Microeconometrics

Edited by

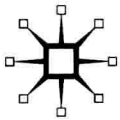
Steven N. Durlauf and
Lawrence E. Blume

Microeconometrics

Edited by
Steven N. Durlauf
University of Wisconsin-Madison, USA

Lawrence E. Blume
Cornell University, USA

palgrave
macmillan



All articles first published in *The New Palgrave Dictionary of Economics*, 2nd Edition
Edited by Steven N. Durlauf and Lawrence E. Blume
in eight volumes, 2008

with the exception of Selection Bias and Self-Selection which first published in
The New Palgrave: A Dictionary of Economics
Edited by John Eatwell, Murray Milgate and Peter Newman
in four volumes, 1987

First published 2010 by
PALGRAVE MACMILLAN

Palgrave Macmillan in the UK is an imprint of Macmillan Publishers Limited, registered in England, company number 785998, of Houndmills, Basingstoke, Hampshire RG21 6XS.

Palgrave Macmillan in the US is a division of St Martin's Press LLC, 175 Fifth Avenue, New York, NY 10010.

Palgrave Macmillan is the global academic imprint of the above companies and has companies and representatives throughout the world.

ISBN 978-0-230-23880-0 hardback
ISBN 978-0-230-23881-7 paperback

This book is printed on paper suitable for recycling and made from fully managed and sustained forest sources. Logging, pulping and manufacturing processes are expected to conform to the environmental regulations of the country of origin.

A catalogue record for this book is available from the British Library.

A catalog record for this book is available from the Library of Congress.

Printed and bound in Great Britain by
CPI Antony Rowe, Chippenham and Eastbourne

List of Contributors

ALBERTO ABADIE
Harvard Kennedy School, USA

JOSHUA D. ANGRIST
Massachusetts Institute of Technology, USA

BADI H. BALTAGI
University of Arizona, USA

MOSHE BUCHINSKY
University of California Los Angeles, USA

A. COLIN CAMERON
University of California Davis, USA

TIMOTHY G. CONLEY
University of Chicago, USA

JOHN DINARDO
University of Michigan, USA

JEFF DOMINITZ
Carnegie Mellon University, USA

DAVID DRAPER
University of California Santa Cruz, USA

JEAN-MARIE DUFOUR
McGill University, USA

ANDREW GELMAN
Columbia University, USA

JINYONG HAHN
University of California Los Angeles, USA

VASSILIS A. HAJIVASSILIOU
London School of Economics, UK

JERRY A. HAUSMAN
Massachusetts Institute of Technology, USA

JAMES J. HECKMAN
The University of Chicago, USA

KEISUKE HIRANO
University of Arizona, USA

CHENG HSIAO
University of Southern California, USA

GUIDO W. IMBENS
University of California Berkeley, USA

YANNIS M. IOANNIDES
Tufts University, USA

EKATERINI KYRIAZIDOU
University of California Los Angeles, USA

EDWARD E. LEAMER
University of California Los Angeles, USA

BRUCE G. LINDSAY
Penn State University, USA

OLIVER B. LINTON
London School of Economics, UK

JOHN A. LIST
University of Chicago, USA

THIERRY MAGNAC
University of Toulouse 1, France

CHARLES F. MANSKI
Northwestern University, USA

ROSA L. MATZKIN
University of California Los Angeles, USA

SALVADOR NAVARRO
University of Madison-Wisconsin, USA

JAMES L. POWELL
University of California Berkeley, USA

DAVID REILEY
University of Arizona, USA

ERIC RENAULT
University of North Carolina Chapel Hill, USA

JEAN-MARC ROBIN
University of Paris 1, France

DONALD B. RUBIN
Harvard University, USA

ROBERT P. SHERMAN
California Institute of Technology, USA

MICHAEL STEWART
University of Sydney, Australia

CHRISTOPHER TABER
The University of Wisconsin-Madison, USA

ELIE TAMER
Northwestern University, USA

PETRA E. TODD
University of Pennsylvania, USA

GERARD J. VAN DEN BERG
University of Amsterdam, The Netherlands

WILBERT VAN DER KLAAUW
Federal Reserve Bank of New York, USA

ARTHUR VAN SOEST
Tilburg University, The Netherlands

TIEMEN M. WOUTERSEN
Johns Hopkins University, USA

General Preface

All economists of a certain age remember the "little green books". Many own a few. These are the offspring of *The New Palgrave: A Dictionary of Economics*; collections of reprints from *The New Palgrave* that were meant to deliver at least a sense of the *Dictionary* into the hands of those for whom access to the entire four volume, four million word set was inconvenient or difficult. *The New Palgrave Dictionary of Economics, Second Edition* largely resolves the accessibility problem through its online presence. But while the online search facility provides convenient access to specific topics in the now eight volume, six million word *Dictionary of Economics*, no interface has yet been devised that makes browsing from a large online source a pleasurable activity for a rainy afternoon. To our delight, *The New Palgrave*'s publisher shares our view of the joys of dictionary-surfing, and we are thus pleased to present a new series, the "little blue books", to make some part of the *Dictionary* accessible in the hand or lap for teachers, students, and those who want to browse. While the volumes in this series contain only articles that appeared in the 2008 print edition, readers can, of course, refer to the online *Dictionary* and its expanding list of entries.

The selections in these volumes were chosen with several desiderata in mind: to touch on important problems, to emphasize material that may be of more general interest to economics beginners and yet still touch on the analytical core of modern economics, and to balance important theoretical concerns with key empirical debates. The 1987 Eatwell, Milgate and Newman *The New Palgrave: A Dictionary of Economics* was chiefly concerned with economic theory, both the history of its evolution and its contemporary state. The second edition has taken a different approach. While much progress has been made across the board in the 21 years between the first and second editions, it is particularly the flowering of empirical economics which distinguishes the present interval from the 61 year interval between Henry Higgs' *Palgrave's Dictionary of Political Economy* and *The New Palgrave*. It is fair to say that, in the long run, doctrine evolves more slowly than the database of facts, and so some of the selections in these volumes will age more quickly than others. This problem will be solved in the online *Dictionary* through an ongoing process of revisions and updates. While no such solution is available for these volumes, we have tried to choose topics which will give these books utility for some time to come.

Steven N. Durlauf
Lawrence E. Blume

Introduction

The 1987 edition of *The New Palgrave* came at a time of some of the most extraordinary post-war developments in microeconometrics, namely work by James Heckman on self-selection and Daniel McFadden on discrete choice. Heckman's entry from the 1987 edition is preserved as a classic in this volume and is joined by an entry on the Roy model which reflects Heckman's subsequent thinking. Coverage of discrete choice has been completely replaced with entries that are included here.

The successes of microeconometrics as of 1987 by no means imply that area has been one of comparative stasis. Much of Heckman's research, for example, has focused on exploring how one can develop inferences which avoid theoretically unmotivated assumptions, just as McFadden's research has focused on exploring how economic theory can be translated into econometric specifications of behaviour. These overarching ways of asking questions continue to generate important advances. One broad area of this type is semiparametric estimation, in which functional forms assumptions are relaxed in statistical analysis. An equally important area of this type is partial identification, which may be thought of as asking what may be learned from data under the most minimal assumptions. Further, new approaches to data acquisition such as survey analysis or the quest for finding interesting natural experiments, complement the methodological advances.

While microeconometrics is widely admired for the sustained pace of advances, there continue to be deep methodological disputes within the field. These are very much manifested in the literature on treatment effects, which now receives very extensive coverage. These disagreements reflect the different beliefs about the role of economic theory in empirical work, both in terms of how empirical exercises should be structured and in terms of the meaning of "statistical" assumptions. We believe this collection communicates the excitement of continuing research on micro-econometrics.

Steven N. Durlauf
Lawrence E. Blume

Contents

categorical data

Categorical outcome models are regression models for a dependent variable that is a discrete variable recording in which of two or more categories, usually mutually exclusive, an outcome of interest lies.

Categorical outcome models are also called discrete outcome models or qualitative response models, and are examples of a limited dependent variable model. Different models specify different functional forms for the probabilities of each category. These models are binomial or multinomial models, usually estimated by maximum likelihood.

Key early econometrics references include McFadden (1974), Amemiya (1981), Manski and McFadden (1981) and Maddala (1983). For textbook treatments see Amemiya (1985), Wooldridge (2002), Greene (2003) and Cameron and Trivedi (2005). The recent econometrics literature has focused on semiparametric estimation (see Pagan and Ullah, 1999) and on simulation-based estimation of multinomial models (see Train, 2003).

Binary outcomes: logit and probit models

Binary outcomes provide the simplest case of categorical data, with just two possible outcomes. An example is whether or not an individual is employed and whether or not a consumer makes a purchase.

For binary outcomes the dependent variable y takes one of two values, for simplicity coded as 0 or 1. If $y_i = 1$ with probability p_i, then necessarily $y_i = 0$ with probability $1 - p_i$, where i denotes the i^{th} of N observations. Regressors $\mathbf{x}_i$ are introduced by parameterizing the probability p_i, with

$$p_i = \Pr[y_i = 1|\mathbf{x}_i] = F(\mathbf{x}_i'\beta),$$

where $F(\cdot)$ is a specified function and a single-index form is assumed.

The obvious choice of $F(\cdot)$ is a cumulative distribution function (CDF) since this ensures that $0 < p_i < 1$. The two standard models are the logit model with $p_i = \Lambda(\mathbf{x}_i'\beta) = e^{\mathbf{x}_i'\beta}/(1 + e^{\mathbf{x}_i'\beta})$, where $\Lambda(z) = e^z/(1 + e^z)$ is the logistic CDF, and the probit model with $p_i = \Phi(\mathbf{x}_i'\beta)$, where $\Phi(\cdot)$ is the standard normal CDF.

Interest usually lies in the marginal effect of a change in regressor on the probability that $y = 1$. For the r^{th} regressor, $\partial p_i/\partial x_{ir} = F'(\mathbf{x}_i'\beta)\beta_r$ where F' denotes the derivative of F. The sign of β_r gives the sign of the marginal effect, if F is a continuous CDF since then $F' > 0$, though the magnitude depends on the point of evaluation $\mathbf{x}_i$. Common methods are to report the average marginal effect over all observations or to report the marginal effect evaluated at $\mathbf{x}$.

Parameter estimates are usually obtained by maximum likelihood (ML) estimation. Given p_i, the density can be conveniently expressed as $f(y_i) = p_i^{y_i}(1 - p_i)^{1-y_i}$. On the

assumption of independence over i, the resulting log-likelihood function is

$$\ln L(\beta) = \sum_{i=1}^{N} \{y_i \ln F(\mathbf{x}_i'\beta) + (1 - y_i)\ln(1 - F(\mathbf{x}_i'\beta))\}.$$

It can be shown that consistency of the ML estimator requires only that $p_i = F(\mathbf{x}_i'\beta)$, that is, that the functional form for the conditional probability is correctly specified.

There is usually little difference between the predicted probabilities obtained by probit or logit, except for very low and high probability events. For the logit model $\ln[p_i/(1 - p_i)] = \mathbf{x}_i'\beta$, so that β_r gives the marginal effect of a change in x_{ir} on the log-odds ratio, a popular interpretation in the biostatistics literature.

A simpler method for binary data is OLS regression of y_i on $\mathbf{x}_i$, with White heteroskedastic robust standard errors used to control for the intrinsic heteroskedasticity in binary data. A serious defect is that OLS permits predicted probabilities to lie outside the (0, 1) interval. But it can be useful for exploratory analysis, as OLS coefficients can be directly interpreted as marginal effects and standard methods then exist for complications such as endogenous regressors.

When one of the outcomes is uncommon, surveys may over-sample that outcome. For example, a survey of transit use may be taken at bus stops to over-sample bus riders. This is a leading example of choice-based sampling. Standard ML estimators are inconsistent and instead one must use alternative estimators such as appropriately weighted ML.

The preceding discussion presumes knowledge of F. A considerable number of semiparametric estimators that provide consistent estimates of β given unknown F have been proposed. Manski's (1975) smooth maximum score estimator was a very early example of semiparametric estimation.

Index models

Define a latent (or unobserved) variable y_i^* that measures the propensity for the event of interest to occur. If y_i^* crosses a threshold, normalized to be zero, then the event occurs and we observe $y_i = 1$ if $y_i^* > 0$ and $y_i = 0$ if $y_i^* \leq 0$. If $y_i^* = \mathbf{x}_i'\beta + u_i$, then

$$p_i = \Pr[y_i^* > 0] = \Pr[-u_i < \mathbf{x}_i'\beta] = F(\mathbf{x}_i'\beta),$$

where $F(\cdot)$ is the CDF of $-u_i$.

The logit model arises if u_i has the logistic distribution. The probit model arises if u_i has the more obvious standard normal distribution, where imposing a unit error variance ensures model identification. The probit model ties in nicely with the Tobit model, where more data are available and we actually observe $y_i = y_i^*$ when $y_i^* > 0$. And it extends naturally to ordered multinomial data.

Random utility models

In many economics applications the binary outcome is determined by individual choice, such as whether or not to work. Then the outcome should be the alternative

with highest utility. The additive random utility model (ARUM) specifies the utility for individual i of alternative j to be $U_{ij} = \mathbf{x}_{ij}'\beta_j + \varepsilon_{ij}$, $j = 0, 1$, where the error term captures factors known by the decision-maker but not the econometrician. Then

$$p_i = \Pr[U_{i1} > U_{i0}] = \Pr[(\varepsilon_{i0} - \varepsilon_{i1}) \leq \mathbf{x}_{i1}'\beta_1 - \mathbf{x}_{i0}'\beta_0] = F(\mathbf{x}_{i1}'\beta_1 - \mathbf{x}_{i0}'\beta_0)$$

where F is the CDF of $(\varepsilon_{i0} - \varepsilon_{i1})$. For components x_{ir} of $\mathbf{x}_i$ that vary across alternatives (so $x_{i0r} \neq x_{i1r}$) it is common to restrict $\beta_{0r} = \beta_{1r} = \beta_r$. For components x_{ir} of $\mathbf{x}_i$ that are invariant across alternatives (so $x_{i0r} = x_{i1r}$) only the difference $\beta_{1r} - \beta_{0r}$ is identified.

The probit model arises, after rescaling, if ε_{i0} and ε_{i1} are i.i.d. standard normal. The logit model arises if ε_{i0} and ε_{i1} are i.i.d. type 1 extreme value distributed with density $f(\varepsilon) = e^{-\varepsilon}\exp(-e^{-\varepsilon})$. The latter less familiar distribution provides more tractable results when extended to multinomial models.

Multinomial outcomes

Multinomial outcomes occur when there are more than two categorical outcomes. With m outcomes the dependent variable y takes one of m mutually exclusive values, for simplicity coded as $1, \ldots, m$. Let p_j denote the probability that the j^{th} outcome occurs. The multinomial density for y can be written as $f(y) = \prod_{j=1}^{m} p_j^{y_j}$ where y_j, $j = 1, \ldots, m$, are m indicator variables equal to 1 if $y = j$ and equal to 0 if $y \neq j$. Introducing a further subscript for the i^{th} individual and assuming independence over i yields log-likelihood

$$\ln L(\beta) = \sum_{i=1}^{N} \sum_{j=1}^{m} y_{ij} \ln p_{ij},$$

where the probabilities p_{ij} are modelled to depend on regressors and unknown parameters β.

There are many different multinomial models, corresponding to different parameterizations of p_{ij}.

Unordered multinomial models

Usually the outcomes are unordered, such as in choice of transit mode to work. The benchmark model for unordered outcomes is the multinomial logit model. When regressors vary across alternatives (such as prices), the conditional logit (CL) model specifies $p_{ij} = e^{\mathbf{x}_{ij}'\beta} / \sum_{k=1}^{m} e^{\mathbf{x}_{ij}'\beta}$. If regressors are invariant across alternatives (such as gender), the multinomial logit (MNL) model specifies $p_{ij} = e^{\mathbf{x}_i'\beta_j} / \sum_{k=1}^{m} e^{\mathbf{x}_i'\beta_k}$, with a normalization such as $\boldsymbol{\beta}_1 = 0$ to ensure identification. In practice some regressors may be a mix of invariant and varying across alternatives; such cases can be re-expressed as either a CL or MNL model.

The CL and MNL models reduce to a series of pairwise choices that do not depend on the other choices available. For example, the choice between use of car or red bus is not affected by whether another alternative is a blue bus (essentially the same as the

red bus). This restriction, called the assumption of independence of irrelevant alternatives, has led to a number of alternative models.

These models are based on the ARUM. Suppose the j^{th} alternative has utility $U_{ij} = \mathbf{x}_{ij}'\beta + \varepsilon_{ij}$, $j = 1, \ldots, m$. Then

$$p_{ij} = \Pr[U_{ij} \geq U_{ik} \text{ for all } k] = \Pr[(\varepsilon_{ik} - \varepsilon_{ij}) \leq (\mathbf{x}_{ij}'\beta - \mathbf{x}_{ik}'\beta) \ \forall \ k].$$

The CL and MNL models arise if the errors ε_{ij} are i.i.d. type 1 extreme value distributed. More general models permit correlation across alternatives j in the errors ε_{ij}.

The most tractable model with error correlation is a nested logit model. This arises if the errors are generalized extreme value distributed. This model is simple to estimate but suffers from the need to specify a particular nesting structure.

The richer multinomial probit model specifies the errors to be m–dimensional multivariate normal with $(m + 1)$ restrictions on the covariances to ensure identification. In practice it has proved difficult to jointly estimate both β and the covariance parameters in this model. A recent popular model is the random parameters logit model. This begins with a multinomial logit model but permits the parameters β to be normally distributed. For these two models there is no closed form expression for the probabilities and estimation is usually by simulation methods or Bayesian methods.

Ordered multinomial models

In some cases the outcomes can be ordered, such as health status being excellent, good, fair or poor.

The starting point is an index model, with single latent variable, $y_i^* = \mathbf{x}_i'\beta + u_i$. As y^* crosses a series of increasing unknown thresholds we move up the ordering of alternatives. For example, for $y^* > \alpha_1$ health status improves from poor to fair, for $y^* > \alpha_2$ it improves further to good, and so on. For the ordered logit (probit) model the error u is logistic (standard normal) distributed.

An alternative model is a sequential model. For example, one may first decide whether or not to go to college ($y = 1$) and if chose college then choose either two-year college ($y = 2$) or four-year college ($y = 3$). The two decisions may be modelled as separate logit or probit models.

A special case of ordered categorical data is a count, such as number of visits to a doctor taking values 0, 1, 2, ... An ordered model can be applied to these data, but it is better to use count models. The simplest count model is Poisson regression with exponential conditional mean $\mathrm{E}[y_i|\mathbf{x}_i] = \exp(\mathbf{x}_i'\beta)$. Common procedures are to use the Poisson but obtain standard errors that relax the Poisson restriction of variance-mean equality, to estimate the richer negative binomial model, or to estimate hurdle or two-part models or with-zeroes models that permit the process determining zero counts to differ from that for positive counts.

Multivariate outcomes and panel data

Multivariate discrete data arise when more than one discrete outcome is modelled. The simplest example is bivariate binary outcome data. For example, we may seek to explain both employment status (work or not work) and family status (children or no children). The standard model is a bivariate probit model that specifies an index model for each dependent variable with normal errors that are correlated. Such models can be extended to permit simultaneity.

For panel binary data the standard model is an individual specific effects model with $p_{it} = F(\alpha_i + \mathbf{x}_{it}'\beta)$ where α_i is an individual specific effect. The random effects model usually specifies $\alpha_i \sim N[0, \sigma_\alpha^2]$ and is estimated by numerically integrating out α_i using Gaussian quadrature. The fixed effects model treats α_i as a fixed parameter. In short panels with few time periods consistent estimation of $\boldsymbol{\beta}$ is possible in the fixed effects logit but not the fixed effects probit model. If $\mathbf{x}_{it}$ includes $y_{i,t-1}$, a dynamic model, fixed effects logit is again possible but requires four periods of data.

A. COLIN CAMERON

See also **logit models of individual choice; maximum score methods; semiparametric estimation; simulation-based estimation.**

Bibliography

Amemiya, T. 1981. Qualitative response models: a survey. *Journal of Economic Literature* 19, 1483–536.

Amemiya, T. 1985. *Advanced Econometrics*. Cambridge, MA: Harvard University Press.

Cameron, A. and Trivedi, P. 2005. *Microeconometrics: Methods and Applications*. Cambridge: Cambridge University Press.

Greene, W. 2003. *Econometric Analysis*, 5th edn. Upper Saddle River, NJ: Prentice-Hall.

Maddala, G. 1983. *Limited-Dependent and Qualitative Variables in Econometrics*. Cambridge: Cambridge University Press.

Manski, C. 1975. The maximum score estimator of the stochastic utility model of choice. *Journal of Econometrics* 3, 205–28.

Manski, C. and McFadden, D., eds. 1981. *Structural Analysis of Discrete Data with Econometric Applications*. Cambridge, MA: MIT Press.

McFadden, D. 1974. Conditional logit analysis of qualitative choice behavior. In *Frontiers in Econometrics*, ed. P. Zarembka. New York: Academic Press.

Pagan, A. and Ullah, A. 1999. *Nonparametric Econometrics*. Cambridge: Cambridge University Press.

Train, K. 2003. *Discrete Choice Methods with Simulation*. Cambridge: Cambridge University Press.

Wooldridge, J. 2002. *Econometric Analysis of Cross Section and Panel Data*. Cambridge, MA: MIT Press.

competing risks model

A competing risks model is a model for multiple durations that start at the same point in time for a given subject, where the subject is observed until the first duration is completed and one also observes which of the multiple durations is completed first.

The term 'competing risks' originates in the interpretation that a subject faces different risks i of leaving the state it is in, each risk giving rise to its own exit destination, which can also be denoted by i. One may then define random variables T_i describing the duration until risk i is materialized. Only the smallest of all these durations $Y := \min_i T_i$ and the corresponding actual exit destination, which can be expressed as $Z := \text{argmin}_i T_i$, are observed. The other durations are censored in the sense that all is known is that their realizations exceed Y. Often those other durations are latent or counterfactual, for example if T_i denotes the time until death due to cause i.

In economics, the most common application concerns individual unemployment durations. One may envisage two durations for each individual: one until a transition into employment occurs, and one until a transition into non-participation occurs. We observe only one transition, namely, the one that occurs first. Other applications include the duration of treatments, where the exit destinations are relapse and recovery, and the duration of marriage, where one risk is divorce and the other is death of one of the spouses. More generally, the duration until an event of interest may be right-censored due to the occurrence of another event, or due to the data sampling design. The duration until the censoring is then one of the variables T_i.

Sometimes one is interested only in the distribution of Y. For example, an unemployment insurance (UI) agency may be concerned only about the expenses on UI and not in the exit destinations of recipients. In such cases one may employ standard statistical duration analysis for empirical inference with register data on the duration of UI receipt. However, in studies on individual behaviour one is typically interested in one or more of the marginal distributions of the T_i. If these variables are known to be independent, then again one may employ standard duration analysis for each of the T_i separately, treating the other variables $T_j (j \neq i)$ as independent right-censoring variables. But often it is not clear whether the T_i are independent. Indeed, economic theory often predicts that they are dependent, in particular if they can be affected by the individual's behaviour and individuals are heterogeneous. It may even be sensible from the individual's point of view to use their privately observed exogenous exit rates into destinations j as inputs for the optimal strategy affecting the exit rate into destination $i (i \neq j)$ (see, for example, van den Berg, 1990). Erroneously assuming independence leads to incorrect inference, and in fact the issue of whether the durations T_i are related is often an important question in its own right.

Unfortunately, the joint distribution of all T_i is not identified from the joint distribution of Y, Z, a result that goes back to Cox (1959). In particular, given any specific joint distribution, there is a joint distribution with independent durations T_i that generates the same distribution of the observable variables Y, Z. In other words, without additional structure, each dependent competing risks model is observationally equivalent to an independent competing risks model. The marginal distributions in the latter can be very different from the true distributions.

Of course, some properties of the joint distribution are identified. To describe these it is useful to introduce the concept of the hazard rate of a continuous duration variable, say W. Formally, the hazard rate at time t is $\theta(t) := \lim_{dt \downarrow 0} \Pr(W \in [t, t+dt))/dt$. Informally, this is the rate at which the duration W is completed at t given that it has not been completed before t. The hazard rate is the basic building block of duration analysis in social sciences because it can be directly related to individual behaviour at t. The data on Y, Z allow for identification of the hazard rates of T_i at t given that $T \geq t$. These are called the 'crude' hazard rates. If the T_i are independent, then these equal the 'net' hazard rates of the marginal distributions of the T_i.

We now turn to a number of approaches that overcome the general non-identification result for competing risks models. In econometrics, one is typically interested in covariate or regressor effects. The main approach has therefore been to specify semi-parametric models that include observed regressors X and unobserved heterogeneity terms V. With a single risk, the most popular duration model is the mixed proportional hazard (MPH) model, which specifies that $\theta(t|X=x,V) = \psi(t)\exp(x'\beta)V$ for some function $\psi(.)$. V is unobserved, and the composition of the survivors changes selectively as time proceeds, so identification from the observable distributions of $T|X$ is non-trivial. However, it holds under the assumptions that $X \perp\!\!\!\perp V$ and $\text{var}(X) > 0$ and some regularity assumptions (see van den Berg, 2001, for an overview of results). With competing risks, the analogue of the MPH model is the multivariate MPH (MMPH) model. With two risks,

$$\theta_1(t|x,V) = \psi_1(t)\exp(x'\beta_1)V_1 \quad \text{and}$$
$$\theta_2(t|x,V) = \psi_2(t)\exp(x'\beta_2)V_2.$$

where $T_1, T_2|X, V$ are assumed independent, so that a dependence of the durations given X is modelled by way of their unobserved determinants V_1 and V_2 being dependent. Many empirical studies have estimated parametric versions of this model, using maximum likelihood estimation.

The semi-parametric model has been shown to be identified, under only slightly stronger conditions than those for the MPH model (Abbring and van den Berg, 2003). Specifically, $\text{Var}(X) > 0$ is strengthened to the condition that the vector X includes two continuous variables with the properties that (*a*) their joint support contains a non-empty open set in $\mathbb{R}^2$, and (*b*) the vectors $\widetilde{\beta}_1, \widetilde{\beta}_2$ of the corresponding elements of β_1 and β_2 form a matrix $(\widetilde{\beta}_1, \widetilde{\beta}_2)$ of full rank. Somewhat loosely, X has two continuous

variables that are not perfectly collinear and that act differently on θ_1 and θ_2. Note that, with such regressors, one can manipulate $\exp(x'\beta_1)$ while keeping $\exp(x'\beta_2)$ constant. The two terms $\exp(x'\beta_i)$ are identified from the observable crude hazards at $t = 0$ because at $t = 0$ no dynamic selection due to the unobserved heterogeneity has taken place yet. Now suppose one manipulates x in the way described above. If $T_1, T_2|X$ are independent, then the observable crude hazard rate of T_2 at $t>0$, given that $T_1 \geq t$, does not vary along. But, if $T_1, T_2|X$ are dependent, then this crude hazard rate does vary along, for the following reason. First, changes in $\exp(x'\beta_1)$ affect the distribution of unobserved heterogeneity V_1 among the survivors at t, due to the well-known fact that V_1 and X are dependent conditional on survival (i.e. conditional on $T_1 \geq t>0$) even though they are independent unconditionally. Second, if V_1 and V_2 are dependent, this affects the distribution of V_2 among the survivors at t, which in turn affects the observable crude hazard of T_2 at t given that $T_1 \geq t$. In sum, the variation in this crude hazard with $\exp(x'\beta_1)$ for given $\exp(x'\beta_2)$ is informative on the dependence of the durations. An analogous argument holds for the crude hazard rate corresponding to cause $i = 1$.

Note that identification is not based on exclusion restrictions of the sort encountered in instrumental variable analysis, which require a regressor that affects one endogenous variable but not the other. Here, all explanatory variables are allowed to affect both duration variables – they are just not allowed to affect the duration distributions in the same way. Identification with regressors was first established by Heckman and Honoré (1989), who considered a somewhat larger class of models than the MMPH model and accordingly imposed stronger conditions on the support of X.

Although the MPH model is identified from single-risk duration data where we observe a single spell per subject, there is substantial evidence that estimates are sensitive to misspecification of functional forms of model elements (see van den Berg, 2001, for an overview). This implies that estimates of MMPH models using competing-risks data should also be viewed with caution. It is advisable to include additional data. For example, longitudinal survey data on unemployment durations subject to right-censoring can be augmented with register data or retrospective data not subject to censoring (see for example van den Berg, Lindeboom and Ridder, 1994). More in general, one may resort to 'multiple-spell competing risks' data, meaning data with multiple observations of Y, Z for each subject. For a given subject, such observations can be viewed as multiple independent draws from the subject-specific distribution of Y, Z, on the assumption that the unobserved heterogeneity terms V_1, V_2 are identical across the spells of the subject. Here, a subject can denote a single physical unit, like an individual, for which we observe two spells in exactly the same state, or it can denote a set of physical units for which we observe one spell each. Multiple-spell data allow for identification under less stringent conditions than single-spell data. Abbring and van den Berg (2003) showed that such data identify models that allow for full interactions between the elapsed durations t and x in $\theta_i(t|x, V)$, and, indeed, allow the corresponding effects to differ between the first and the second spell. The assumptions on the support of X are similar to above. Fermanian (2003)