Bivariate Geometric Distribution and Competing Risks: Statistical Analysis and Application

Petr Volf[1]

Abstract. The present contribution studies the statistical model for discrete time two-variate duration (time-to-event) data. The analysis is complicated by just partial data observation caused either by the right-side censoring or even by the presence of dependent competing events. The case is modeled and analyzed with the aid of a two-variate geometric distribution. The model identifiability is discussed and it is shown that the model is not identifiable without proper additional assumptions. The method of analysis is illustrated both on artificially generated example and on real unemployment data.

Keywords: bivariate geometric distribution, discrete time, event-history analysis, competing risks, unemployment data.

JEL Classification: C41, J64
AMS Classification: 62N02, 62P25

1 Introduction

In many time-to event data cases, more than one potential event can be associated with observed objects. For example, we can follow X and Y being the times to failure of a device caused by different reasons. To cope with such cases, various bivariate probability distributions have been introduced in the literature (see for example Marshall and Olkin [6] and many other sources). However, most of them were developed for continuous time cases. In the discrete-time lifetime data (often originated from continuous time processes with observations aggregated to intervals), it is assumed that the lifetimes X and Y are discrete random variables (attaining positive integer values, measured in corresponding time units). Then, the distribution of the time to event can be modeled by different variants of geometric distribution. Naturally, corresponding discrete bivariate distributions have been introduced in the literature as well, as the bivariate geometric distribution versions of Basu and Dhar [1]. However, it might be said that despite frequent discrete measuring of lifetime and other duration data, very common in applications, few papers related to discrete lifetime data appears in the literature (see for example, Grimshaw et al. [4], Davarzani et al. [3]). One of the reasons is that the continuous time models are often more "comfortable" both from the point of analysis and of theoretical knowledge (compare for instance the frequent use of the Cox regression model).

When each of followed events terminates the observation (like, for instance, a critical failure of a device) the events are competing, just the first occurring is observed. When corresponding random variables, times to these events, are independent, one variable censors (randomly, from the right side) the other. However, if they are dependent mutually, we deal with more complicated case of dependent competing risks. Discrete time case is further complicated (when compared to the continuous time setting) by potential occurrence of both events in the same time interval, though just one of them really happens. Consequently, in general, in such a case the model parameters are not identifiable. This could be easily shown by the analysis of log-likelihood, which is then evidently over-parametrized.

In the present contribution, first, a version of two-variate geometric distribution is recalled. Then, the setting of two competing risk is described, together with well know problems with model identification. The main part then studies the assumptions under which the model identifiability holds. The analysis method and the impact of additional assumptions will be illustrated on artificial data and the results discussed. Finally, a real data example will be presented.

2 Bivariate geometric distribution

Standard univariate geometric distribution corresponds to the order of the first "1" in the series of Bernoulli "0-1" attempts. Thus, when U is the Bernoulli r.v. with $P(U=1)=p$, the geometric r.v. X "starting at 1" has $P(X=k)=(1-p)^{k-1}\cdot p$, for $k=1,2,\ldots$, $\mathrm{E}X=1/p$, $\mathrm{var}(X)=(1-p)/p^2$.

[1] Department of Stochastic Informatics, ÚTIA AV ČR, Pod vodárenskou věží 4, 182 08 Praha 8, Czech Republic, volf@utia.cas.cz

Similarly, a basis for the bivariate geometric model is the bivariate Bernoulli random variable (U, V) with joint probability distribution $p_{00},\ p_{10},\ p_{01},\ p_{11}$, where $p_{jk} = P(U = j, V = k)$, j, k are 0 or 1. The marginal probabilities are $p_{1.} = P(U = 1) = p_{10} + p_{11}$, $p_{.1} = P(V = 1) = p_{01} + p_{11}$ and the covariance $\text{cov}(U, V) = p_{11} - p_{1.} \cdot p_{.1}$. Recall also that in the case $p_{11} = p_{1.} \cdot p_{.1}$, i.e. when $U,\ V$ are non-correlated, they are already independent, too.

Let us denote corresponding geometric random variables $X,\ Y$. Then, as in Basu and Dhar [1], the following form of bivariate geometric distribution will be considered, for $s, t = 1, 2, \ldots$:

$$P(X = s, Y = t) = p_{00}^{s-1} \cdot p_{10} \cdot p_{.0}^{(t-s-1)} \cdot p_{.1} \quad \text{for } s < t,$$

$$P(X = s, Y = t) = p_{00}^{t-1} \cdot p_{01} \cdot p_{0.}^{(s-t-1)} \cdot p_{1.} \quad \text{for } s > t,$$

$$P(X = s, Y = t) = p_{00}^{s-1} \cdot p_{11} \quad \text{in the case } s = t. \tag{1}$$

Covariance can be also computed easily, its sign is the same as of $\text{cov}(U, V)$. Namely

$$\text{cov}(X, Y) = (p_{10} + p_{01}) * (1 + p_{00})/(1 - p_{00})^3 + (p_{10}/p_{.1} +$$

$$+ p_{01}/p_{1.})/(1 - p_{00})^2 + p_{11}(1 + p_{00})/(1 - p_{00})^3 - 1/(p_{.1}\, p_{1.}).$$

Correlation then equals $\text{corr}(X, Y) = \text{cov}(X, Y) \cdot p_{1.}\, p_{.1} / \sqrt{(1 - p_{1.})\,(1 - p_{.1})}$.

3 Competing risks problem

The interest in the problem of mutually dependent competing risks dates back to 70-ties of the last century. Formally, there are K random variables $T_j, j = 1, \ldots, K$, running simultaneously, each representing the time to certain event. The occurrence of events, therefore also corresponding random variables, can be dependent mutually. Observation is terminated at the minimum of them. Let $\overline{F}_K(t_1, \ldots, t_K) = P(T_1 > t_1, \ldots, T_K > t_K)$ be the joint survival function of $\{T_j\}$. However, instead the 'net' survivals T_j we observe just 'crude' data (sometimes called also 'the identified minimum') $Z = \min(T_1, \ldots, T_K)$ and the indicator $\delta = j$ if $Z = T_j$. Such data allow for direct estimation of the distribution of $Z^* = \min(T_1, \ldots, T_K)$, for instance its survival function $S(t) = P(Z^* > t) = \overline{F}_K(t, \ldots, t)$.

Generally, however, from data $(Z_i, \delta_i),\ i = 1, \ldots, N$ it is not possible to identify the marginal or joint distributions of $\{T_j\}$. Tsiatis in [8] has shown that for arbitrary joint model we can find a model with independent components having the same incidences, i.e. we cannot distinguish among the models. It follows that it is necessary to make certain functional assumptions about the form of both marginal and joint distribution in order to identify them. Several such cases are specified for instance in Basu and Ghosh [2]. Later on the case of competing risks with covariates was studied by many other authors, in a more precise way, already with the aid of a copula describing the dependence. However, the results concern mostly the continuous time setting. On the contrary, our interest lies in the analysis of discrete models.

4 Bivariate geometric distribution under competing risks

As it has already been said, in the competing risks case only $T = \min(X, Y)$ is observed. However, in the discrete time setting a serious problem arises, that both X, Y may occur in the same time interval, i.e. $X = Y = t$, however just one of them is observed, the "first" one. The probability of such an instance equals $p_{11} \cdot \sum_{t=1}^{\infty} p_{00}^{t-1} = p_{11}/(1 - p_{00})$. In other words, if $X = t$ is observed, on Y we know only that $Y \geq t$ and the probability of such a result is between $p_{00}^{t-1} \cdot p_{10}$ and $p_{00}^{t-1} \cdot (p_{10} + p_{11})$. In fact, we are not able to evaluate it precisely, without an additional assumption. Let me recall here a similar problem from the analysis of discrete-time life tables with censoring. We know the number of item at the beginning of interval, and numbers of items failed or censored during. However, for accurate estimation of survival function, for instance by the Product Limit Estimate, also the order of them is necessary. There are two boundary instances, namely that censoring precedes failures, or vice versa, the reality is between (c.f. Prentice and Gloeckler [7]). In order to cope with this problem here, let us first formulate the following:

Assumption A1. Let us assume that the probabilities of observed data can be expressed with the aid of a (known) constant $C \in [0, 1]$ as

$$P(T = X = s, Y \geq T) = p_{00}^{s-1} \cdot (p_{10} + C\, p_{11}),$$

$$P(T = Y = t, X \geq T) = p_{00}^{t-1} \cdot (p_{01} + \bar{C} p_{11}).$$

Here $\bar{C} = 1 - C$.

Let us denote these two cases by an indicator δ: $\delta = 1$ in the first case, $\delta = 2$ in the second. Further, let $p_1 = p_{10} + Cp_{11}$, $p_2 = p_{01} + \bar{C}p_{11}$, $p_0 = 1 - p_1 - p_2 = p_{00}$. These probabilities are identifiable and characterize the **incidence** of really observed events.

The sense of Assumption 1 could be clarified, at least to certain extent, by following examples:

Example 1. Let us consider a case of employment data The events are "to leave voluntarily", X, or "to be fired", Y. One can imagine that one leaves the company just before being dismissed, i.e. $X = Y$, both are (Y just potentially) in the same time interval, but only X is observed. On the other hand, it is hard to imagine an opposite case that one is fired despite he already announced his decision to leave. In such a case, $C = 1$ could be a reasonable choice.

Example 2. X, Y are the times of the first (or next) goal in an ice-hockey match, of both teams. If one team scored first, still there was a potential possibility that the second team would score first, in the same interval (the same minute, say). Its chance can be estimated comparing "scoring strengths" of teams, for instance putting $C = p_{10}/(p_{10} + p_{01})$, $\bar{C} = p_{01}/(p_{10} + p_{01})$.

Nevertheless, Assumption 1 itself does not suffice to model identification, as it quantifies just relative proportions of p_{11}, not its relation to other probabilities. Let observed data consist in N independent replications of T, δ, $T = \min(X, Y)$, $\delta = 1, 2$. The likelihood of unknown probabilities p_{jk}, $j, k = 0, 1$, under these data is

$$\mathcal{L} = \prod_{i=1}^{N} \left(p_0^{(T_i - 1)} \cdot p_1^{I[\delta_i = 1]} \cdot p_2^{I[\delta_i = 2]} \right), \tag{2}$$

where $p_{00} = p_0$ and $p_1 = p_{10} + C p_{11}$, $p_2 = p_{01} + \bar{C} p_{11}$. Hence we have 3 unknown parameters related to 2 known (well estimated) incidence probabilities p_1, p_2. Therefore p_{jk} are not identifiable.

4.1 Special cases

The simplest case arises when $p_{11} = 0$, i.e. there is no chance of both events occurrence in one time interval. In such a case there is no identification problem. Notice also that then the $\mathrm{corr}(U, V)$ as well as $\mathrm{corr}(X, Y) < 0$.

Independent variables. Another particular case occurs when U, V are not correlated. They are then also independent and X, Y are independent as well. In this case Assumption 1 suffices to identification of their distribution, as there are in fact just two probabilities to be estimated, namely $p_{1.}$ and $p_{.1}$, the rest can be derived from them. This instance covers also the case of right-side random censoring: For example, when X is a variable of our interest and Y is censoring variable, we are in fact not interested in estimation of characteristics of Y. Nevertheless, we need to assume something about order of values X, Y when occurring potentially in the same time interval. The instance is often encountered when life tables are analyzed, and standardly it is assumed (as in Prentice, Gloeckler [7]) that censoring occurred at the intervals end. In our setting it means $C = 1$. Naturally, this is just an assumption, other border case can assume $C = 0$, the reality is between.

4.2 An assumption guarantying identifiability

Let us try to propose another kind of limitation to model parameters. In fact, the meaning of constant C is

$$C = P(\delta = 1 | X = Y), \quad \bar{C} = P(\delta = 2 | X = Y).$$

Using the Bayes formula, we are able to evaluate the opposite,

$$P(X = Y | \delta = 1) = \frac{P(\delta = 1 | X = Y) \cdot P(X = Y)}{P(\delta = 1)}.$$

Denote this by α. As $P(\delta = 1) = P(\delta = 1 | X < Y) \cdot P(X < Y) + P(\delta = 1 | X = Y) \cdot P(X = Y) + P(\delta = 1 | X > Y) \cdot P(X > Y) = 1 \cdot p_{10} + C \cdot P_{11} + 0$, then

$$\alpha = \frac{C p_{11}}{p_{10} + C p_{11}} \quad \text{and} \quad C = \frac{\alpha \, p_{10}}{(1 - \alpha) \, p_{11}}.$$

Quite similarly, for $\beta = P(X = Y|\delta = 2)$ we have

$$\beta = \frac{\bar{C}\, p_{11}}{p_{01} + \bar{C}\, p_{11}} \quad \text{and} \quad \bar{C} = \frac{\beta\, p_{01}}{(1-\beta)\, p_{11}}. \tag{3}$$

Another consequence is that

$$p_{11} = \frac{\alpha}{1-\alpha} p_{10} + \frac{\beta}{1-\beta} p_{01} = \alpha\, p_1 + \beta\, p_2. \tag{4}$$

Constants $\alpha,\ \beta$ characterize the proportion of hidden events $X = Y$ in observed $\delta = 1$ or $\delta = 2$, resp. In fact, as seen also from (4), assumption on knowledge of $\alpha,\ \beta$ is much stronger that just Assumption 1 on C. Nevertheless, it can be considered to be realistic, obtained from a prior knowledge, experience, for instance. Let us formulate it as:

Assumption A2. Let us assume that both constants $\alpha,\ \beta$ defined above are known.

Notice that for instance $\alpha = 1$ means that $p_{10} = 0$, all cases with $\delta = 1$ are caused by events $X = Y$. Similarly, $\beta = 1$ is equivalent to $p_{01} = 0$, in fact we then deal with degenerate cases.

Now, it is easy to show that the model is identifiable. The log-likelihood is now

$$L = \sum_{i=1}^{N} \Big((T_i - 1) \ln p_{00} + I[\delta_i = 1] \ln(p_{10}/(1-\alpha)) + I[\delta_i = 2] \ln(p_{01}/(1-\beta)) \Big), \tag{5}$$

with $p_{00} = 1 - p_{10}/(1-\alpha) - p_{01}/(1-\beta)$. Hence, there are just 2 parameters to be estimated, the rest are then obtained from them. If both $\alpha, \beta < 1$, there is no problem to get consistent estimates of $p_{10}/(1-\alpha)$, $p_{01}/(1-\beta)$, and then of all original p_{jk}, also p_{11} from (4). Naturally, a wrong specification of α, β, leads to error in estimates of p_{jk}.

Let us also return to examples 1 and 2 from above: In Example 1, with $C = 1$, we obtain that $\alpha = p_{11}/(p_{10} + p_{11})$, $\beta = 0$, and $p_{11} = \alpha \cdot p_{10}/(1-\alpha)$. Similarly for $C = 0$, i.e. $\bar{C} = 1$.

In Example 2 we obtain that $\alpha = \beta = p_{11}/(1 - p_{00})$.

4.3 Artificial example

The data were generated from the bivariate geometric model (1) with parameters $p_{00} = 0.7,\ p_{10} = 0.1,\ p_{01} = 0.15,\ p_{11} = 0.05$. Further, we selected $C = 0.5$ which yields $\alpha = 0.2000$ and $\beta = 0.1429$, from (3). These values were taken as known constants, in accord with Assumption 2. The estimation followed the standard MLE scheme, with the aid of the Newton-Raphson iteration. As it uses the first and second derivatives of log-likelihood, then the Fisher information matrix, and, consequently, asymptotic variances of estimates, can be estimated, too.

Two results are displayed below, with $N = 100$ and 1000 generated values.

For $N = 100$:
Estimated $p_{00},\ p_{10},\ p_{01},\ p_{11} = 0.6732,\ 0.0967,\ 0.1765,\ 0.0536$.
Their asymptotic standard deviations were estimated as 0.0301, 0.0150, 0.0242, 0.0048.
Further, estimated covariance and correlation of X and Y was $\text{cov}(X, Y) = 1.6816$, $\text{corr}(X, Y) = 0.0719$, while the real 'true' values were 2.2222, 0.0808, resp.

For $N = 1000$ the following values were obtained:
Estimates $p_{00},\ p_{10},\ p_{01},\ p_{11} = 0.7032,\ 0.0969,\ 0.1506,\ 0.0493$,
with asymptotic standard deviations estimated as 0.0085, 0.0045, 0.0065, 0.0014. Finally, covariance and correlation of X and Y were estimated as $\text{cov}(X, Y) = 2.3153$, $\text{corr}(X, Y) = 0.0819$.

It is possible to say that the precision of estimates is rather good and increases with growing extent of data, N. In both cases the values of (2-dimensional here) score function, i.e. the first derivatives of the log-likelihood, which should be 0 at the log-likelihood maximum, were of order $1.0e - 004$.

Summary. In studied setting we are able to estimate consistently just incidence probabilities $p_1,\ p_2$ corresponding to observed events $\delta = 1$ or 2, resp., hence also $p_{00} = p_0 = 1 - p_1 - p_2$. Further, we know that $p_1 + p_2 = p_{10} + p_{01} + p_{11}$.

Under Assumption 1 $p_1 = p_{10} + C \cdot p_{11}$, $p_2 = p_{01} + \bar{C} \cdot p_{11}$, however even the knowledge of C does not suffice to reliable estimation of p_{jk}, except in the case of independent variables U, V. Just under much stronger Assumption 2 we have that $p_{10} = (1 - \alpha) \cdot p_1$, $p_{01} = (1 - \beta) \cdot p_2$, and $p_{11} = \alpha\, p_1 + \beta\, p_2$.

A wrong specification of α, β leads to biased estimation of p_{jk} (except p_{00}), while incidence probabilities p_1, p_2 are available directly from the maximization of likelihood (2). Simultaneously they suffice to replication of competing risk data, without knowledge of other parameters.

5 Application

Han and Hausman [5] have analyzed the data on unemployment duration, namely the records on 1051 people, collected there in Table III (with several insignificant misprints which were corrected). The time is discrete, as the information was gathered on a week basis, for 70 weeks. More about data origin can be found in the paper [5]. The data show certain non-regularities, visible also in Figures 1 and 2 below, for instance significantly larger numbers of events in 26-th week which is related to a change of support after the first half-year of unemployment. In general, the censoring (exit from the study) is caused mostly by the termination of unemployment insurance benefits (it concerns to weeks 39 and 52, too), and, naturally, by the end of study. Han and Hausman had also an information on several covariates which was not available to us. They used a discrete version of Cox regression model, with certain not fully justified approximations (e.g. substituting Gauss distribution instead the Gumbel one). The other problem are ties of events limiting the correct use of continuous time model. This problem, which we have discussed in previous sections, has been omitted.

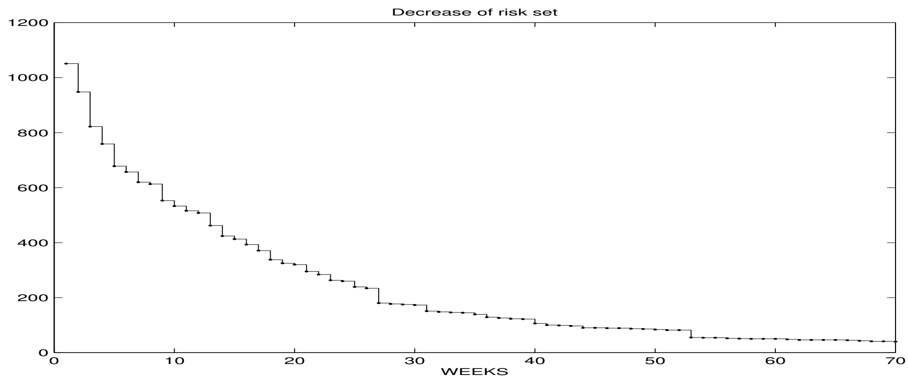

Figure 1 Decrease of number of persons at risk set

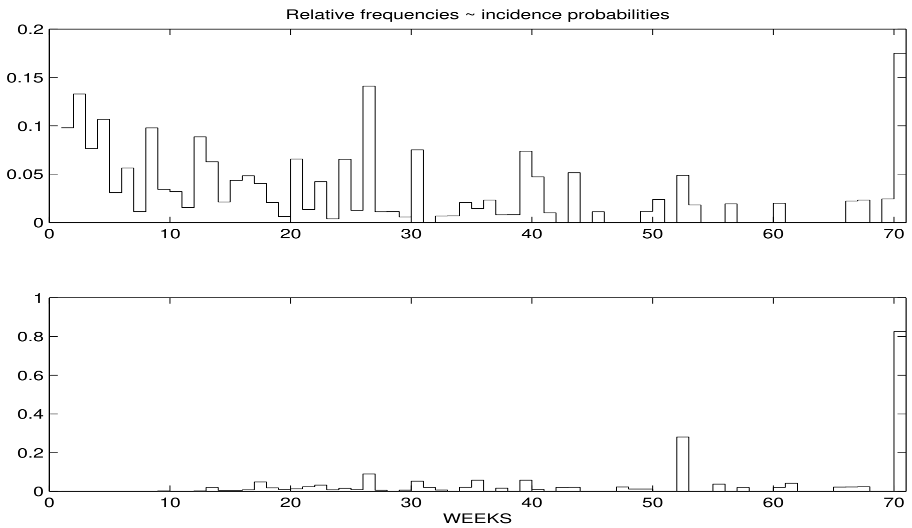

Figure 2 Incidence probabilities estimated by relative frequencies, of re-employment (above), censoring (below).

We have concentrated here to the analysis of unemployment termination and censoring as two independent competing events. Further, it was assumed that censoring occurred at the intervals end, hence constant $C = 1$. From this point of view, the situation was simplified and leading to a unique solution. On the other hand, distributions of Bernoulli random variables $U = U(t)$ for re-employment and $V = V(t)$ for censoring depended on time – weeks from 1 till $T = 70$. Let us denote N_t the number of people staying in the study at the t-th week beginning. Thus, $N_1 = 1051,\ N_{t+1} = N_t - n_t - m_t,$, where $n_t,\ m_t$ are the numbers of persons re-employed and censored, respectively, in t-th week.

First, the incidence probabilities were estimated as $p_1(t) = n_t/N_t$, $p_2(t) = m_t/N_t$ for variables $U(t),\ V(t)$, respectively. These estimates are displayed in Figure 2, while Figure 1 shows the decrease of N_t. In the next step, estimates of marginal probabilities $p_{1.}(t)$ of $U(t)$ and $p_{.1}(t)$ of $V(t)$ were computed. Due the assumptions of independence and of $C = 1$, it holds that $p_1(t) = p_{10}(t) + p_{11}(t) = p_{1.}(t)p_{.0}(t) + p_{1.}(t)p_{.1}(t) = p_{1.}(t)$ and $p_2(t) = p_{01}(t) = p_{0.}(t)p_{.1}(t) = (1 - p_{1.}(t))p_{.1}(t)$. Then

$$p_{1.}(t) = p_1(t),\ \ p_{.1}(t) = p_2(t)/(1 - p_1(t)).$$

On the basis of these results, distributions of two independent random variables, the time to re-employment, X, and the time to censoring, Y, were derived easily. Finally, we can construct a distribution of probabilities of competing events, i.e. of being re-employed at s, before censoring, or being censored at t, still unemployed. They are

$$P(X = s, Y \geq s) = \prod_{j=1}^{s-1}(p_{0.}(j)p_{.0}(j)) \cdot p_{1.}(s),$$

$$P(Y = t, X > t) = \prod_{j=1}^{t-1}(p_{0.}(j)p_{.0}(j)) \cdot p_{0.}(t) \cdot p_{.1}(t).$$

6 Concluding remarks

It has to be said that the presented results on identifiability under mutually dependent competing risks and discrete time are not satisfactory. Potential occurrence of unobserved events $X = Y$ is the cause of problems which were overcame just with rather strong assumptions. Naturally, the basic model with constant probabilities can be generalized. Thus, the final example considered time-dependent probabilities. In another setting, the probabilities can depend on explanation variables, covariates, for instance via the logistic regression model. In the continuous time cases, it was proved that the presence of covariates and assumption of a regression model type (e.g. the Cox one) can support the model identifiability (see e.g. Volf [9] for references). However, as the discrete time allows for more events at the same time interval, the situation is worse. In fact, we were not able to show any facilitation caused by the information provided by covariates.

Acknowledgements

The research was supported by the grant No. 18-02739S of the Grant Agency of the Czech Republic.

References

[1] Basu, A. P. & Dhar, S. (1995). Bivariate geometric distribution. *Journal of Applied Statistical Science*, 2, 33–44.

[2] Basu, A. P. & Ghosh, J. K. (1978). Identifiability of the multinormal and other distributions under competing risks model. *Journal of Multivariate Analysis*, 8, 413–429.

[3] Davarzani, N., Achcar, J. A., Smirnov, E. N. & Peeters, R. (2015). Bivariate lifetime geometric distribution in presence of cure fractions. *Journal of Data Science*, 13, 755–770.

[4] Grimshaw, S. D., McDonald, J., McQueen, G. R. & Thorley, S. (2005). Estimating hazard functions for discrete lifetimes. *Communications in Statistics, Part B-Simulation and Computation*, 34, 451–463.

[5] Han, A. & Hausman J. A. (1990). Flexible parametric estimation of duration and competing risk models. *J. of Applied Econometrics*, 5, 1–28.

[6] Marshall, A. W. & Olkin, I. (1985). A family of bivariate distributions generated by the bivariate bernoulli distribution. *Journal of the American Statistical Association*, 80, 332–338.

[7] Prentice, R. & Gloeckler, L. (1978). Regression analysis of grouped survival data with application to the breast cancer data. *Biometrics*, 34, 57–67.

[8] Tsiatis, A. (1975). A nonidentifiability aspects of the problem of competing risks. textitProc. Nat. Acad. Sci. USA, 72, 20–22.

[9] Volf, P. (2018). Problem of competing risks with covariates: Application to an unemployment study. In *Proceedings of the 36th International Conference Mathematical Methods in Economics, 2018.* FM VŠE Jindřichuv Hradec, 624–629.

Mendel University in Brno
Faculty of Business and Economics

Svatopluk Kapounek
Hana Vránová (eds.)

38th INTERNATIONAL CONFERENCE ON MATHEMATICAL METHODS IN ECONOMICS (MME 2020)

Conference Proceedings

September 9–11, 2020
Mendel University in Brno
Czech Republic

Organisation Committee:

Technical Editors:

ISBN 978-80-7509-734-7