BIOMETRICS 56, 775–778
September 2000

Proportional Means Regression for Censored Medical Costs

D. Y. Lin

Department of Biostatistics, Box 357232,
University of Washington, Seattle, Washington 98195, U.S.A.
email: danyu@biostat.washington.edu

SUMMARY. The semiparametric proportional means model specifies that the mean function for the cumulative medical cost over time conditional on a set of covariates is equal to an arbitrary baseline mean function multiplied by an exponential regression function. We demonstrate how to estimate the vector-valued regression parameter using possibly censored lifetime costs. The estimator is consistent and asymptotically normal with an easily estimable covariance matrix. Simulation studies show that the proposed methodology is appropriate for practical use. An application to AIDS is provided.

KEY WORDS: Censoring; Cost analysis; Counting process; Economic evaluation; Health care; Survival analysis.

1. Introduction

The rising cost of health care has created serious national concern in the United States and other industrialized countries. This concern has prompted tremendous recent interest in the economic evaluation of medical care. Proper understanding of medical cost plays a crucial role in the search for cost-effective intervention/prevention strategies.

Medical cost data have now been routinely collected in clinical trials, disease registries, and health insurance records. A common feature with these available data sources is that not all patients are followed until the endpoint of interest, which is the time of death, assuming that one is interested in the lifetime cost. This is the well-known phenomenon of censoring. If a patient's survival time is censored, then his/her lifetime cost is also censored.

There is a fundamental difference between censored survival time and censored lifetime cost. Because a patient who accumulates costs over time at relatively higher rates tends to generate larger cumulative costs at both the survival time and censoring time, the cumulative cost at the survival time (i.e., the lifetime cost) is positively correlated with the cumulative cost at the censoring time (i.e., the censoring variable for the lifetime cost) even if the underlying survival time and censoring time are independent. Due to this dependent censoring, standard methods for handling censored survival data, such as the Kaplan–Meier estimator and Cox regression, cannot be used to analyze censored lifetime cost data, although this approach has been erroneously suggested in the literature.

So far, the only formal investigation of statistical methods for analyzing censored cost data is that done by Lin et al. (1997). That investigation was focused on the estimation of the mean lifetime cost for a group of patients. Currently there does not exist any valid regression method for assessing the effects of covariates (e.g., therapies and patient characteristics) on medical cost.

In the next section, we present a proportional means regression model for the cumulative medical cost and develop the corresponding semiparametric inference procedures based on the possibly censored observations of the lifetime cost. In Section 3, we report some simulation results and present an application to data from an AIDS clinical trial. A few concluding remarks are given in Section 4.

2. Regression Methodology

Let $N^*(t)$ be the cumulative cost up to time t. The cost is normally measured in financial terms, such as the amount of hospital charges, but may sometimes be measured in more basic units, such as the number of hospital admissions and days of hospitalization. Naturally, there is no further medical cost after death so that $N^*(\cdot)$ does not jump after T, where T is the survival time. Let $\mathbf{Z}$ be the set of covariates of interest. Define $\mu(t \mid \mathbf{Z}) = \mathrm{E}\{N^*(t) \mid \mathbf{Z}\}$, which is the mean cumulative cost at t conditional on the covariates. We specify that

$$\mu(t \mid \mathbf{Z}) = \mu_0(t)e^{\beta'\mathbf{Z}}, \tag{1}$$

where $\mu_0(\cdot)$ is an arbitrary baseline mean function and β is a set of unknown regression parameters.

Model (1) is referred to as the proportional means model, which is similar to the Cox proportional hazards model for survival data. This model is semiparametric in that the baseline mean function is completely unspecified while the effects of covariates are represented by a limited number of regression parameters. If $N^*(\cdot)$ represents the cumulative number of hospital admissions, then equation (1) is the counting process model studied by Lin et al. (2000), among others. Note that we do not impose any dependence structure between T and $N^*(\cdot)$ or among the increments of $N^*(\cdot)$.

Because of censoring, $N^*(\cdot)$ may not be fully observed. Let $N(t) = N^*(t \wedge C)$, where C is the censoring time and $a \wedge b = \min(a, b)$. Thus, $N(t) = N^*(t)$ if $t \leq C$ and $N(t) = N^*(C)$ if $t > C$. If $N^*(\cdot)$ pertains to hospital admissions, then $N(\cdot)$ is the familiar counting process used by Andersen and Gill (1982) and Lin et al. (2000).

Although Lin et al. (2000) confined their attention to counting processes, their theoretical development did not rely on the counting process feature of $N^*(\cdot)$. It is not difficult to verify that the results of Lin et al. (2000) hold for any nondecreasing process $N^*(\cdot)$ satisfying equation (1). To be specific, if the data consist of n independent random triplets $\{N_i(\cdot), C_i, \mathbf{Z}_i\}$ $(i = 1, \ldots, n)$, then a possible estimating function for β of model (1) is

$$\mathbf{U}^*(\beta) = \sum_{i=1}^{n} \int_0^\infty \{\mathbf{Z}_i - \bar{\mathbf{Z}}(\beta, t)\} dN_i(t), \qquad (2)$$

where $\bar{\mathbf{Z}}(\beta, t) = \Sigma_{j=1}^n I(C_j \geq t) e^{\beta' \mathbf{Z}_j} \mathbf{Z}_j / \Sigma_{j=1}^n I(C_j \geq t) \times e^{\beta' \mathbf{Z}_j}$ and $I(\cdot)$ is the indicator function. Simple algebraic manipulation yields

$$\mathbf{U}^*(\beta) = \sum_{i=1}^{n} \int_0^\infty \{\mathbf{Z}_i - \bar{\mathbf{Z}}(\beta, t)\} dM_i(t),$$

where

$$M_i(t) = \int_0^t I(C_i \geq s) \left\{ dN_i^*(s) - e^{\beta' \mathbf{Z}_i} d\mu_0(s) \right\}, \quad i = 1, \ldots, n.$$

Under model (1), $M_i(t)$ $(i = 1, \ldots, n)$ are zero-mean stochastic processes. It then follows from the arguments given in Appendix A.1–A.2 of Lin et al. (2000) that $n^{-1/2}\mathbf{U}^*(\beta)$ is asymptotically zero-mean normal and the solution to $\mathbf{U}^*(\beta) = \mathbf{0}$ is consistent and asymptotically normal.

In general, the data $\{N_i(\cdot), C_i, \mathbf{Z}_i\}$ $(i = 1, \ldots, n)$ needed for the calculation of $\mathbf{U}^*(\beta)$ are not available. First, the censoring time C_i is normally unknown if $T_i < C_i$ (i.e., the patient dies before he/she is censored), one exception being administrative censoring, i.e., censoring due to staggered patient entries and early study termination. (In the recurrent event problem studied by Lin et al. (2000) and others, it is implicitly assumed that no subject dies during the study so that the censoring time is known for every subject.) Second, the evaluation of $\mathbf{U}^*(\beta)$ requires that the sample paths (i.e., the entire histories) of $\{N_i(\cdot); i = 1, \ldots, n\}$ be known. This is usually not a problem if $N^*(\cdot)$ represents the number of hospital admissions but is unrealistic if $N^*(\cdot)$ pertains to the amount of hospital charges. One may be able to ascertain the amount of charges for each episode of hospitalization but rarely for each day. In many applications, the database may only contain information about the total cost at the last contact date, i.e., $N^*(T \wedge C)$.

Suppose that censoring arises in a completely random fashion. It is then possible to estimate β using only the data $(\tilde{N}_i, \mathbf{Z}_i)$ $(i = 1, \ldots, n)$, where $\tilde{N}_i = N_i^*(T_i \wedge C_i)$ or $N_i(\infty)$. Specifically, let $G(t) = \Pr(C \geq t)$. By replacing the indicator functions $I(C_j \geq t)$ $(j = 1, \ldots, n)$ involved in $\bar{\mathbf{Z}}(\beta, t)$ with their common expectation $G(t)$, we derive the following estimating function from (2):

$$\mathbf{U}(\beta) = \sum_{i=1}^{n} \tilde{N}_i \{\mathbf{Z}_i - \bar{\mathbf{Z}}(\beta)\},$$

where $\bar{\mathbf{Z}}(\beta) = \Sigma_{j=1}^n e^{\beta' \mathbf{Z}_j} \mathbf{Z}_j / \Sigma_{j=1}^n e^{\beta' \mathbf{Z}_j}$. The solution to $\mathbf{U}(\beta) = \mathbf{0}$ is denoted by $\hat{\beta}$. Let $\mathcal{I}(\beta) = -n^{-1} \partial \mathbf{U}(\beta) / \partial \beta'$, i.e.,

$$\mathcal{I}(\beta) = n^{-1} \sum_{i=1}^{n} \tilde{N}_i \times \sum_{j=1}^{n} e^{\beta' \mathbf{Z}_j} \{\mathbf{Z}_j - \bar{\mathbf{Z}}(\beta)\}^{\otimes 2} \Big/ \sum_{j=1}^{n} e^{\beta' \mathbf{Z}_j},$$

where $\mathbf{a}^{\otimes 2} = \mathbf{a}\mathbf{a}'$. Clearly, $\mathcal{I}(\beta)$ is positive semidefinite, which implies that $\hat{\beta}$ can be obtained by the standard Newton–Raphson algorithm.

Simple algebraic manipulation yields

$$\mathbf{U}(\beta) = \sum_{i=1}^{n} \tilde{M}_i \{\mathbf{Z}_i - \bar{\mathbf{Z}}(\beta)\},$$

where

$$\tilde{M}_i = \tilde{N}_i - e^{\beta' \mathbf{Z}_i} \int_0^\infty G(t) d\mu_0(t). \qquad (3)$$

Because $\mathrm{E}\{dN(t) \mid \mathbf{Z}\} = \mathrm{E}\{I(C \geq t) dN^*(t) \mid \mathbf{Z}\} = G(t) \times e^{\beta' \mathbf{Z}} d\mu_0(t)$, the $\tilde{M}_i$'s are zero-mean random variables. Thus, by the law of large numbers, $n^{-1}\mathbf{U}(\beta)$ converges to $\mathbf{0}$. Assume that $\mathcal{I}(\beta)$ is nonsingular in the limit. It then follows from convex analysis that $\hat{\beta}$ is a consistent estimator of β.

Let $\bar{\mathbf{z}}(\beta)$ be the limit of $\bar{\mathbf{Z}}(\beta)$. By Slutsky's theorem,

$$n^{-1/2}\mathbf{U}(\beta) = n^{-1/2} \sum_{i=1}^{n} \tilde{M}_i \{\mathbf{Z}_i - \bar{\mathbf{z}}(\beta)\} + o_p(1),$$

which is essentially a normalized sum of n independent zero-mean random vectors. Therefore, the multivariate central limit theorem implies that $n^{-1/2}\mathbf{U}(\beta)$ is asymptotically zero-mean normal with covariance matrix $\mathbf{B} = \lim_{n \to \infty} n^{-1} \Sigma_{i=1}^n \tilde{M}_i^2 \times \{\mathbf{Z}_i - \bar{\mathbf{z}}(\beta)\}^{\otimes 2}$. Consequently, $n^{1/2}(\hat{\beta} - \beta)$ is asymptotically zero-mean normal with covariance matrix $\mathbf{D} = \mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1}$, where $\mathbf{A}$ is the limit of $\mathcal{I}(\beta)$.

Naturally, $\mathbf{A}$ is estimated by $\mathcal{I}(\hat{\beta})$. Because the $\tilde{M}_i$'s given in (3) have zero means, we estimate the integral $\int_0^\infty G(t) \times d\mu_0(t)$ involved in the $\tilde{M}_i$'s by $\Sigma_{j=1}^n \tilde{N}_j / \Sigma_{j=1}^n e^{\hat{\beta}' \mathbf{Z}_j}$. It then follows from the law of large numbers that $\mathbf{B}$ can be consistently estimated by

$$\hat{\mathbf{B}} = n^{-1} \sum_{i=1}^{n} \hat{M}_i^2 \{\mathbf{Z}_i - \bar{\mathbf{Z}}(\hat{\beta})\}^{\otimes 2},$$

where

$$\hat{M}_i = \tilde{N}_i - e^{\hat{\beta}' \mathbf{Z}_i} \sum_{j=1}^{n} \tilde{N}_j \Big/ \sum_{j=1}^{n} e^{\hat{\beta}' \mathbf{Z}_j}, \qquad i = 1, \ldots, n.$$

Thus, a consistent estimator of $\mathbf{D}$ is $\hat{\mathbf{D}} = \mathcal{I}^{-1}(\hat{\beta})\hat{\mathbf{B}}\mathcal{I}^{-1}(\hat{\beta})$. The variance of $\hat{\beta}$ is estimated by $\hat{\mathbf{V}} = n^{-1}\hat{\mathbf{D}}$.

It is numerically simple to obtain $\hat{\beta}$ and $\hat{\mathbf{V}}$. If Z is a 0/1 indicator, then $U(\hat{\beta}) = 0$ yields $\hat{\beta} = \log(\bar{N}_1 / \bar{N}_0)$, where $\bar{N}_0$

Table 1
Summary statistics for the simulation studies[a]

n	σ^2	Model I				Model II			
		Bias	SE	SEE	CP	Bias	SE	SEE	CP
100	0	0.0007	0.174	0.172	0.949	0.0002	0.035	0.035	0.946
	0.25	0.0022	0.202	0.197	0.944	0.0001	0.039	0.038	0.945
	0.5	0.0030	0.226	0.218	0.940	0.0002	0.041	0.041	0.943
	1	0.0036	0.259	0.252	0.940	0.0001	0.045	0.044	0.940
200	0	−0.0003	0.122	0.122	0.950	−0.0001	0.025	0.025	0.951
	0.25	0.0011	0.142	0.141	0.946	−0.0001	0.028	0.027	0.946
	0.5	0.0009	0.159	0.157	0.942	0.0001	0.029	0.029	0.947
	1	−0.0038	0.184	0.181	0.945	−0.0002	0.031	0.031	0.948

[a] Bias is the sampling mean of $\hat{\beta}$ minus β, SE is the sampling standard error of $\hat{\beta}$, SEE is the sampling mean of the standard error estimator $\hat{V}^{1/2}$, and CP is the coverage probability of the 95% confidence interval $(\hat{\beta} \pm 1.96\hat{V}^{1/2})$. Each entry is based on 10,000 simulation samples.

and $\tilde{N}_1$ are the sample means of $\tilde{N}$ for groups 0 and 1, respectively. Thus, no iteration is required to obtain $\hat{\beta}$ in the two-sample case.

3. Numerical Results

3.1 *Simulation Studies*

We used Monte Carlo simulation to evaluate the finite-sample properties of the proposed regression methodology. Two models for medical cost were considered. For both models, the number of hospital admissions is a random-effect Poisson process that is terminated by death. Under model I, the intensity function for the underlying Poisson process is $\xi e^{\beta' \mathbf{Z}}$ and the hazard function for death is 0.1ξ, where ξ is a gamma random variable with mean one and variance σ^2; the cost is simply measured by the number of hospital admissions. Under model II, the intensity function for the Poisson process is ξ and the hazard function for death is again 0.1ξ; the cost is measured in financial terms: $\$1000e^{\beta' \mathbf{Z}}$ for each hospitalization, plus $\$5000e^{\beta' \mathbf{Z}}$ diagnostic cost at $t = 0$ and $\$10{,}000e^{\beta' \mathbf{Z}}$ final cost around the time of death. The random effect creates heterogeneity among the patients or the dependence of hospital admissions and death within the same patient. It is easy to verify that both models I and II satisfy equation (1). We generated censoring times from the uniform $[0, 10]$ distribution, corresponding to an average of 4.87 observed hospital admissions per patient.

Table 1 displays the simulation results for randomized clinical trials, in which Z is the treatment indicator with $n/2$ patients in each of the two treatments and $\beta = 0.5$. As evident from the table, the parameter estimator $\hat{\beta}$ is virtually unbiased. The variance estimator $\hat{V}$ provides accurate estimation of the true variance of $\hat{\beta}$, and the corresponding confidence intervals have proper coverage probabilities. Note that the variance of $\hat{\beta}$ is much smaller under model II than under model I.

3.2 *Application*

Cytomegalovirus (CMV) retinitis is a sight-threatening opportunistic infection affecting a great proportion of patients with acquired immunodeficiency syndrome (AIDS). Ganciclovir is a pharmacologic agent that was approved for treating CMV retinitis in the United States. Standard therapy involves lifelong intravenous (IV) infusion. There was widespread interest in an orally active agent that would ease administration and potentially reduce the cost of therapy. Thus, a clinical economic study was conducted to compare oral and IV ganciclovir for the maintenance treatment of newly diagnosed CMV retinitis (Sullivan et al., 1996). The primary outcome measures were time to first retinitis progression and associated direct medical care expenditure.

The study randomly assigned 57 and 60 patients to IV and oral therapies, respectively. By the end of the trial, 44 patients on IV therapy and 43 patients on oral therapy had experienced progression of CMV retinitis. The hazard ratio of progression for oral versus IV therapies is estimated at 1.054, associated with a p-value of 0.808. Thus, there is no significant efficacy difference between the two therapies.

As mentioned above, the primary outcome measure for economic evaluation was the direct medical care expenditure up to time of first retinitis progression. The database contains information on the observed total expenditure for each patient. To compare the costs of the two therapies, we fit model (1) in which Z indicates, by the values one versus zero, whether or not the patient received oral therapy. It is determined that $\tilde{N}_1 = \$4574.50$ and $\tilde{N}_0 = \$7737.26$. Thus, $\hat{\beta} = -0.526$. The standard error estimate $\hat{V}^{1/2}$ is 0.096. The standard-normal test statistic is -5.45, which is associated with an extremely small p-value. The 95% confidence interval for β is $(-0.715, -0.337)$.

It is more intuitive to express the treatment difference in terms of e^{β} than β since e^{β} is the ratio of the two mean functions. By exponentiation, the estimate of e^{β} is 0.591, which suggests that the oral therapy reduces the cost by about 40% compared with the standard IV administration. The 95% confidence interval for e^{β} is $(0.489, 0.714)$.

4. Discussion

Cumulative medical cost is not the only type of accumulation measure encountered in scientific studies. In clinical research, there is an increasing interest in the so-called quality-adjusted lifetime, which is a weighted sum of the numbers of days the patient lives in various health states (a consistent estimator for the distribution of quality-adjusted lifetime for a group of patients was provided by Zhao and Tsiatis (1997)). In industrial reliability, one may be interested in the mileage of an automobile or the operating hours of other machines. The regression methodology presented in this paper applies to all accumulation measures.

The proposed methodology requires that censoring arises in a completely random fashion. This assumption certainly holds for administrative censoring but may not be satisfied for voluntary patient withdrawals. For well-controlled clinical trials, voluntary withdrawals are minimal. For studies concerned with insurance payments, voluntary withdrawal is effectively treated as death rather than censoring because the costs incurred after withdrawals are of no interest. It would be possible to estimate β by allowing censoring to depend on $\mathbf{Z}$, but then the sample paths of $\{N_i(\cdot); i = 1, \ldots, n\}$ would have to be observed. In the AIDS example given in Section 3.2, censoring did not appear to depend on Z, the p-value of the log-rank test for comparing the censoring distributions of the two treatments being 0.812.

If $\{N_i(\cdot); i = 1, \ldots, n\}$ are available, then one can estimate $\mu_0(t)$ consistently by

$$\hat{\mu}_0(t) = \sum_{i=1}^{n} \int_0^t \frac{dN_i(s)}{\hat{G}(s) \sum_{j=1}^{n} e^{\hat{\beta}' \mathbf{Z}_j}},$$

where $\hat{G}(\cdot)$ is the Kaplan-Meier estimator of $G(\cdot)$. If $\{N_i(\cdot); i = 1, \ldots, n\}$ are partially observed, then $\hat{\mu}_0(\cdot)$ can be approximated. This estimator enables one to display the time pattern of cost accumulation for patients with specific covariate values. One may also use such estimators to check the proportional means assumption, as in the case of the proportional hazards model (Fleming and Harrington, 1991, pp. 173–174). For the one-sample case, Cook and Lawless (1997) suggested a different estimator of $\mu_0(t)$.

Model (1) may also be called the proportional rates model in that the covariates have multiplicative effects on the rate of accumulation. In Section 3.1, we describe two simple and realistic situations in which model (1) holds exactly. Model (1) can hold approximately and thus provide useful inference as long as the effects of each covariate on the rate of accumulation do not change drastically over time. When model (1) fails, $\hat{\beta}$ pertains to an average of the effects of the covariate on the rate of accumulation over time and $\hat{\mathbf{V}}$ remains a valid estimator for the variance of $\hat{\beta}$ (see Lin and Wei (1989) for discussion of this issue in the similar context of misspecified proportional hazards model).

After the submission of this paper, Lin (2000) developed a linear regression methodology that specifies additive rather than multiplicative covariate effects on the mean medical cost. Both of these two papers pertain to the marginal mean/rate of medical cost, which reflects the actual cost accumulated by the patient in the presence of death. The estimation of the marginal mean cost is essential to cost-effectiveness analysis and is important to public health. Naturally, the marginal measure tends to be larger if the patient survives longer. We are currently investigating semiparametric models to study the effects of covariates on the rates of cost accumulation among survivors.

ACKNOWLEDGEMENTS

This research was supported by the U.S. National Institutes of Health. The author thanks Dr Sean Sullivan for providing the AIDS data set. Helpful comments from the editor, associate editor, and two referees are greatly appreciated.

RÉSUMÉ

Dans le modèle semi-paramétrique des moyennes proportionnelles, on suppose que la fonction moyenne des coûts médicaux cumulés dans le temps, conditionnellement à un ensemble de covariables, se décompose en le produit d'une fonction moyenne d'échelle arbitraire par une fonction de régression exponentielle. Nous montrons comment estimer le vecteur des paramètres de la régression en utilisant les coûts de survie censurés. L'estimateur est consistant et asymptotiquement normal. Des études de simulation montrent que la méthode est adaptée aux utilisations courantes. On présente une application relative au SIDA.

REFERENCES

Andersen, P. K. and Gill, R. D. (1982). Cox's regression model for counting processes: A large sample study. *The Annals of Statistics* **10,** 1100–1120.

Cook, R. J. and Lawless, J. F. (1997). Marginal analysis of recurrent events and a terminating event. *Statistics in Medicine* **16,** 911–924.

Fleming, T. R. and Harrington, D. P. (1991). *Counting Processes and Survival Analysis.* New York: Wiley.

Lin, D. Y. (2000). Linear regression of censored medical costs. *Biostatistics* **1,** 35–47.

Lin, D. Y. and Wei, L. J. (1989). The robust inference for the Cox proportional hazards model. *Journal of the American Statistical Association* **84,** 1074–1078.

Lin, D. Y., Etzioni, R., Feuer, E. J., and Wax, Y. (1997). Estimating medical costs from incomplete follow-up data. *Biometrics* **53,** 419–434.

Lin, D. Y., Wei, L. J., Yang, I., and Ying, Z. (2000). Semiparametric regression for the mean and rate functions of recurrent events. *Journal of the Royal Statistical Society, Series B,* in press.

Sullivan, S. D., Mozaffari, E., Johonson, E. S., Wolitz, R., and Follansbee, S. E. (1996). An economic evaluation of oral compared with intravenous ganciclovir for maintenance treatment of newly diagnosed cytomegalovirus retinitis in AIDS patients. *Clinical Therapeutics* **18,** 546–558.

Zhao, H. and Tsiatis, A. A. (1997). A consistent estimator for the distribution of quality-adjusted survival time. *Biometrika* **84,** 339–348.

Received January 1999. *Revised July* 1999.
Accepted January 2000.