Likelihood-Based Estimation of a Proportional-Hazard, Competing-Risk Model with Grouped Duration Data

Mark Yuying An[1]

Abstract: This short paper demonstrates two important results related to the estimation of competing-risk models under the proportional-hazards assumption with grouped duration data. First I show that the model with non-parametric baseline hazards is **unidentifiable** with only grouped duration data. Therefore one has to make functional form assumption for any meaningful inference. Secondly I demonstrate that under some parametric assumption such as piecewise constant baseline hazards, the sample likelihood function has explicit analytical form. Therefore there is no need for approximation. The approximation formula adopted by Deng *et al* (2000) and religiously followed by others is only a quasi likelihood function.

1. Introduction

Following the seminal work of Han and Hausman (1990), Sueyoshi (1992), and McCall (1996), competing-risk models have become very popular in economic analysis of duration data when the duration of economic event has multiple causes of termination. For example, an unemployment spell may end either when the unemployed accepts a job offer or when s/he drops out of labor force completely. A mortgage loan contract can terminate either when the borrower defaults on the loan (surrendering the collateral and walks away) or when s/he prepays the loan completely. Each of these causes is labeled a risk. Typically, one thinks about the underlying durations associated with all the risks. It is called a competing-risk model, because the smallest realized risk-specific duration makes the durations for other risks right censored. In another word, the analyst only observes the minimal of the risk-specific durations.

The latent risk-specific durations are often modeled as continuous time variables. In a competing-risk model, the effects of observed and unobserved factors are usually accommodated using a proportional-hazard specification. This model is fit with the duration data that are grouped into intervals between integers due to the particular nature of data. Grouping is very common in economic duration data. Unemployment spells are typically measured in weeks. Mortgage payments are typically observed in monthly intervals. Fitting competing risks model with grouped duration data is a natural extension of single-risk duration analysis with grouped duration data (Prentice and Gloakler, 1978, Kiefer, 1988, Ryu 1994, An 2000, 2002).

---

1. Mark An is Director of Econometric Research at Fannie Mae. The opinion expressed in this paper does not necessary represent those of Fannie Mae's. For correspondence send email to mark_y_an@fanniemae.com or ordinary mail to Dr Mark Y An, Fannie Mae, Mail Stop 8H 203, 3900 Wisconsin Ave., NW, Washington, DC, 20016.

Recently the competing risks model under proportional hazards assumption has been applied to modeling loan performance where a mortgage can terminate by either prepayment or default (Deng *et al.*, 2000, Ciochetti *et al*. 2001, Ambrose and LaCour-Little 2001). Borrower of a mortgage decides when to stop the monthly payment of principal and interest. Prepayment refers to total payoff of the outstanding loan. Default refers to stop of payment and surrender the collateral to the lender. Prepayment is triggered either by a substantial decrease in market mortgage interest rate, by the need to take cash outs of the accumulated home equity, or by the need to move away from the current residence. Default is either triggered sudden family event (such as devoice) or made as ruthless financial decision due to negative equity in the house. In these analyses, the researcher observes monthly or quarterly loan payment histories.

This short paper is concerned with some methodological issues related to likelihood-based estimation of competing-risks models under proportional-hazards assumption with grouped duration data. I plan to clear up two common confusions.

- First, I show that the models with non-parametric baseline hazards are unidentifiable with grouped duration data. This implies that any consistent estimation and meaningful inference have to hinge on and stem from assumption of the shape of the baseline hazards. This is quite contrary to the single-risk duration case, where consistent estimation model regression coefficients do not rely on parametric assumptions of the baseline hazard even with grouped duration data.

- Second, I point out that under some parametric assumption, such as piecewise constant baseline hazards, the sample likelihood function has an explicit analytical form. An immediate implication is that there is no need for approximation once one makes the assumption that the baseline hazards are piece-wise linear.

Section 2 of the paper introduces the basic notation and describes the framework of statistical inference with grouped duration data. Section 3 discusses the main non-identification result. In Section 4, I comment on the piece-wise constant case and derive the implied analytical likelihood function. Concluding remarks are made in Section 5 with some extensions to the basic setting.

2. Competing-Risks Model under Proportional Hazards Specification

To focus on the presentation of the main ideas, I restrict my attentions to situations when

- all explanatory variables are time-invariant,

- there are only two competing risks,

- the duration variables are grouped into regular intervals quantified by positive integers,

- the heterogeneity distribution is either degenerate (non-existent), or has a known bivariate parametric distribution.

Let (T_1, T_2) be the two risk-specific (and latent) durations. Let $Y=\min\{T_1, T_2\}$ be the observable duration. Let R=1 if it is known that $Y=T_1$, R=2 if it is known that $Y=T_2$, and R=0 if both T_1 and T_2 are right-hand censored, in this case we observe some value c with the knowledge that $T_1>c$ and $T_2>c$. Let X be a vector of weakly exogenous covariates. Let $V=(V_1, V_2)$ be two unobserved heterogeneity factors. The leading example in this paper is mortgage loan termination. Specifically for this example, T_1 would be the duration until prepayment; T_2 would be the duration until default; Y would be the observed duration until the loan is terminated. If it is known that the loan is terminated due to prepayment, then R=1. If it is known that the loan is terminated due to default, then R=2. However, if by the time the survey ends, the loan of age c is still actively performing, then R=0 we say the both latent durations are right-hand censored at c.

A continuous-time competing-risks model under proportional hazard specification has the following three components:

Assumption 1 (**Conditional Independence**) Conditional on the observed and unobserved heterogeneity, (X, V), the two risk-specific durations T_1 and T_2 are independent.

Assumption 2 (Proportional Hazards) Conditional on (X, V)=(x, v), the hazard rates for T_1 and T_2 are, respectively,

$$(1) \qquad h_j(t|x, v)= \lambda_j(t) \exp\{x \beta_j+v_j\}, \quad j=1,2.$$

A**ssumption 3 (Heterogeneity Distribution)** The heterogeneity vector (V_1, V_2) is independent from X, and is distributed with a bivariate distribution function $G(v_1, v_2)$ which is either

(a) $G(v_1, v_2)$ is degenerate, i.e., $P(V_1 =0, V_2 =0) =1$, or

(b) $G(v_1, v_2; \gamma)$ has a parametric form with parameter γ.

The parameters of primary interest are the regression coeffiecients $\beta_1$and β_2 together with possibly γ in the heterogeneity distribution. Following the tradition in single-risk setting due to the seminal work of Cox (1972), it is now customary to leave the two baseline hazard functions $\lambda_1(t)$ and $\lambda_2(t)$ in (1) unspecified to enhance the robustness of estimating $\beta_1$and β_2.

In the next section I will show why this effort is not fruitful when the model is to be fit with grouped duration data. Data grouping arises when the duration variable Y is not observed in continuous time. In clinical trial studies, subject's blood sample is only tested

on scheduled time intervals. In economic analysis, the unemployment duration is only registered in weeks. In the context of competing risks model, I will first assume, with out of generality, that the duration variable is grouped in two time intervals bounded by integers. Specifically,

Assumption 4 (Data Grouping) Every observation in the entire sample can be classified in the one of the following three types of grouping:

	Explanation of the Situation	**Y Value**	**R Value**	**Knowledge of T_1 and T_2**
Type P	A loan is prepaid in Period K_n[2]	$\in (K_n -1, K_n]$	= 1	$T_1 \in (K_n -1, K_n]$ & $T_2 > T_1$
Type D	A loan defaults in Period K_n	$\in (K_n -1, K_n]$	= 2	$T_2 \in (K_n -1, K_n]$ & $T_1 > T_2$
Type C	A loan is still performing at the time of observation in period K_n	$\in (K_n -1, \infty)$	= 0	$T_1 > K_n -1$ & $T_2 > K_n -1$

These three types of data grouping are illustrated in Figure 1.

(INSERT FIGURE 1 HERE)

Using the above notation, for each individual n in the sample we have the following information (X_n, K_n, R_n), whereby the value of R_n corresponds to the whether n belongs to Type P, Type D, or Type C. Notice that the heterogeneity vector (V_1, V_2) is unobserved.

3. Non-identification

Under Assumptions 1 and 2, conditional on (X_n, V_{n1}, V_{n2}), the joint density function of (T_1, T_2) is

$$f(s,t|X_n, V_{n1},V_{n2})=h_1(s|X_n, V_{n1})\, h_2(t|X_n, V_{n2})\exp\{ - \Lambda_1(s)\phi_{1n} - \Lambda_2(t)\phi_{2n}\}, \quad (2)$$

and the joint survivor function is

2 Our convention is to name the interval (0, 1] the first period, the interval (1, 2] the second period, and so on.

(3) $S(s, t|X_n, V_{n1,}V_{n2}) \equiv P(T_1 > s, T_2 > t| X_n, , V_{n1},V_{n2})$

$= \exp\{ -\Lambda_1(s)\phi_{1n}-\Lambda_2(t)\phi_{2n}\},$

where we use $\Lambda_j(s)= \int_0^s \lambda_j(t)dt$ to denote the risk specific integrated baseline hazard, and use $\phi_{jn} =\exp\{X_n\beta_j+V_{nj}\}$ to denote the covariates effect.

To derive the likelihood function based on the model and the data set up in the previous section, it is helpful to start with a Type C observation. Such an observation contributes to the sample likelihood function in the following form,

(4)
$$P(Y_n \geq K_n - 1 | X_n) = E_G\left[P(Y_n \geq K_n - 1 | X_n, V)\right] = E_G\left[S(K_n - 1, K_n - 1 | X_n, V)\right].$$

Where the expectation is, whenever necessary if with respect to the heterogeneity distribution G(.), because V is by construction not observed.[3]

The contribution to the sample likelihood of a Type P observation is more can be derived with a little algebra.

(5) $\Pr(K_n - 1 < Y_n \leq K_n, R_n=1|X_n)$

$= E_G [\Pr(K_n - 1 < Y_n \leq K_n, R_n=1|X_n,V]$

$$= E_G\left[\int_{k-1}^{k} \lambda_1(t)\phi_{1n} \exp\{-\Lambda_1(t)\phi_{1n} - \Lambda_2(t)\phi_{2n}\}dt\right],$$

Similarly for a Type D observation, its contribution to the sample likelihood is

(6) $\Pr(K_n - 1 < Y_n \leq K_n, R_n=2|X_n)$

$$= E_G\left[\int_{k-1}^{k} \lambda_2(t)\phi_{2n} \exp\{-\Lambda_1(t)\phi_{1n} - \Lambda_2(t)\phi_{2n}\}dt\right].$$

To illustrate the fundamental non-identification, let us take the simplest case when the unobserved heterogeneity is absent. Assume Assumption 3(a). The equation (5) simplifies to

(7) $\Pr(K_n - 1 < Y_n \leq K_n, R_n=1|X_n)$

$$= \int_{k-1}^{k} \lambda_1(t)\phi_{1n} \exp\{-\Lambda_1(t)\phi_{1n} - \Lambda_2(t)\phi_{2n}\}dt .$$

3 When the heterogeneity distribution is degenerate, the expectation is trivialized.

The above integral depends on the values of $\lambda_1(t)$ and $\lambda_2(t)$ for all t between the interval (k-1, k]. With that, we have arrived at the following result.

Proposition 1. Without the parameterization of $\lambda_1(t)$ and $\lambda_2(t)$ the competing-risks model under proportional hazard specification is unidentified by grouped duration data.

This is quite a different picture from the single-risk setting. In the latter, it is well known that with grouped duration data, the likelihood function depends on the baseline hazard only through the discrete values of integrated hazard function, that is,

$$\Pr(K_n - 1 < Y_n \le K_n, R_n=1|X_n) = \int_{k-1}^{k} \lambda_1(t)\phi_{1n} \exp\{-\Lambda_1(t)\phi_{1n}\}dt$$

$$= \exp\{-\Lambda_1(k-1)\phi_{1n}\} - \exp\{-\Lambda_1(k)\phi_{1n}\}$$

Therefore provided the number of cut-off points is either fixed or grows slower to infinity than the sample size n, consistent estimation of the regression coefficients β is achievable, even without the specification of the baseline hazard function.[4]

Notice that the non-identification for the competing-risk world is purely due to data grouping. It has nothing to do with whether or not the unobserved heterogeneity is present. This identification is also qualitatively different from the non-identification concept of Tsiatis (1975), as here the un-identification arises even under conditional independence between the two risks and enough variation of the X vector.

4. Exact Solution under Piece-wise Constant Baseline Hazards

The direct implication of Proposition 1 is that it is necessary to make functional form assumption about the baseline hazards $\lambda_1(t)$ and $\lambda_2(t)$. Any meaningful inference comes from that assumption, and also hinges on that assumption.

One of the commonly used assumption is the piece-wise constant assumption, popularized after Han and Hausman (1990). In this section I comment on the piece-wise constant baseline hazard and derive the exact likelihood function associated with this assumption.

A**ssumption 5 (Piece-wise Constant Baseline Hazards)** For j=1,2, the baseline hazard function $\lambda_j(t)$ is piece-wise constant, that is, there exit constants such that

$$(8) \qquad \lambda_j(t) = \sum_{k=1}^{M} \alpha_{jk} 1_{t\in[k-1,k)}, \qquad j = 1,2,$$

where M is the total number of the distinct integers in the set $\{K_n\}$.

4 For an intuition about the non-parametric identification in the single risk setting and how to fully exploit that feature for statistical inference purpose, see An (2000).

Under Assumption 5, the two integrated baseline hazard functions are *piece-wise linear* with interval-specific slopes α_{jk}.

Proposition 2. Under Assumptions 1-5, the integral appearing in equation (5) has an analytical expression, that is,

(9)
$$\int_{k-1}^{k} \lambda_1(t)\phi_{1n} \exp\{-\Lambda_1(t)\phi_{1n} - \Lambda_2(t)\phi_{2n}\}dt$$

$$= \frac{\alpha_{1k}\phi_{1n}}{\alpha_{1k}\phi_{1n} + \alpha_{2k}\phi_{2n}} \exp\{-\Lambda_1(K_n - 1)\phi_{1n} - \Lambda_2(K_n - 1)\phi_{2n}\}[1 - \exp\{-\alpha_{1k}\phi_{1n} - \alpha_{2k}\phi_{2n}\}].$$

The result is proved by simple algebra. Because under Assumption 5,

$$\int_{k-1}^{k} \lambda_1(t)\phi_{1n} \exp\{-\Lambda_1(t)\phi_{1n} - \Lambda_2(t)\phi_{2n}\}dt$$
$$= \int_{k-1}^{k} \alpha_{1k}\phi_{1n} \exp\{-\{\Lambda_1(K_n - 1) + \alpha_{1k}[t - (K_n - 1)]\}\phi_{1n} - \{\Lambda_2(K_n - 1) + \alpha_{2k}[t - (K_n - 1)]\}\phi_{2n}\}dt$$

To gain intuition of the above expression, denote

(10)
$$\theta_n = \frac{\alpha_{1k}\phi_{1n}}{\alpha_{1k}\phi_{1n} + \alpha_{2k}\phi_{2n}}.$$

Notice that under the Assumption 5, probability that the duration ends in interval $[K_n\text{-}1, K_n)$ conditional on (X_n, V) is

(11)
$$\Pr(K_n - 1 < Y_n \le K_n | X_n, V)$$
$$= \Pr(K_n - 1 < Y_n | X_n, V) - \Pr(K_n < Y_n | X_n, V)$$
$$= \exp\{-\Lambda_1(K_n - 1)\phi_{1n} - \Lambda_2(K_n - 1)\phi_{2n}\}[1 - \exp\{-\alpha_{1k}\phi_{1n} - \alpha_{2k}\phi_{2n}\}].$$

Equation (9) and equation (11) make clear that Assumption 5 calls for a division of this probability mass according to the weights θ_n and $1 - \theta_n$ respectively.

McCall (1996) proposes an ad hoc approximation of the likelihood contribution of a Type P or Type D observation by essentially fixing $\theta_n = ½$ for all n. The corresponding formula under McCall (1996) is

$$\Pr(K_n - 1 < Y_n \le K_n | X_n, V)$$
$$= 0.5 \exp\{-\Lambda_1(K_n - 1)\phi_{1n} - \Lambda_2(K_n - 1)\phi_{2n}\}[1 - \exp\{-\alpha_{1k}\phi_{1n} - \alpha_{2k}\phi_{2n}\}].$$

In recent papers on loan performance models, Deng *et al* (2000), Ciochetti *et al*. 2001 and Ambrose and LaCour-Little 2001 for example, all adopt McCall's formula explicitly with their piece-wise constant assumption of the baseline hazards.

(1) In mortgage termination models, compared with prepayments, loan default is an extremely rare event. It is well known that default hazard rate is only a tiny fraction (1/50, say) of the prepayment hazard rate. In this case, 50-50 split of the probability is way is quite inaccurate.

(2) According to Proposition 2, the split ratio θ_n is individual specific, therefore cannot be fixed once for all for all observations.

Notice also that under Assumption 5, the joint survivor function, $S(K_n,K_n|X, V)$, depends on the baseline hazards only through the 2M discrete values of the integrated baseline hazards. Define

$$\rho_{jk} = \log[\Lambda_j(K_n) - (\Lambda_j(K_n-1)],$$

as the logarithm consecutive increments of Λ_j from k-1 to k. With this parameterization, the full parameter vector is

$$\delta = (\beta_1, \beta_2, \rho_{11}, \rho_{12,...} \rho_{1M}, \rho_{21}, \rho_{22,...,} \rho_{2M}, \gamma).$$

Estimation of δ can be carried out by maximizing the sample log likelihood function. The optimization routine depends on how the heterogeneity distribution is specified. The most convenient case is when the heterogeneity distribution.

The most convenient way to specify the heterogeneity distribution is the two-dimensional discrete distribution. For example, on a 3x3 grids, there are 15 parameters,

$\mathbf{V_1}$	$\mathbf{V_2}$	**P**
a1	b1	p11
a1	b2	p12
a1	b3	p13
a2	b1	p21
……		
a3	b3	p33

satisfying three constraints: (1) the probabilities sum to 1; (2) the mean of V_1 is zero; and (3) the mean of V_2 is zero. With these restrictions, there would only be 12 free parameters in the γ vector. If past experience is our guide, then there is unlikely a need to increase

the grid points. Typically a 2x2 grid with 8-3 =5 free parameters should be enough (An 2002).

5. Conclusions

The previous two sections delivered two main messages. First, the models with nonparametric baseline hazards are fundamentally unidentifiable with grouped duration data. When a competing-risks model is fit with grouped duration data, any meaningful inference has to stem from and hinge on parametric assumption of the baseline hazard. Second, under parametric assumption such as the piece-wise linear baseline hazards, the sample likelihood function has explicit analytical functional form. Direct estimation using the full likelihood function is feasible and easy. Under this assumption, approximation of the likelihood function is no longer necessary. Specifically, when the two risks are very different in hazard rate, the folk approximation using a 50-50 split can be very damaging.

Throughout this short paper I have limited to the case where there are two competing risks, where all the observed covariates are time-invariant, where the data grouping is regular in the sense that the continuous duration variable falls into intervals bounded by whole integers. Generalization to more than two competing risks only involves notational complication. Treatment of time-varying covariates can be typically accommodated by making assumptions that the time trajectories of the X's are also piece-wise constant whose value changes are conforming to the interval of the duration variable. Non-regular data grouping can also be easily handled without much of difficulty, just as in the case of single-risk models (An, 2002).

References

Ambrose, Brent W. and Michael LaCour-Little (2001): "Prepayment Risk in Adjustable Rate Mortgages Subject to Initial Year Discount: Some New Evidence," forthcoming in ***Real Estate Economics,* 29.**

An, Mark Y. (2000): "A Semiparametric Distribution of Willingness to Pay and Statistical Inference with Dichotomous Choice CV Data", ***American Journal of Agricultural Economics***, **82**, pp 487-500.

An, Mark Y. (2002): "Statistical Inference of Mixed Proportional Hazard Models with Grouped Data," Manuscript, Fannie Mae.

Ciochetti, Brian A., Bin Gao, and Rui Yao (2001): "The Termination of Lending Relationships through Prepayment and Default in the Commercial Mortgage Markets: A Proportional Hazard Approach with Competing Risks," Working Paper, University of North Carolina at Chapel Hill.

Cox, David R. (1972): “Regression Models and Life Tables,” ***Journal of the Royal Statistical Society,*** **Series B 34**, 187-220.

Deng, Yongheng, John M. Quigley, and Robert van Orders (2000): “Mortgage Terminations, Heterogeneity and the Exercise of Mortgage Options” ***Econmetrica***, **68,** 275-307.

Han, Aaron and Jerry Hausman (1990):”Flexible Parametric Estimation of Duration and Competing Risk Models," ***Journal of Applied Econometrics,*** **5:** 325-353.

Kiefer, Nicholas M. (1988): “Analysis of Grouped Duration Data***,” Contemporary Mathematics,*** **80**, pp. 107-137.

McCall, Brian P. (1996): “Unemployment Insurance Rules, Joblessness, and Part-Time Work,” ***Econometrica,*** **64:** 647-682.

Prentice R.L. and L.A. Gloackler (1978): “Regression Analysis of Grouped Survival Data with Application to Breast Cancer Data,” ***Biometrics,*** **34:** 57-67.

Ryu, Keunkwan (1994): “Group Duration Analysis of the Proportional Hazard Model: Minimum chi-Squared Estimators and Specification Tests," ***Journal of the American Statistical Association,*** **89**, 1386-1397.

Sueyoshi, Glenn (1992): “Semiparametrics proportional Hazards estimation of Competing Risks Models with Time-Varying Covariates," ***Journal of Econometrics,*** **51**, 25-58.

Tsiatis, A. A. (1975), A Nonidentifiability Aspect of the Problem of Competing Risks, ***Proceedings of National Academy of Sciences USA,*** **V. 72,** Page 20-2.

Figure 1 Three Types of Grouped Duration Data