gamma\} \right] I\{e_{Y_j}(\beta) \leq e_{Y_i}(\beta)\}.
\end{aligned}
$$

Inference via Multiplier Bootstrap

Jin et al. (2006) presented a resampling scheme for the Buckley-James estimator and we can adopt a similar strategy here.

1. Simulate $(Z_1, \ldots, Z_n)$ i.i.d. such that $E(Z_i) = \text{var}(Z_i) = 1$
2. Compute the estimate $\widehat{\gamma}^*_\phi$ by minimizing the k-step perturbed weighted Gehan loss function
3. Compute the estimate $\widehat{\beta}^*_\phi$ by solving the perturbed system $0 = \mathbb{S}^*_\beta(\beta, \widehat{\gamma}^*_\phi; O, Z)$,

$$\mathbb{S}^*_\beta(\beta, \gamma; O, Z) = \sum_{i=1}^{n} Z_i \int_{-\infty}^{\infty} \phi^*(t, \gamma) \left\{ X_i - \bar{X}^*(t, \gamma) \right\} R\{e_{Y_i}(\beta)\} \, dN_i(t, \gamma)$$

4. Repeat B times and use the Wald or percentile method

Scenario

Simulate data according to the model

$$\begin{pmatrix} Y_i \\ T_i \end{pmatrix} = X_i' \begin{pmatrix} \beta \\ \gamma \end{pmatrix} + \varepsilon_i, \ (i = 1, \ldots, n)$$

where $\beta = (1, 0.5)'$, $\gamma = (1, 1)'$,

- X_i are standard normal;
- C_i Un$(0, 6)$;
- and ε_i are i.i.d.

$$(1 - \pi) \times N(0, \Omega) + \pi \times N\left[\begin{pmatrix} 10 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right],$$

with $\Omega = \{\omega_{jk}\}$, $\omega_{11} = \omega_{22} = 1$, and $\omega_{12} = \omega_{21} = 0.5$.

TABLE: Monte Carlo simulation results

			Contamination Probability								
			0%			1%			2.5%		
Method	Weight	Parameter	Bias	SE	SEE	Bias	SE	SEE	Bias	SE	SEE
$n = 40$											
Huang	Gehan	β_1	5	201	184	-5	247	205	5	313	256
		β_2	-5	202	185	6	256	206	16	301	254
	Log-rank	β_1	1	175	170	-7	209	190	6	273	241
		β_2	-6	189	173	8	235	190	18	297	248
Robust	Gehan	β_1	7	220	204	-8	234	216	-2	245	230
		β_2	-7	220	205	-6	227	214	-5	232	230
	Log-rank	β_1	-1	196	236	-7	204	248	-3	208	265
		β_2	-8	206	248	-2	211	252	-2	206	271
$n = 60$											
Huang	Gehan	β_1	3	185	153	12	203	177	8	287	238
		β_2	7	179	154	-16	222	181	-38	274	230
	Log-rank	β_1	1	162	141	9	193	163	2	261	225
		β_2	1	155	142	-13	208	166	-25	255	226
Robust	Gehan	β_1	1	183	165	4	176	166	2	178	178
		β_2	7	179	163	-13	184	172	-19	184	181
	Log-rank	β_1	-2	172	169	4	166	175	3	169	184
		β_2	-2	150	168	-8	167	174	-11	169	182

Analysis of SWOG 9509

- Randomized clinical trial of untreated patients with advanced nonsmall cell lung cancer
- 408 study participants randomized to two groups:
  1. Paclitaxel plus Carboplatin
  2. Vinorelbine plus Cisplatin
- **Objective**: What is the expected difference in lifetime medical costs between the two treatments after adjusting for age and LDH?

More Notes on Analysis

Cost of Resource Utilization

- Resource utilization was monitored as part of study
- Collected at 3, 6, 12, 18, & 24 mos.
- Cost assigned using standard procedures
- Re-calibrated back to 1998 dollars

• 10 participants had insufficient data and were removed

TABLE: Analysis results for the lung cancer data on log-scale

		Huang			Robust		
Weight	Variable	Estimate	SE	95% CI	Estimate	SE	95% CI
Gehan	Treatment	0.405	0.134	(0.140, 0.652)	0.357	0.131	(0.103, 0.612)
	LDH	0.164	0.138	(-0.094, 0.446)	0.156	0.131	(-0.102, 0.390)
	Age	-0.069	0.066	(-0.202, 0.066)	-0.081	0.062	(-0.200, 0.046)
Log-rank	Treatment	0.338	0.121	(0.109, 0.582)	0.300	0.113	(0.086, 0.532)
	LDH	0.141	0.121	(-0.117, 0.366)	0.118	0.112	(-0.102, 0.336)
	Age	-0.050	0.056	(-0.168, 0.058)	-0.059	0.052	(-0.168, 0.035)

Table: Analysis results for the lung cancer data on natural scale

		Huang			Robust		
Weight	Variable	Estimate	SE	95% CI	Estimate	SE	95% CI
Gehan	Treatment	8.251	3.485	(1.266, 14.757)	9.195	3.308	(2.974, 15.680)
	LDH	2.763	3.465	(-3.892, 9.118)	3.516	3.285	(-2.918, 9.640)
	Age	-1.897	1.652	(-5.219, 1.406)	-2.085	1.580	(-5.353, 1.075)
Log-rank	Treatment	8.227	3.551	(1.064, 15.415)	8.998	3.267	(2.612, 15.799)
	LDH	3.588	3.559	(1.064, 15.415)	3.380	3.201	(-3.234, 9.486)
	Age	-1.232	1.677	(-5.219, 1.406)	-1.793	1.516	(-5.103, 0.993)

Some Remarks about SWOG 9509

- Lifetime medical cost significantly higher in the Paclitaxel plus Carboplatin group as compared to the Vinorelbine plus Cisplatin group
- Using results on natural scale, the average difference is $\approx \$9,100\text{USD}$
- The length of confidence interval using the rank-based estimator is about $\$1,200$ less than Huang's estimator

Some Concluding Remarks

- Proposed a robust, rank-based extension of an estimator proposed by Huang (2002)
- In simulation studies, the new estimator performed as well or better than Huang's "semi-rank-based" estimator
- Proposed a resampling scheme to approximate the sampling distribution of the coefficient estimator
- The resampling scheme performed well in simulation studies. It should work equally well for any style of calibration estimator.

What about Cost Accumulation?

Let $A(t)$ be cost accumulation at time t.

- $Y = A(T)$.
- $Y^L = A(T \wedge L)$
- $W^L = A(T \wedge L \wedge C)$
- but $W \equiv Y \cdot I(T \leq C) \neq A(T \wedge C)$

So, if cost data is collected on participants with censored failure times, then the current approach may be somewhat wasteful.

For other problems, like HIV resistance, marks may not be collected/available until failure occurs.