# The Econometrics of Data Combination

**Geert Ridder**

Department of Economics,
University of Southern California, Los Angeles
E-mail: ridder@usc.edu

**Robert Moffitt**

Department of Economics
Johns Hopkins University,Baltimore
E-mail: moffitt@jhu.edu

**Chapter for the Handbook of Econometrics** *

April 1, 2005

*We thank J. Angrist, J. Currie, J. J. Heckman, C. F. Manski, and a referee for helpful comments on an earlier draft.

# Contents


## 1. Introduction

## 2. Merging samples with common units

### 2.1. Broken random samples

### 2.2. Probabilistic record linkage

#### 2.2.1. Matching with imperfect identifiers

#### 2.2.2. Matching errors and estimation

## 3. Independent samples with common variables

### 3.1. Fréchet bounds and conditional Fréchet bounds on the joint distribution

### 3.2. Statistical matching of independent samples

## 4. Estimation from independent samples with common variables

### 4.1. Types of inference

### 4.2. Semi- and non-parametric inference

#### 4.2.1. Conditional independence

#### 4.2.2. Exclusion restrictions

### 4.3. Parametric inference

4.3.1. Conditional independence

4.3.2. Exclusion restrictions

4.4. The origin of two-sample estimation and applications

4.5. Combining samples to correct for measurement error

## 5. Repeated cross-sections

5.1. General principles

5.2. Consistency and related issues

5.3. Binary choice models

5.4. Applications

# 1 Introduction

Economists who use survey or administrative data for inferences regarding a population may want to combine information obtained from two or more samples drawn from the population. This is the case if there is no single sample that contains all relevant variables. A special case occurs if longitudinal or panel data are needed, while only repeated cross-sections are available.

There are good reasons why data sets often do not have all relevant variables. If the data are collected by interview, it is advisable to avoid long questionnaires. If the data come from an administrative file, usually only variables that are relevant for the eligibility for a program and for the determination of the benefits or payments associated with that program are included. Hence, unless a survey was designed to include all the relevant variables for a particular research project, there is no single data set that contains all variables of interest. However, often the variables are available in two or more separate surveys. In that case it is natural to try to combine the information in the two surveys to answer the research question.

In this chapter we survey sample combination. What can be learned by combining two or more samples depends on the nature of the samples and the assumptions that one is prepared to make. If two (or more) samples from the same population are combined, there are variables that are unique to one of the samples and variables that are observed in each sample. To be specific, consider a population and assume that for each member of the population we can define the variables $Y, Z, X$. Sample A contains the variables $Y, Z$ and sample B the

variables $X, Z$. The variables in $Y$ are unique to sample A and those in $X$ are unique to sample B. Hence, we have random samples from overlapping (in variables) marginal distributions.

How one uses this information depends on the goal of the study. We distinguish between

(i) Identification and estimation of the joint distribution of $X, Y, Z$. This was the original motivation for the type of sample merging that is discussed in section 3.2. The hope was that with the merged sample the distributional impact of taxes and social programs could be studied. An example is a study of the effect of a change in the tax code on the distribution of tax payments. In principle, tax returns contain all the relevant variables. However, if the change depends on variables that did not enter the tax code before, or if it is desired to estimate the effect for specific subgroups that are not identifiable from the tax returns, the need arises to obtain the missing information from other sources. The joint distribution is also the object of interest in nonparametric (conditional) inference. This is obviously the most ambitious goal.

(ii) Estimation of economic models that involve $X, Y, Z$ (or a subset of these variables). Such models are indexed by a vector of parameters $\theta$ that is of primary interest, and, as will become clear in section 4.3, parametric restrictions are helpful (but not necessary) in securing identification by sample combination. An example is the estimation of the effect of age at school entry on the years of schooling by combining data from the US

censuses in 1960 and 1980 (Angrist and Krueger, 1992).

(iii) Estimation of an economic model with mismeasured variables. In this case sample A contains $Y, X, Z$ and sample B $X^*, Z$ with $X^*$ the correct value and $X$ the mismeasured value of the same variable, e.g. income. If $X$ is self-reported income, this variable may be an imperfect indicator of true income $X^*$. A better indicator is available in administrative data, e.g. tax records. Hence, it is desirable to combine these samples to obtain a dataset that has both the correctly measured variable and $Y$. Again this was a motivation for the type of sample merger discussed in section 3.2. In section 4.5 we show that sample merger is not necessary to avoid measurement error bias.

For problems of type (i) there are a number of methods that merge the samples A and B into one sample that is treated as a random sample from the joint distribution of $X, Y, Z$. Because the common variables $Z$[1] are often not of independent interest, we assume for the moment that the researcher is satisfied with a random sample from the joint distribution of $X, Y$. Sample merging is discussed in sections 2 and 3. Its success depends on two factors: (i) the number of members of the population that are in both samples, and (ii) the degree to which these common members can be identified from the common variables $Z$. In the simplest case $Z$ identifies members of the population uniquely, for instance if $Z$ is an individual's Social Security Number or some other unique

[1]Sometimes variables have to be transformed to make them equal in both samples. For instance, $A$ may contain the age and $B$ the year of birth.

identifier (measured without error). If the common members are a random sample from the population, then the merged sample is indeed a random sample from the population distribution of $X, Y$. Complications arise if the number of population members that are in both samples is substantial, but they cannot be identified without error. We discuss estimation in samples that have been merged. Because the matching process is not perfect the merging introduces a form of measurement or matching error. The analogy is almost complete because the bias is similar to the attenuation bias in models with mismeasured independent variables

The merger of samples has also been attempted in the case that the fraction of units that are in both samples is negligible. Indeed the techniques that have been used to merge such samples are the same as for samples with common units that cannot be identified with absolute certainty. Only under the strong assumption of conditional independence of $Y$ and $X$ given $Z$, we can treat the merged or matched sample as a random sample from the joint distribution of $Y, Z, X$ (section 4). As shown in section 4 it is preferable not to merge the two samples, even if the assumption of conditional independence is correct. Under conditional independence we can estimate the joint distribution of $Y, Z, X$ and any identified conditional model without merging the samples. If the assumption of conditional independence does not hold and our goal is to recover the joint distribution of $Y, Z_0, X$ with $Z_0$ a subvector of $Z$, then the two samples give bounds on this joint distribution. Point identification is possible if we specify a parametric model for the conditional distribution of $Y$ given $X, Z_0$, $f(y \mid x, z_0; \theta)$

or moments of that distribution, e.g. the conditional mean. In both cases, it is essential that some of the common variables in $Z$ are not in $Z_0$, i.e. that there are exclusion restrictions. In section 4.5 we also consider the case that one or more of the variables of a survey is subject to measurement error, while there is a second survey that has error free data on these variables, but does not contain data on the other relevant variables in the first survey. We show that the merger of the two samples is again not the solution, but that such data are helpful in reducing or even eliminating the errors-in-variables bias.

A special case of sample combination with some distinct variables are synthetic cohorts obtained from repeated cross-sections. In that case $Y$ and $X$ are the same variables in two time periods and $Z$ is the variable that identifies the cohort. This special case deserves separate consideration and is discussed in section 5.

This chapter provides a common framework for research in different fields of economics and statistics. It is mostly a survey, but we also point at some areas, for instance nonparametric identification of joint distributions by exclusion restrictions, that have not been explored yet. Although we survey empirical applications we have not attempted to include all studies that use some form of data combination. By bringing together research that until now was rather disjoint we hope to stimulate further research on data combination.

# 2 Merging samples with common units

An obvious way to combine information in two samples is to merge the samples. If the two samples have a substantial number of common units, the natural action is to link the records relating to the same unit. The linkage of records for the same unit is usually called *exact matching*. This term is misleading, because it suggests that the linkage is without errors. Record linkage is easy if both records contain a unique identifier, e.g. an individual's social security number, that is observed without error. Card, Hildreth, and Shore-Sheppard (2001) match survey to administrative data, and find that even in the administrative data the social security numbers are often misreported. If the two surveys are independently drawn samples from two overlapping populations, the linked records are a sample from the intersection of the two populations.

## 2.1 Broken random samples

DeGroot, Feder, and Goel (1971), DeGroot and Goel (1976) and DeGroot and Goel (1980)) consider the reconstruction of a broken random sample, i.e. a random sample in which the identity of the members is observed with error. Besides its intrinsic interest, we discuss their method because of its similarity to methods used to merge samples that have no common units.

Consider a random sample of size N from a population and assume that the identity of the units in the random sample is observed with error, i.e. a record consist of $(Y_i, Z_{1i}, Z_{2j}, X_j)$ with

$$Z_{ki} = Z_i + \varepsilon_{ki}, \qquad k = 1, 2 \tag{1}$$

The identifier $Z$ is observed with error and unit $i$ is erroneously linked to unit $j$. We ignore for the moment $Y, X$[2]. We also assume that $Z, \varepsilon_1, \varepsilon_2$ are jointly normally distributed[3], and as a consequence the observed $Z_1, Z_2$ have a bivariate normal distribution with means $\mu_1, \mu_2$, standard deviations $\sigma_1, \sigma_2$, and correlation coefficient $\rho$. Let $\phi$ denote a permutation of $1, \ldots, N$ so that $Z_{1i}$ is linked with $Z_{2\phi(i)}$. The loglikelihood of the sample $Z_{1i}, Z_{2\phi(i)}, i = 1, \ldots, N$ is

$$\begin{aligned} \ln L(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho, \phi) = C - \frac{N}{2}\log(1-\rho^2) - \frac{N}{2}\log\sigma_1^2 - \frac{N}{2}\log\sigma_2^2 - \\ -\frac{1}{2(1-\rho)^2}\sum_{i=1}^{N}\left\{\frac{(z_{1i}-\mu_1)^2}{\sigma_1^2} + \frac{(z_{2\phi(i)}-\mu_2)^2}{\sigma_2^2} - 2\rho\frac{(z_{1i}-\mu_1)(z_{2\phi(i)}-\mu_2)}{\sigma_1\sigma_2}\right\} \end{aligned} \tag{2}$$

Note that the vector $\phi$ is treated as a vector of parameters, i.e. the likelihood is the joint distribution if $\phi$ is the correct linkage. Maximizing the loglikelihood with respect to the means and variances yields the usual MLE for these parameters. If we substitute these MLE and maximize with respect to $\rho$ we obtain the concentrated loglikelihood that only depends on $\phi$

$$L(\phi) = C - \frac{N}{2}\log(1-\rho_\phi^2) \tag{3}$$

with $\rho_\phi$ the sample correlation coefficient between $Z_{1i}, Z_{2\phi(i)}, i = 1, \ldots, N$. This sample correlation coefficient depends on the permutation $\phi$. It is easily verified

[2]If $Y, X$ are correlated (given $Z_1, Z_2$) they could be helpful in reconstructing the correctly linked sample.

[3]This assumption can be relaxed, see DeGroot, Feder, and Goel (1976)

for $N = 2$ and it can be shown for all $N$ (Hájek and Šidak, 1967) that the average of the sample correlation coefficient over all permutations is equal to 0. Hence the smallest value for $\rho_\phi$ is $\rho_{min} < 0$ and the largest $\rho_{max} > 0$. If the order statistics of $Z_1, Z_2$ are denoted by $Z_{1(i)}, Z_{2(i)}$, then it is intuitively clear that the sample correlation coefficient is maximal if $Z_{1(i)}$ is linked with $Z_{2(i)}$, and minimal if $Z_{1(i)}$ is linked with $Z_{2(N-i+1)}$. The first permutation is denoted by $\phi_{max}$, the second by $\phi_{min}$. Because the concentrated loglikelihood increases with $\rho^2_\phi$, the MLE of $\rho$ is $\rho_{max}$ if $\rho^2_{max} > \rho^2_{min}$ and $\rho_{min}$ if the reverse inequality holds. In the first case the likelihood is maximized if we link according to the order statistics, and in the second case if we link in the reverse order. As is obvious from the loglikelihood in (2) the nature of the linkage, i.e. the choice of $\phi$, depends only on the sign of $\rho$. The MLE for $\rho$ suggests the following rule to decide on this sign: if $\rho^2_{max} > \rho^2_{min}$ then we estimate the sign of $\rho$ as $+1$, while we use the opposite sign if the reverse inequality holds. DeGroot and Goel (1980) conduct some sampling experiments that show that for values of $\rho$ of .9, i.e. a relatively small measurement error in the identifier , this procedure yields the correct sign in more than 75% of the replications (for sample sizes ranging from 5 to 500).

Obviously, if the $Z_1, Z_2$ are observations on a common identifier, we do not have to estimate the sign of $\rho$, because the correlation is positive, unless we make extreme assumptions on the correlation between the two measurement errors. The optimal linkage is then on the order statistic of $Z_1$ and $Z_2$. Maximization of the loglikelihood (2) with respect to the permutation $\phi$ is equivalent

to maximization of

$$\sum_{i=1}^{N} z_{1i} z_{2\phi(i)} \tag{4}$$

and this is in turn equivalent to minimization of

$$\sum_{i=1}^{N} z_{1i}^2 + \sum_{i=1}^{N} z_{2i}^2 - 2\sum_{i=1}^{N} z_{1i} z_{2\phi(i)} = \sum_{i=1}^{N} (z_{1i} - z_{2\phi(i)})^2 \tag{5}$$

Hence the Euclidean or $L^2$ distance between the vectors of observed identifiers is minimized. As we shall see, this rule that is derived for the case of exact matching with mismeasured identifiers, is also used in the case that there are no common units in the samples.

If there are multiple identifiers, i.e. if $Z$ is a $K$ vector and $Z_1, Z_2$ have a multivariate normal distributions with means $\mu_1, \mu_2$, variance matrices $\Sigma_{11}, \Sigma_{22}$, and covariance matrix $\Sigma_{12}$, the factor of the likelihood function that depends on the permutation $\phi$ is

$$\ln L(\mu, \Sigma_{12}) = \exp\left\{ -\frac{1}{2} \sum_{i=1}^{N} z_{1i}' \Sigma^{12} z_{2\phi(i)} \right\} \tag{6}$$

In this expression

$$\Sigma^{12} = -\Sigma_{11}^{-1} \Sigma_{12} (\Sigma_{22} - \Sigma_{21} \Sigma_{11}^{-1} \Sigma_{12})^{-1} \tag{7}$$

This likelihood factor is the probability that the permutation $\phi$ is the correct match and hence maximization of the likelihood function is equivalent to maximization of the probability of a correct match.

The maximization of the likelihood factor in (6) is equivalent to the maximization of

$$\sum_{i=1}^{N} z_{1i} C_{12} z_{2\phi(i)} \tag{8}$$

with $C_{12} = -\Sigma^{12}$. This is equivalent to the minimization of

$$\sum_{i=1}^{N} (z_{1i} - z_{2\phi(i)})' C_{12} (z_{1i} - z_{2\phi(i)}) \tag{9}$$

i.e. the quadratic distance with matrix $C_{12}$ between the vectors of identifiers. The same distance measure is sometimes used if the samples have no common units and $Z$ is a vector of common characteristics (see section 3.2).

Because all units must be matched the maximization of (8) is equivalent to the minimization of

$$\sum_{i=1}^{N} \sum_{j=1}^{N} d_{ij} z_{1i} C_{12} z_{2j} \tag{10}$$

subject to for $i = 1, \ldots, N$, $j = 1, \ldots, N$

$$\sum_{i=1}^{N} d_{ij} = \sum_{j=1}^{N} d_{ij} = 1 \tag{11}$$

and $d_{ij} = 0, 1$. This is a linear assignment problem, an integer programming problem for which efficient algorithms are available.

This procedure requires an estimate of $\Sigma_{12}$, the covariance matrix of $Z_1$ and $Z_2$. Note that in the case of a single identifier only the sign of this covariance was needed. If the errors in the identifiers are independent in the two samples,

an estimate of the variance matrix of the true identifier vector $Z$ suffices. The extension of DeGroot and Goel's MLE to the multivariate case has not been studied.

## 2.2 Probabilistic record linkage

### 2.2.1 Matching with imperfect identifiers

The ML solution to the reconstruction of complete records assumes that the mismeasured identifiers are ordered variables. The method of probabilistic record linkage can be used if the matching is based on (mismeasured) nominal identifiers, such as names, addresses or social security numbers. Probabilistic record linkage has many applications. It is used by statistical agencies to study the coverage of a census, by firms that have a client list that is updated regularly, and by epidemiologists who study the effect of a potentially harmful exposure (see Newcombe, 1988) for a comprehensive survey of the applications). In epidemiological studies a sample of individuals who have been exposed to an intervention is linked with a population register to determine the effects on fertility and/or mortality, the latter possibly distinguished by cause (Newcombe, Kennedy, Axford and James, 1959 ; Buehler et al., 2000; Fair et al., 2000). Probabilistic record linkage is also used in queries from a large file, e.g. finding matching fingerprints or DNA samples. The implementation of probabilistic record linkage depends on the specific features of the data. In this survey we only describe some general ideas. We use the setup of Fellegi and Sunter (1969), although we change it to stress the similarity with the reconstruction of broken random

samples (section 2.1) and statistical matching (section 3.2).

Initially we assume that there is a single identifier $Z$ that identifies each member of the population uniquely. We have two samples of sizes $N_1$ and $N_2$ from the population. These samples need not be of equal size and, although it is assumed that a substantial fraction of the units in both samples are common, the remaining units are unique to one of the samples. This is a second departure from the assumptions made in the case of a broken random sample. A key ingredient of probabilistic matching is the record generating model that describes how the observed identifiers in the records are related to the unique true identifier. It is obvious that errors in names and reported social security numbers cannot be described by a simple model with additive measurement error (Fellegi and Sunter , 1969; Copas and Hilton, 1990; and Newcombe, Fair and Lalonde, 1992) develop alternative record generating models). To keep the exposition simple, we will stick with the additive model of equation (1). The main ideas can be explained with this model and are independent of a specific model of the record generating process.

The first step is to define a comparison vector $W_{ij}$ for each pair $i, j$, with $i$ with identifier $Z_{1i}$ in the first and $j$ with identifier $Z_{2j}$ in the second random sample. An obvious choice is $W_{ij} = Z_{2j} - Z_{1i}$, but we can also include $Z_1$ and use the comparison vector $W_{ij} = (Z_{2j} - Z_{1i}, Z_{1i})'$. Define $M_{ij}$ as the indicator of the event that $i$ and $j$ are matched, i.e. are the same unit. If we assume that the measurement errors in the two samples are independent of each other and of the true identifier $Z$, and that the identifiers of distinct units are independently

distributed in the two samples, we have, for $W_{ij} = Z_{2j} - Z_{1i}$, with $f$ the density of $\varepsilon_2 - \varepsilon_1$ and $G_k$ the cdf of $Z$ in sample $k$,

$$\begin{aligned} h(w_{ij} \mid M_{ij} = 1) &= f(w_{ij}) \\ h(w_{ij} \mid M_{ij} = 0) &= \int\int f(w_{ij} - z' + z) d\, G_1(z) d\, G_2(z') \end{aligned} \tag{12}$$

For every pair $i, j$ we consider the density ratio, provided that the denominator is greater than 0 (if the denominator is 0, the match can be made without error),

$$\frac{h(w_{ij} \mid M_{ij} = 1)}{h(w_{ij} \mid M_{ij} = 0)} \tag{13}$$

This ratio gives the relative likelihood that the comparison vector is from a matched pair. Just as in a statistical test of the null hypothesis that $i, j$ refer to the same unit, we decide that the pair is matched if the density ratio exceeds a threshold. Note that with this matching rule unit $i$ may be matched with more than one unit in sample 2 and unit $j$ may be matched with more than one unit in sample 1.

To illustrate the procedure we consider a simple case. The distribution of the identifier is usually discrete. Here we assume that there is a superpopulation of identifiers from which the identifiers in the (finite) population are drawn. In particular, we assume that the $Z$'s in both samples are independent draws from a normal distribution with mean $\mu$ and variance $\sigma^2$. A uniform distribution

may be a more appropriate choice in many instances. The measurement errors are also assumed to be normally distributed with mean 0 and variances $\sigma_1^2, \sigma_2^2$.

Under these assumptions, the density ratio is

$$\frac{\phi(z_{2j} - z_{1i}; \sigma_1^2 + \sigma_2^2)}{\phi(z_{2j} - z_{1i}; 2\sigma^2 + \sigma_1^2 + \sigma_2^2} = \tag{14}$$

$$= \sqrt{\frac{2\sigma^2 + \sigma_1^2 + \sigma_2^2}{\sigma_1^2 + \sigma_2^2}} \exp\left\{ -\frac{\sigma^2}{(2\sigma^2 + \sigma_1^2 + \sigma_2^2)(\sigma_1^2 + \sigma_2^2)} (z_{2j} - z_{1i})^2 \right\}$$

The cutoff value for the density ratio can also be expressed as

$$(z_{2j} - z_{1i})^2 < C \tag{15}$$

and we match if this inequality holds. $C$ is a constant that is chosen to control either the probability of a false or a missed match. If we take the first option we choose $C$ such that

$$2\Phi\left( \frac{\sqrt{C}}{\sqrt{2\sigma^2 + \sigma_1^2 + \sigma_2^2}} \right) - 1 = \alpha \tag{16}$$

The advantage of this choice is that the cutoff value can be computed with the (estimated) variances of the observed identifiers $Z_{1i}$ and $Z_{2j}$ which are $\sigma^2 + \sigma_1^2$ and $\sigma^2 + \sigma_2^2$ respectively. Estimation of the variances of the measurement errors is not necessary. If there are multiple identifiers, the criterion for matching $i$ and $j$ is

$$(z_{2j} - z_{1i})' \left( (\Sigma_1 + \Sigma_2)^{-1} - (2\Sigma + \Sigma_1 + \Sigma_2)^{-1} \right) (z_{2j} - z_{1i}) < C \tag{17}$$

i.e. the quadratic distance with the specified matrix between the observed identifiers is less than a threshold. To use this criterion we need estimates of $\Sigma$ and $\Sigma_1 + \Sigma_2$. If $\Sigma \gg \Sigma_1 + \Sigma_2$ the criterion can be approximated by a quadratic form with matrix $(\Sigma_1 + \Sigma_2)^{-1}$, and the distance is chi-squared distributed for matches. In that case it is more convenient to choose $C$ to control the probability of a missed match.

In general, the estimation of the parameters that enter the density ratio is the most problematic part of probabilistic linkage. Tepping (1968), Copas and Hilton (1990) and Belin and Rubin (1995) propose estimation methods that use a training sample in which it is known which pairs are matched to estimate the parameters of the distribution of the comparison vector among matched and unmatched pairs.

It is interesting to compare probabilistic record linkage to the method that was proposed for the reconstruction of a broken random sample. Instead of minimizing the (average) distance between the identifiers as in (5), we choose a cutoff value for the distance and match those pairs with a distance less than the cutoff value. In probabilistic record linkage a record may be linked with two or more other records. If the true identifiers are sufficiently distinct and/or if the measurement errors are relatively small the probability of this event is negligible. Alternatively, we can choose the record that has the largest match probability.

### 2.2.2 Matching errors and estimation

The term exact matching is a misnomer when dealing with samples that have been matched using identifiers that are subject to error. Matching error biases estimates of parameters. In this section we consider the case that a random sample from a population is matched (with error) to a register that contains each unit in the sample. There has been very little work on biases due to matching errors. Usually, matched samples are analyzed as if there are no mismatches. This section provides a framework that can be used to assess potential biases and to obtain unbiased estimates if some knowledge of the matching process is available.

We assume that a random sample of size $N_1$ is matched with a register of size $N_2$ that is a random sample from the target population or the complete target population ($N_2 > N_1$). For example, we have a sample of taxpayers that is matched with the register of tax returns. The sample contains a variable $X$ and an identifier $Z_1$ that is measured with error and the register contains a variable $Y$ and an identifier $Z_2$ that is also measured with error. The true identifier is denoted by $Z$. We want to study the relation between $X$ and $Y$ or in general statistics defined for the joint distribution of $X, Y$. In fact, we show that the joint distribution of $X, Y$ is (nonparametrically) identified, if the matching probabilities are available.

The data are generated as follows. First, a sample of size $N_2$ is drawn from the joint distribution of $X, Y, Z$. This sample is the register. Next, we generate the mismeasured identifiers $Z_1, Z_2$, e.g. according to (1) or some other record

generating model discussed in the previous section. We observe $Y_j, Z_{2j}, j = 1, \ldots, N_2$. The next step is to draw $N_1 < N_2$ observations from the register without replacement. This is the sample, for which we observe $X_i, Z_{1i}, i = 1, \ldots, N_1$. Note that in this case all members in the sample are represented in the register.

The bias induced by the matching errors depends on the relation between the mismeasured identifier and the variables of interest. For instance, if the identifier is a (misreported) social security number, then it is reasonable to assume that both the identifier $Z$ and the observed values $Z_1, Z_2$ are independent of the variables of interest. If, in addition, there is a subsample with correctly reported identifiers $Z_1 = Z_2 = Z$, e.g. the subsample with $Z_1 = Z_2$ (this is an assumption), then this subsample is a random sample from the joint distribution of the variables of interest. However, often common variables beside the identifier are used to match units $i$ and $j$ with $z_{1i} \neq z_{2j}$, e.g. we match $i$ and $j$ if $z_{1i}$ and $z_{2j}$ are close and $i$ and $j$ have the same gender, age, and location etc. Note that the additional common variables need not be observed with error in the two samples. However, the probability that the match is correct depends on these additional common variables that in general are correlated with variables of interest. In this case, even if we can identify a subsample in which all matches are correct, this subsample is not a random sample from the joint distribution of the variables of interest.

Here we only consider the case that $Z, Z_1, Z_2$ are independent of $X, Y$. The general case can be analyzed in much the same way. Note that this the simplest

case for probabilistic record linkage. There is an interesting contrast with statistical matching, as discussed in the next section, because there the quality of the approximation relies heavily on the correlation between the identifiers and the variables of interest.

The quality of the matches depends on the matching method that in turn depends on the record generating model. We use the same example that was considered in section 2.2.1. The record generating model is as in (1) and $Z$, $\varepsilon_1$ and $\varepsilon_2$ are all independently normally distributed. Under these assumptions $i$ in the sample is matched with $\phi(i)$ in the register if and only if $|z_{2\phi(i)} - z_{1i}| < C$ with $C$ determined e.g. as in (16) or by some other rule. We can derive an expression for the probability that the match is correct given that we use this matching rule, i.e. the probability of the event that $Z_i = Z_{\phi(i)}$ given that $|Z_{2\phi(i)} - Z_{1i}| \leq C$. Substitution of (1) and using the independence of the reporting errors and the true value gives by Bayes' theorem

$$\Pr(M_{i\phi(i)} = 1) = \Pr(Z_i = Z_{\phi(i)} \mid |Z_{2\phi(i)} - Z_{1i}| \leq C) = \tag{18}$$

$$= \frac{\Pr(Z_i = Z_{\phi(i)}) \Pr(|\varepsilon_{2\phi(i)} - \varepsilon_{1i}| < C)}{\Pr(Z_i = Z_{\phi(i)}) \Pr(|\varepsilon_{2\phi(i)} - \varepsilon_{1i}| < C) + \Pr(Z_i \neq Z_{\phi(i)}) \Pr(|Z_{\phi(i)} + \varepsilon_{2\phi(i)} - Z_i - \varepsilon_{1i}| < C)} =$$

$$= \frac{\frac{1}{N_2} \Phi\left(\frac{C}{\sqrt{\sigma_1^2+\sigma_2^2}}\right)}{\frac{1}{N_2} \Phi\left(\frac{C}{\sqrt{\sigma_1^2+\sigma_2^2}}\right) + \frac{N_2-1}{N_2} \Phi\left(\frac{C}{\sqrt{\sigma_1^2+\sigma_2^2+2\sigma^2}}\right)}$$

This expression for the probability of a correct match under the given matching rule has a Bayesian flavor. The probability of a correct match, if a unit in the sample is matched at random with a unit in the register is $\frac{1}{N_2}$. This is also the limit of the probability of a correct match if $C \to \infty$. The probability

decreases in $C$. If $C \downarrow 0$ we obtain the limit

$$\frac{\frac{1}{N_2}}{\frac{1}{N_2} + \frac{N_2-1}{N_2}\sqrt{\frac{\sigma_1^2+\sigma_2^2}{\sigma_1^2+\sigma_2^2+2\sigma^2}}} \tag{19}$$

and this probability approaches 1 if the reporting error in the identifier is small. Hence, we improve on random matching by using the noisy identifiers. Of course, if we choose $C$ too small, there will be few matches. As will be seen below, the variance of estimators is inversely proportional to the probability of a correct match, so that if our goal is to estimate parameters accurately we face a trade-off between the number of matched observations and the probability that the match is correct. Although this analysis is for a specific record generating model, the trade-off is present in all matched samples.

If we match $i$ in the sample to $\phi(i)$ in the register, if $|Z_{2\phi(i)} - Z_{1i}| \leq C$, then the conditional probability of a correct match given the identifiers $Z_1, Z_2$ is

$$\Pr(M_{i\phi(i)} = 1 \mid Z_{1i}, Z_{2\phi(i)}) = \Pr(Z_i = Z_{\phi(i)} || Z_{2\phi(i)} - Z_{1i}| \leq C, Z_{1i}, Z_{2\phi(i)}) = \tag{20}$$

$$= \frac{\Pr(M_{i\phi(i)} = 1)\phi_1(Z_{2\phi(i)} - Z_{1i})}{\Pr(M_{i\phi(i)} = 1)\phi_1(Z_{2\phi(i)} - Z_{1i}) + \Pr(M_{i\phi(i)} = 0)\phi_2(Z_{2\phi(i)} - Z_{1i})}$$

with

$$\phi_1(Z_{2\phi(i)} - Z_{1i}) = \phi(Z_{2\phi(i)} - Z_{1i} \mid |Z_{2\phi(i)} - Z_{1i}| \leq C; \sigma_1^2 + \sigma_2^2)$$

$$\phi_2(Z_{2\phi(i)} - Z_{1i}) = \phi(Z_{2\phi(i)} - Z_{1i} \mid |Z_{2\phi(i)} - Z_{1i}| \leq C; 2\sigma^2 + \sigma_1^2 + \sigma_2^2)$$

Now we are in a position to discuss estimation. Consider a pair $i, \phi(i)$ matched according to a matching rule, e.g. the rule above, from the $N_1 \times N_2$ possible

pairs. The joint distribution of $X_i, Z_{1i}, Y_{\phi(i)}, Z_{2\phi(i)}$ has density $g(x_i, z_{1i}, y_{\phi(i)}, z_{2\phi(i)})$ with

$$g(x_i, z_{1i}, y_{\phi(i)}, z_{2\phi(i)}) = g(x_i, z_{1i}, y_{\phi(i)}, z_{2\phi(i)}, M_{i\phi(i)} = 1) + \tag{21}$$

$$+ g(x_i, z_{1i}, y_{\phi(i)}, z_{2\phi(i)}, M_{i\phi(i)} = 0)$$

If the joint density of $X, Y$ is $f(x, y)$, then because we assume that $X, Y$ and $Z, Z_1, Z_2$ are independent,

$$g(x_i, z_{1i}, y_{\phi(i)}, z_{2\phi(i)}, M_{i\phi(i)} = 1) = \tag{22}$$

$$= f(x_i, y_{\phi(i)}) \Pr(M_{i\phi(i)} = 1 \mid z_{1i}, z_{2\phi(i)}) g(z_{1i}, z_{2\phi(i)})$$

and

$$g(x_i, z_{1i}, y_{\phi(i)}, z_{2\phi(i)}, M_{i\phi(i)} = 0) = \tag{23}$$

$$= f_1(x_i) f_2(y_{\phi(i)}) \Pr(M_{i\phi(i)} = 0 \mid z_{1i}, z_{2\phi(i)}) g(z_{1i}, z_{2\phi(i)})$$

Substituting (22) and (23) in (21), and using $g(x_i, z_{1i}, y_{\phi(i)}, z_{2\phi(i)} = f(x_i, y_{\phi(i)}) g(z_{1i}, z_{2\phi(i)})$, we can solve for $f(x_i, y_{\phi(i)})$

$$f(x_i, y_{\phi(i)}) = \frac{g(x_i, y_{\phi(i)}) - \Pr(M_{i\phi(i)} = 0 \mid z_{1i}, z_{2\phi(i)}) f_1(x_i) f_2(y_{\phi(i)})}{\Pr(M_{i\phi(i)} = 1 \mid z_{1i}, z_{2\phi(i)})} = \tag{24}$$

$$= f_1(x_i) f_2(y_{\phi(i)}) + \frac{g(x_i, y_{\phi(i)}) - f_1(x_i) f_2(y_{\phi(i)})}{\Pr(M_{i\phi(i)} = 1 \mid z_{1i}, z_{2\phi(i)})}$$

if the denominator is greater than 0, which is the case for any sensible matching rule.

The distributions on the right-hand side of this expression are all observed. Hence this identification result is nonparametric, although it requires that the matching probabilities are known or that they can be estimated.

Often we are not interested in the joint distribution of $Y, X$, but in a population parameter $\theta_0$ that is the unique solution to a vector of population moment conditions

$$\mathrm{E}\left[m(X_i, Y_i; \theta)\right] = 0 \tag{25}$$

These population moment conditions refer to the correctly matched observations. If two observations are incorrectly matched, they are stochastically independent. In general for $i \neq j$

$$\mathrm{E}\left[m(X_i, Y_j; \theta)\right] = 0 \tag{26}$$

is solved by $\theta_1 \neq \theta_0$. In other words, the parameter cannot be identified from the two marginal distributions.

The solution for the joint population distribution in (24) suggests the sample moment conditions that combine information from the sample and the register

$$\begin{aligned} &\frac{1}{N_1}\sum_{i=1}^{N_1}\frac{m(x_i, y_{\phi(i)}; \theta)}{\Pr(M_{i\phi(i)} = 1 \mid z_{1i}, z_{2\phi(i)})} - \\ &-\frac{1}{N_1^2}\sum_{j=1}^{N_1}\sum_{k=1}^{N_1}\frac{1 - \Pr(M_{j\phi(k)} = 1 \mid z_{1j}, z_{2\phi(k)})}{\Pr(M_{j\phi(k)} = 1 \mid z_{1j}, z_{2\phi(k)})} m(x_j, y_{\phi(k)}; \theta) \end{aligned} \tag{27}$$

and the weighted GMM estimator of $\theta$ either makes (27) equal to 0 or is the minimizer of a quadratic form in these sample moment conditions. In this

expression (but not in (24)) it is implicitly assumed that the probability that a unit in the sample is matched with two or more units in the register is negligible. This simplifies the notation.

We obtain a particularly simple result if we use the identifiers to match the sample to the register, but ignore them in the inference, i.e. in (21) we start with the joint distribution of $X_i, Y_{\phi(i)}$, so that

$$f(x_i, y_{\phi(i)}) = f_1(x_i) f_2(y_{\phi(i)}) + \frac{g(x_i, y_{\phi(i)}) - f_1(x_i) f_2(y_{\phi(i)})}{\Pr(M_{i\phi(i)} = 1)}$$

This will give consistent, but less efficient, estimates. Let the probability of a correct match $\Pr(M_{i\phi(i)} = 1) = \lambda$. If $X$ and $Y$ have mean 0, then

$$\operatorname{cov}(X_i, Y_i) = \frac{\operatorname{cov}(X_i, Y_{\phi(i)})}{\lambda} \tag{28}$$

With the same assumption we find for the moment conditions of a simple linear regression with an intercept

$$\mathrm{E}\left[(Y_i - \alpha - \beta X_i) X_i\right] = \tag{29}$$
$$= \frac{\mathrm{E}\left[(Y_{\phi(i)} - \alpha - \beta X_i) X_i\right] - (1-\lambda)\left[\mathrm{E}(Y_{\phi(i)})\mathrm{E}(X_i) - \alpha \mathrm{E}(X_i) - \beta \mathrm{E}(X_i^2)\right]}{\lambda}$$

$$\mathrm{E}\left[Y_i - \alpha - \beta X_i\right] = \tag{30}$$
$$= \frac{\mathrm{E}\left[Y_{\phi(i)} - \alpha - \beta X_i\right] - (1-\lambda)\left[\mathrm{E}(Y_{\phi(i)}) - \alpha - \beta \mathrm{E}(X_i)\right]}{\lambda} =$$
$$= \mathrm{E}\left[Y_{\phi(i)} - \alpha - \beta X_i\right]$$

Setting these conditions equal to 0 and solving for the parameters we find that

$$\begin{aligned} \beta &= \frac{\text{cov}(X_i, Y_{\phi(i)})}{\lambda \text{var}(X_i)} \\ \alpha &= \text{E}(Y_{\phi(i)}) - \beta \text{E}(X_i) \end{aligned} \tag{31}$$

and, if we substitute the sample statistics for the population statistics,we obtain the estimator suggested by Neter, Maynes and Ramanathan (1965) and Scheuren and Winkler (1993). The results in this section generalize their results to arbitrary moment conditions and less restrictive assumptions on the sampling process. In particular, we show that the matching probabilities that are computed for probabilistic linkage can be used to compute the moment conditions for the matched population. This is important because the simulation results in Scheuren and Winkler (1993) show that the bias induced by false matches can be large.

The asymptotic variance of the estimator for $\beta$ is

$$\text{var}(\hat{\beta}) = \frac{\sigma^2}{N_1 \lambda^2 \text{var}(X)} \tag{32}$$

The variance decreases with the matching probability. The GMM estimator is consistent if the matching probability is positive.

# 3 Independent samples with common variables

## 3.1 Fréchet bounds and conditional Fréchet bounds on the joint distribution

Exact or probabilistic matching is not advisable if the fraction of units that are in both samples is small. If the fraction is negligible, we may treat the two random samples as independent samples that have no units in common. Although exact or probabilistic matching produces more informative data, the fear that linked files pose a threat to the privacy of individuals who, with some effort, may be identifiable from the linked records, has prevented the large scale matching of administrative and survey data[4]. As a consequence, often the only available samples that contain all relevant variables are relatively small random samples from a large population. It is safe to assume that these random samples have no common units.

The two independent random samples identify the marginal distributions of $X, Z$ (sample A) and $Y, Z$ (sample B). If there are no common variables $Z$, the marginal distributions put some restrictions on the joint distribution of $X, Y$. These Fréchet (1951) bounds on the joint distribution are not very informative. For example, if the marginal and joint distributions are all normal, there is no restriction on the correlation coefficient of $X$ and $Y$, i.e. it can take any value

[4]Fellegi (1999) notes that public concern with file linkage varies over place and time and that, ironically, the concern is larger if the linkage is performed by government agencies than if private firms are involved. Modern data acquisition methods like barcode scanners and the internet result in large files that are suitable for linkage.

between -1 and 1.

With common variables $Z$ the Fréchet bounds can be improved. The bounds for the joint conditional cdf of $X, Y$ given $Z = z$ are

$$\max\{F(x \mid z) + F(y \mid z) - 1, 0\} \leq F(x, y \mid z) \leq \min\{F(x \mid z), F(y \mid z)\} \quad (33)$$

Taking the expectation over the distribution of the common variables $Z$ we obtain

$$\mathrm{E}\left[\max\{F(x \mid Z) + F(y \mid Z) - 1, 0\}\right] \leq F(x, y) \leq \mathrm{E}\left[\min\{F(x \mid Z), F(y \mid Z)\}\right] \quad (34)$$

The bounds are sharp, because the lower and upper bounds, $\mathrm{E}\left[\max\{F(x \mid Z) + F(y \mid Z) - 1, 0\}\right]$ and $\mathrm{E}\left[\min\{F(x \mid Z), F(y \mid Z)\}\right]$ are joint cdf's of $X, Y$ with marginal cdf's equal to $F(x)$ and $F(y)$. Note that because the expectation of the maximum is greater than the maximum of the expectations (the reverse relation holds for the expectation of the minimum), the Fréchet bounds with common variables are narrower than those without. If either $X$ or $Y$ are fully determined by $Z$, then the joint cdf is identified. To see this let the conditional distribution of $X$ given $Z = z$ be degenerate in $x(z)$. Define $A(x) = \{z \mid x(z) \leq x\}$. Then $F(x \mid z) = 1$ if $z \in A(x)$ and $F(x \mid z) = 0$ if $z \in A(x)^c$. Substitution in (34) gives that the lower and upper bound coincide and that

$$F(x, y) = \mathrm{E}\left[F(y \mid Z) \mid Z \in A(x)\right) \Pr(Z \in A(x))] \quad (35)$$

In the special case that the population distribution of $X, Y, Z$ is trivariate normal, the only parameter that can not be identified is the correlation between

$X$ and $Y$. We have

$$\rho_{XY} = \rho_{XY|Z}\sqrt{1-\rho_{XZ}^2}\sqrt{1-\rho_{YZ}^2} + \rho_{XZ}\rho_{YZ} \tag{36}$$

This gives the bounds

$$\rho_{XZ}\rho_{YZ} - \sqrt{1-\rho_{XZ}^2}\sqrt{1-\rho_{YZ}^2} \leq \rho_{XY} \leq \rho_{XZ}\rho_{YZ} + \sqrt{1-\rho_{XZ}^2}\sqrt{1-\rho_{YZ}^2} \tag{37}$$

The lower bound reaches its minimum -1 if $\rho_{XZ} = -\rho_{YZ}$ (the upper bound is $1 - 2\rho_{XZ}^2$) and the upper bound reaches its maximum 1 if $\rho_{XZ} = \rho_{YZ}$ (the lower bound is $-1 + 2\rho_{XZ}^2$. Also if either $\rho_{XZ}$ or $\rho_{YZ}$ is equal to 1, then $\rho_{XY} = \rho_{XZ}\rho_{YZ}$. The length of the interval is $2\sqrt{1-\rho_{XZ}^2}\sqrt{1-\rho_{YZ}^2}$ and hence the the bound is narrower if the correlation between either $Z$ and $X$ or $Z$ and $Y$ is high.

An example illustrates how much correlation between $X, Y$ and $Z$ is required to obtain sufficiently narrow bounds. Consider a linear regression model

$$Y = \alpha + \beta X + U \tag{38}$$

where $X$ and $U$ are independent and normally distributed. If $\sigma_X, \sigma_Y$ denote the standard deviation of $X$ and $Y$, respectively, we have

$$\frac{\sigma_Y}{\sigma_X} = \frac{|\beta|}{\sqrt{R^2}} \tag{39}$$

with $R^2$ the coefficient of determination of the regression. If we multiply the bounds in (38) by $\frac{\sigma_Y}{\sigma_X}$ we obtain an interval for the slope $\beta$. If $p$ denotes the

relative (with respect to $\beta$) length of the interval and we consider the case that the correlation between $X$ and $Z$ and $Y$ and $Z$ are equal, we obtain the following expression for the required correlation

$$\rho_{XZ} = \sqrt{1 - \frac{p\sqrt{R^2}}{2}} \tag{40}$$

The correlation decreases with $R^2$ and the (relative) length of the interval for $\beta$. For instance, if we want a .20 relative length for a regression with an $R^2$ of .9, we need that $\rho_{XZ} = \rho_{YZ} = .95$. In general, the correlation that is needed to obtain informative bounds is rather high, and this illustrates the limited information about the relation between $X$ and $Y$ in the combined sample.

The Fréchet bounds on the joint cdf in (34) treat the variables $X$ and $Y$ symmetrically. As the notation suggests, often $Y$ is the dependent and $X$ the independent variable in a relation between these variables, and we focus on the conditional distribution of $Y$ given $X$. An important reason to do this, is that we may assume that this conditional distribution is invariant under a change in the marginal distribution of $X$. For example, Cross and Manski (2002) consider the case that $Y$ is the fraction of votes for a party and $X$ is the indicator of an ethnic group. It is assumed that the ethnic groups vote in the same way in elections,but that the ethnic composition of the voters changes over time. If we have the marginal distributions of $Y$ (election results by precinct) and $X$ (ethnic composition by precinct), what can we say about future election results, if we have a prediction of the future composition of the population, i.e. the future marginal distribution of $X$?

Horowitz and Manski (1995) and Cross and Manski (2002) have derived

bounds for the case that $X$ is a discrete variable with distribution

$$\Pr(X = x_k) = p_k \;\; k = 1, \ldots, K \tag{41}$$

We first derive their bounds for the case that there are no common variables $Z$. They consider bounds on the conditional expectation

$$\mathrm{E}\left[g(h(Y), X)|X = x\right]$$

with $g$ bounded and monotone in $h$ for almost all $x$. A special case is $g(h(Y), X) = I(Y \leq y)$ which gives the conditional cdf. Because the conditional expectation above is continuous and increasing in $F(y|x)$, in the sense that the expectation with respect to $F_1(y|x)$ is not smaller than that with respect to $F_2(Y|x)$, if $F_1(y|x)$ first-order stochastically dominates $F_2(y|x)$, we can derive bounds on this expectation from bounds on the conditional cdf.

In the sequel we derive bounds both on the conditional cdf $F(y|x)$ and on $F(y; x_k) = \Pr(Y \leq y, X = x_k)$. We first derive bounds on these cdf's for a given $k$. Next we consider the $K$-vector of these cdf's. Note that by the law of total probability

$$\sum_{k=1}^{K} F(y; x_k) = F(y)$$

which imposes an additional restriction on the vector $F(y; x_k), k = 1, \ldots, K$. The Fréchet bounds on $F(y; x_k)$ are

$$\max\{F(y) - (1 - p_k), 0\} \leq F(y; x_k) \leq \min\{F(y), p_k\} \tag{42}$$

For each $k$ these bounds are sharp, because both the lower and upper bound are increasing in $y$, and they both increase from 0 to $p_k$, i.e. they are $\tilde{F}(y; x_k)$ for some random variables $\tilde{Y}$ and $\tilde{X}$.

The bounds in (42) imply that if $p_k \leq \frac{1}{2}$

$$\begin{aligned} 0 &\leq F(y;x_k) \leq F(y), \quad y < F^{-1}(p_k) \\ 0 &\leq F(y;x_k) \leq p_k, \quad F^{-1}(p_k) \leq y < F^{-1}(1-p_k) \\ F(y)-(1-p_k) &\leq F(y;x_k) \leq p_k, \quad y \geq F^{-1}(1-p_k) \end{aligned} \tag{43}$$

with an obvious change if $p_k > \frac{1}{2}$. Upon division by $p_k$ we obtain bounds on the conditional cdf of $Y$ given $X = x_k$

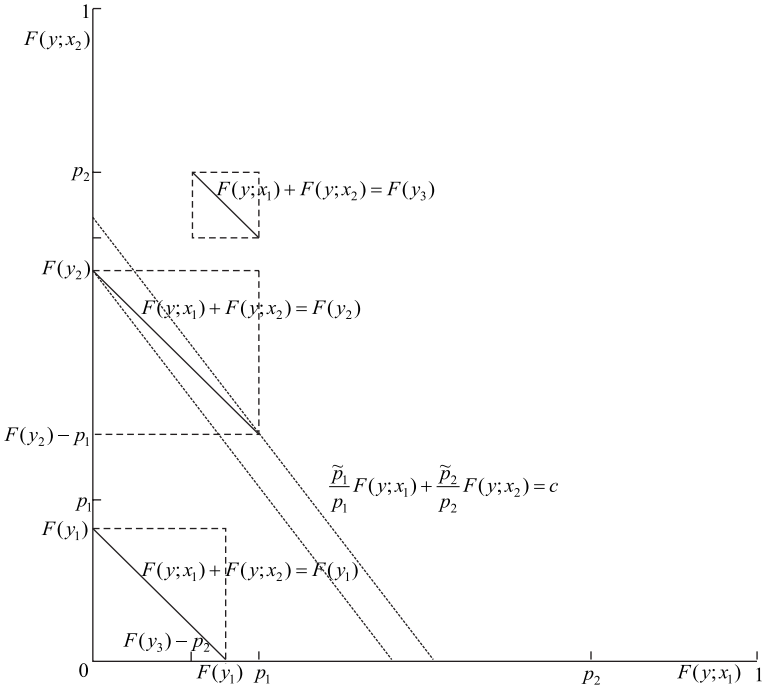
$$\begin{aligned} 0 &\leq F(y \mid x_k) \leq \frac{F(y)}{p_k}, \quad y < F^{-1}(p_k) \\ 0 &\leq F(y \mid x_k) \leq 1, \quad F^{-1}(p_k) \leq y < F^{-1}(1-p_k) \\ \frac{F(y)-(1-p_k)}{p_k} &\leq F(y \mid x_k) \leq 1, \quad y \geq F^{-1}(1-p_k) \end{aligned} \tag{44}$$

The bounds have an appealing form. The lower bound is the left truncated cdf of $Y$ where the truncation point is the $(1-p_k)$-th quantile of the distribution of $Y$ and the upper bound is the right truncated cdf with truncation point equal to the $p_k$-th quantile. These bounds on the conditional cdf of $Y$ were derived by Horowitz and Manski (1995) and Cross and Manski (2002). They are essentially Fréchet bounds on the joint distribution.

Next, we consider bounds on the vector $F(y;.) = (F(y;x_1) \ldots F(y;x_K))'$. For $K = 2$ the bounds in (42) are (without loss of generality we assume $p_1 < \frac{1}{2}$, i.e. $p_2 = 1 - p_1 > p_1$)

$$\begin{aligned} 0 &\leq F(y;x_1) \leq F(y), \quad y < F^{-1}(p_1) \\ 0 &\leq F(y;x_1) \leq p_1, \quad F^{-1}(p_1) \leq y < F^{-1}(p_2) \end{aligned}$$

Figure 1: Bounds on $(F(y;x_1), F(y;x_2))$ for three values of $y$.




$$
\begin{aligned}
F(y)-p_2 \quad &\leq F(y;x_1) \leq \quad p_1, \qquad y \geq F^{-1}(p_2) \\
0 \quad &\leq F(y;x_2) \leq \quad F(y), \quad y < F^{-1}(p_1) \\
F(y)-p_1 \quad &\leq F(y;x_2) \leq \quad F(y), \quad F^{-1}(p_1) \leq y < F^{-1}(p_2) \\
F(y)-p_1 \quad &\leq F(y;x_2) \leq \quad p_2, \qquad y \geq F^{-1}(p_2)
\end{aligned} \tag{45}
$$

By the law of total probability $F(y;.)$ satisfies for all $y$

$$
\sum_{k=1}^{K} F(y;x_k) = F(y) \tag{46}
$$

Hence, the vector of conditional cdf's is in a set that is the intersection of the Fréchet bounds in (45) and the hyperplane in (46). The resulting bounds

on $(F(y;x_1), F(y;x_2))$ are given in figure 1 for three values of $y$ with $y_1 < F^{-1}(p_1)$, $F^{-1}(p_1) \leq y_2 < F^{-1}(p_2)$, and $y_3 \geq F^{-1}(p_2)$. The Fréchet bounds on $(F(y;x_1), F(y;x_2))$ are the squares. The law of total probability selects two vertices of these squares as the extreme points of the set of $(F(y;x_1), F(y;x_2))$ that satisfy both the Fréchet bounds and the law of total probability. Bounds on the conditional cdf's $F(y \mid x_1)$ and $F(y \mid x_2)$ are obtained upon division by $p_1$ and $p_2$ respectively. This amounts to a change in the units in figure 1 and except for that the figure is unchanged.

From (45) the lower bound on $F(y;x_1)$ is

$$\begin{aligned} F_L(y;x_1) &= 0 & y < F^{-1}(p_2) \\ &= F(y) - p_2 & y \geq F^{-1}(p_2) \end{aligned}$$

and the upper bound is

$$\begin{aligned} F_U(y;x_1) &= F(y) & y < F^{-1}(p_1) \\ &= p_1 & y \geq F^{-1}(p_1) \end{aligned}$$

Note that both the lower and upper bound increase from 0 to $p_1$ with $y$, and hence are equal to $\tilde{F}(y;x_1)$ for some random variables $\tilde{Y}$ and $\tilde{X}$. The corresponding upper and lower bounds on $F(y;x_2)$ are $F_U(y;x_2) = F(y) - F_L(y;x_1)$ and $F_L(y;x_2) = F(y) - F_U(y;x_1)$, and these bounds are equal to $\tilde{F}(y;x_2)$ for some random variables $\tilde{Y}$ and $\tilde{X}$.This establishes that the bounds are sharp. A general proof of this statement can be found in Cross and Manski.

The bounds on the conditional cdf's $F(y|x_1)$ and $F(y|x_2)$ are also given in figure 2. By the law of total probability, the lower bound of $F(y|x_1)$ corresponds

with upper bound of $F(y|x_2)$ and the other way around. Note that the bounds are narrower for $F(y|x_2)$ because $x_2$ has a higher probability than $x_1$. From this figure we can obtain bounds on the conditional median of $Y$ given $X$. We find that the change in this conditional median has bounds

$$F^{-1}(\frac{1}{2}-\frac{1}{2}p_1)-F^{-1}(1-\frac{1}{2}p_1) \leq \text{med}(Y|x_2)-\text{med}(Y|x_1) \leq F^{-1}(\frac{1}{2}+\frac{1}{2}p_1)-F^{-1}(\frac{1}{2}p_1) \tag{47}$$

Note that the lower bound is negative and the upper bound positive for all $p_1$, so that it is impossible to sign the change of the conditional median with this information. This suggests that the relation between $Y$ and $X$ cannot be inferred from two marginal distributions without common variables.

If $K \geq 3$ the bounds can be derived in the same way. First, we order the $p_k$ by increasing size. Next, we find the hypercubes that correspond to the Fréchet bounds on $F(y;.)$. As in figure 1 the vertices depend on the value of $y$, i.e. for which $k$ we have $F^{-1}(p_k) \leq y < F^{-1}(p_{k+1})$. Finally, we select the vertices that satisfy the law of total probability. These are the extreme points of the set of admissible $F(y;x_k), k = 1,\ldots,K$. To be precise, the set is the convex hull of these extreme points. As we shall see below, for prediction purposes it is sufficient to find the vertices.

The main reason for bounds on the conditional cdf of $Y$ given $X$, instead of on the joint cdf of $Y, X$, is that it is usually assumed that the conditional cdf is invariant with respect to changes in the distribution of $X$. Of course, this is a common assumption in conditional econometric models with fixed parameters. An obvious application is to conditional prediction. Cross and Manski consider

the prediction of the outcome of a future election assuming that the voting behavior of demographic groups remains the same, but that the composition of the population changes and the future composition of the population can be predicted accurately.

The predicted distribution of the future outcome $\tilde{F}(y)$ satisfies

$$\tilde{F}(y) = F(y; x_1)\frac{\tilde{p}_1}{p_1} + F(y; x_2)\frac{\tilde{p}_2}{p_2} \tag{48}$$

with $\tilde{p}_1$ the future fraction with $X = x_1$. Again, without loss of generality we assume $p_1 < \frac{1}{2}$. We can further distinguish between $\tilde{p}_1 \leq p_1$ and $\tilde{p}_1 > p_1$. In the former case the bounds on the predicted cdf can be found as in figure 1. In that figure we indicate the bounds for $F^{-1}(p_1) \leq y < F^{-1}(p_2)$. The bounds are obtained by intersecting the set of feasible $(F(y; x_1), F(y; x_2))$ with (48). We find

$$\begin{aligned}
\frac{\tilde{p}_1}{p_1}F(y) \quad &\leq \tilde{F}(y) \leq \quad \min\left\{\frac{\tilde{p}_2}{p_2}F(y), 1\right\}, \quad y < F^{-1}(p_1) \\
1 - \frac{\tilde{p}_2}{p_2}(1 - F(y)) \quad &\leq \tilde{F}(y) \leq \quad \min\left\{\frac{\tilde{p}_2}{p_2}F(y), 1\right\}, \quad F^{-1}(p_1) \leq y < F^{-1}(p_2) \\
1 - \frac{\tilde{p}_2}{p_2}(1 - F(y)) \quad &\leq \tilde{F}(y) \leq \quad 1 - \frac{\tilde{p}_1}{p_1}(1 - F(y)), \quad y \geq F^{-1}(p_2)
\end{aligned} \tag{49}$$

As is obvious from figure 1, the bounds increase with the difference between $p_1$ and $\tilde{p}_1$. For $K \geq 3$ the bounds on the predicted cdf are found by evaluating

$$\sum_{k=1}^{K} \frac{\tilde{p}_k}{p_k} F(y; x_k) \tag{50}$$

at the extreme points of the set of feasible $F(y;.)$.

As noted, a key assumption in the derivation of the bounds is that $X$ is a discrete variable. From (44) it is obvious that the bounds on the conditional cdf become uninformative if $p_k$ goes to 0, i.e the bounds become $0 \leq F(y \mid x_k) \leq 1$ for all $y$. Hence, if $X$ is close to continuous the bounds on the conditional cdf's are not useful. If the support of $Y$ is bounded, e.g. if it is a dichotomous variable, then the bounds on the support can be used to obtain bounds on conditional expectations. Such bounds are of a different nature and beyond the scope of this chapter.

## 3.2 Statistical matching of independent samples

The Fréchet bounds exhaust the information on the joint distribution of $X, Y$. If we merge the samples A and B no information is added, and our knowledge of the joint distribution of $X$ and $Y$ does not increase. How much we can learn about the joint distribution of $X, Y$ is completely determined by the relation between $X$ and $Z$ in sample A and that between $Y$ and $Z$ in sample B.

In spite of this, the temptation to match two samples that do not have common units as if they were two samples with a substantial degree of overlap has been irresistible. A number of authors have proposed methods for this type of file matching (Okner, 1972; Ruggles and Ruggles, 1974 ; Radner, 1974; Ruggles, Ruggles, and Wolff, 1977; Barr and Turner, 1978; Kadane, 1978; see also the survey in Radner et al. ,1980). These methods are direct applications of those that are used in the reconstruction of broken random samples and

probabilistic matching. Let the sample A be $x_i, z_{1i}, i = 1, \ldots, N_1$ and the sample B be $y_i, z_{2i}, i = 1, \ldots, N_2$. The vectors $z_1$ and $z_2$ contain the same variables and the subscript only indicates whether the observation is in sample A or B. Because the samples A and B do not contain common units, the fact that $z_{1i}$ and $z_{2j}$ are close does not imply that they refer to the same unit or even similar (except for these variables) units. If we match unit $i$ in A to unit $j$ in B we must decide which of the vectors $z_{1i}$ or $z_{2j}$ we include in the matched file. If we use the observation for file A, then this file is referred as the base file, and file B is called the supplemental file.

The two methods that have been used in the literature are constrained and unconstrained matching. Both methods require the specification of a distance function $D(z_{1i}, z_{2j})$. In (9) (for broken random sample) and (18) (for probabilistic record linkage) we specify the distance function as a quadratic function of the difference, but other choices are possible[5]. In practice, one must also decide on which variables to include in the comparison, i.e. in the $z$ vector. The Fréchet bounds suggest that the joint distribution of $X, Y$ is best approximated, if the correlation between either $X$ or $Y$ and $Z$ or the $R^2$ in a regression of either $X$ or $Y$ on $Z$ is maximal. Often, the units that can be matched are restricted to e.g. units that have the same gender. In that case gender is called a cohort variable.

With constrained matching every unit in sample A is matched to exactly

[5]Rodgers (1984) finds no systematic differences in the performance of distance functions, although he comments that the Mahalanobis distance using an estimated variance matrix does not perform well.

one unit in sample B. Often A and B do not have an equal number of units. However, both are random samples from a population and hence the sampling fraction for both samples is known (assume for the moment that the sample is obtained by simple random sampling) . The inverse of the sampling fraction is the sample weight, $w_A$ for sample A and $w_B$ for sample B. Assume that the weights are integers. Then we can replicate the units in sample A $w_A$ times and those in sample B $w_B$ times to obtain two new samples that have the same number of units $M$ (equal to the population size). Now we match the units in these samples as if they were a broken random sample, i.e. we minimize over $d_{ij}, i = 1, \ldots, M, \quad j = 1, \ldots, M$ with $d_{ij} = 1$ if $i$ and $j$ are matched

$$\sum_{i=1}^{M} \sum_{j=1}^{M} d_{ij} D(z_{1i}, z_{2j}) \tag{51}$$

subject to

$$\begin{aligned} \sum_{k=1}^{M} d_{ik} &= 1 \\ \sum_{k=1}^{M} d_{kj} &= 1 \end{aligned} \tag{52}$$

for all $i = 1, \ldots, M$, $j = 1, \ldots, M$. If we choose distance function (9) we obtain the same solution as in a broken random sample. Of course, there is little justification for this matching method if the samples A and B have no common units.

The method of constrained matching was first proposed by Barr and Turner

(1980). An advantage of this method is that the marginal distributions of $X$ and $Y$ in the merged file are the same as those in the samples A and B. A disadvantage is that the optimization problem in (51) is computationally burdensome.

In unconstrained matching the base file A and the supplemental file B are treated asymmetrically. To every unit $i$ in file A we match the unit $j$ in file B, possibly restricted to some subset defined by cohort variables, that minimizes $D(z_{1i}, z_{2j})$. It is possible that some unit in B is matched to more than one unit in A, and that some units in B are not matched to any unit in A. As a consequence, the distribution of $Z_2, Y$ in the matched file may differ from that in the original sample B. Note that if we use the distance function (18), unconstrained matching is formally identical to probabilistic record linkage. Of course, there is no justification for this method, if the samples A and B have no common units. The first application of unconstrained matching was by Okner (1972) who used the 1967 Survey of Economic Opportunity as the base file and the 1966 Tax File as the supplemental file to create a merged file that contained detailed data on the components of household income.

The merger of two files using either unconstrained or constrained matching has been criticized since its first use. In his comment on Okner's (1972) method, Sims (1972) noted that an implicit assumption on the conditional dependence of $X, Y$ given $Z$ is made, usually the assumption that $X, Y$ are independent conditional on $Z$. A second problem is best explained if we consider matching as an imputation method for missing data. File A contains $X, Z_1$ and $Y$ is

missing. If we assume conditional independence, an imputed value of $Y$ is a draw from the conditional distribution of $Y$ given $Z_1 = z_1$. Such a draw can be obtained from file B, if for one of the units in file B $Z_2 = z_1$. If such a unit is not present in file B, we choose a unit with a value of $Z_2$ close to $z_1$. This is an imperfect imputation, and we can expect that the relation between $Z_1$ and $Y$ in the merged file is biased. Indeed, Rodgers (1984) reports that the covariance between $Z_1$ and $Y$ is underestimated, as one would expect. An alternative would be to estimate the relation between $Y$ and $Z_2$ in sample B, e.g. by a linear regression, and use the predicted value for $Z_1 = z_1$, or preferably a draw from the estimated conditional distribution of $Y$ given $Z_1 = z_1$, i.e. include the regression disturbance variability in the imputation[6]. The imputation becomes completely dependent on model assumptions, if the support of $Z_1$ is larger than that of $Z_2$. In general the distribution of $X, Y, Z$ can only be recovered on the intersection of the supports of $Z_1$ and $Z_2$. If both samples are random samples from the same population, as we assume here, then the supports coincide.

It is possible to evaluate the quality of the data produced by a statistical match, by matching two independent subsamples from a larger dataset. The joint distribution in the matched sample can be compared to the joint distribution in the original dataset. Evaluation studies have been performed by, among others, Ruggles, Ruggles, and Wolff (1977), and Rodgers and DeVol (1982). It comes as no surprise that the conclusion from these evaluations is that the joint distribution of $X, Y$ cannot be estimated from the joint marginal distributions

[6]Even better: also include the variability due to parameter uncertainty.

of $X, Z$ and $Y, Z$.

As noted, matching can be considered as an imputation method for missing data. Rubin (1986) has suggested that instead of merging the files A and B, it is preferable to concatenate them, and to impute the missing $Y$ in file A and missing $X$ in file B using the estimated relations between $X$ and $Z_1$ (file A) and $Y$ and $Z_2$ (file B). In particular, he suggests not to use a single draw from the (estimated) conditional distribution of $X$ given $Z_1 = z_2$ and of $Y$ given $Z_2 = z_1$, effectively assuming conditional independence, but to add draws from the distributions of $X$ given $Z_1 = z_2, Y = y$ and $Y$ given $Z_2 = z_1, X = x$ assuming a range of values for the conditional correlation. The resulting datasets reflects the uncertainty on the conditional correlation and the variability of parameter estimates over the imputations indicates the sensitivity of these estimates to assumptions on the conditional correlation. Further developments along these lines can be found in Raessler (2002).

# 4 Estimation from independent samples with common variables

## 4.1 Types of inference

Without further assumptions the (conditional) Fréchet bound on the joint cdf is all that can be learned from the two samples. These bound is usually not sufficiently narrow, unless the common variables are highly correlated with $Y$ and $X$. In this section we explore what additional assumptions are needed to

improve the inference.

We consider (i) conditional independence, and (ii) exclusion restrictions. Exclusion restrictions refer to the situation that the distribution of $Y$ given $X, Z$ is independent of a subvector $Z_0^c$ of $Z$, and hence depends only on the other variables $Z_0$ in $Z$. We also consider both nonparametric inference, i.e. the goal is to estimate the joint distribution of $Y, X, Z_0$ or the conditional distribution of $Y$ given $X, Z_0$ or moments of these distributions, and parametric inference, i.e. the joint distribution of $Y, X, Z_0$ or the conditional distribution of $Y$ given $X, Z_0$ is in a parametric class. Parametric assumptions play an important role in inference from independent samples, a theme that is repeated in section 5 on inference in repeated cross-sections.

None of the methods discussed below requires that the two samples are merged. All computations can be done on the two samples separately.

## 4.2 Semi- and non-parametric inference

### 4.2.1 Conditional independence

If $Y, X$ are stochastically independent given the common variables $Z$, then the joint density of $X, Y$ is

$$f(x, y) = \mathrm{E}(f(x|Z)f(y|Z)) \tag{53}$$

Although the joint distribution is identified, often we just want to compute an expectation $E(g(X, Y)$. We have

$$\mathrm{E}(g(X,Y)) = \mathrm{E}_{YZ}(\mathrm{E}(g(X,Y) \mid Y,Z)) = \mathrm{E}_{YZ}(\mathrm{E}(g(X,Y) \mid Z)) \qquad (54)$$

where the last equality holds by conditional independence. Note that the inner conditional expectation is with respect to the distribution of $X$ given $Z$ that is identified from sample A, and that the outer expectation is with respect to the joint distribution of $Y, Z$ that is identified from sample B. We implicitly assume that the distributions of $Z_1$ and $Z_2$ in the samples A and B are identical. This is true if both samples are from the same population

For a fixed value of $Y$, we can estimate the inner conditional expectation by a nonparametric regression (e.g. kernel or series) estimator of $g(X, y)$ on $Z$ using sample A. The estimator of $E(g(X,Y)$ is then obtained by averaging this regression estimator over $Y, Z$ in sample B. The analysis of this estimator is beyond the scope of this chapter. It is similar to the semi-parametric imputation estimator proposed by Imbens, Newey, and Ridder (2004) and Chen, Hong, and Tarrozi (2004) who establish semi-parametric efficiency for their estimator. Their results can be directly applied to this estimator. In the literature it has been suggested that for the estimation of $E(g(X,Y)$ we must first estimate the joint distribution of $X, Y$ (see Sims ,1972 and Rubin ,1986), but this is not necessary. Note that a similar method can be used to estimate $E(g(X,Y,Z_0))$ with $Z_0$ a subvector of $Z$.

### 4.2.2 Exclusion restrictions

If we are not prepared to assume that $X, Y$ are conditionally independent given $Z$, we can only hope for bounds on the expected value $E(g(X, Y, Z_0))$. Such bounds are given by Horowitz and Manski (1995) and Cross and Manski (2002) and can be derived in the same way as the bounds in section 3.1. In particular, they derive bounds on $\mathrm{E}\left[g(h(Y, Z_0), X, Z_0) \mid X = x, Z_0 = z_0\right]$ with $g$ bounded and monotone in $h$ for (almost all) $x, z_0$.

We consider two possibilities: (i) the conditional distribution of $Y$ given $X, Z$ depends on all variables in $Z$, (ii) this conditional distribution only depends on a subvector $Z_0$ of $Z$ and is independent of the other variables $Z_0^c$ in $Z$. Note that the possibilities are expressed in terms of the conditional distribution of $Y$ given $X$ (and $Z$ or $Z_0$). This suggests that $Y$ is considered as the dependent variable and that $X, Z$ are explanatory variables.

If assumption (i) applies, the bounds derived above are bounds on $F(y; . \mid Z = z)$ or $F(y \mid ., Z = z))$. If we are interested in $F(y; .)$ or $F(y \mid .)$, we have to average over the marginal distribution of $Z$ or the conditional distribution of $Z$ given $X = x_k$ ($F(y \mid X = x_k, Z)$ has to be averaged over this distribution). As noted in section 3.1 this averaging results in narrower bounds, but as noted in that section the correlation between $Y$ and $Z$ and $X$ and $Z$ must be high to obtain informative bounds.

Assumption (ii) that states that the vector of common variables $Z_0^c$ can be omitted from the relation between $Y$ and $X, Z$ is more promising. As stated, assumption (ii) focuses on conditional (in)dependence of $Y$ and $Z_0^c$ given

$X, Z_0$. Alternatively, the assumption can be expressed as conditional mean (in)dependence or conditional quantile (in)dependence. In that case, we identify or obtain bounds on the conditional mean or quantile. We only discuss conditional (in)dependence. The derivation of bounds on the conditional mean from bounds on the conditional cdf is complicated by the fact that the conditional mean is not a continuous function of the conditional cdf. However, if the assumptions are expressed as restrictions on the conditional mean, this does not matter.

Assumption (ii) is an exclusion restriction. If we decompose $Z = (Z_0' Z_0^{c'})'$, then $Z_0^c$ is excluded from the conditional distribution of $Y$ given $X, Z$. Exclusion restrictions are powerful and often are sufficient to identify $F(y \mid x, z_0)$. We maintain the assumption that $X$ is discrete. This simplifies the analysis substantially. This is not surprising, because nonparametric identification under exclusion restrictions is an inverse problem, and it is well-known that inverse problems are much harder for continuous distributions (see e.g. Newey and Powell (2003)). First, we consider conditions under which $F(y \mid x, z_0)$ is nonparametrically identified. Next, we consider the underidentified case, and we show that we can find bounds that improve on the bounds that hold without an exclusion restriction.

Without loss of generality we omit $Z_0$. The common variable $Z$ is excluded from the conditional cdf of $Y$ given $X, Z$. We denote

$$\Pr(X = x_k \mid Z = z) = p_k(z) \tag{55}$$

With the exclusion restriction we have that for all $z$

$$F(y \mid z) = \sum_{k=1}^{K} F(y \mid x_k) p_k(z) \tag{56}$$

If $Z$ is also discrete, (56) is a linear system of equations with unknowns $F(y \mid x_k)$, i.e. $K$ unknowns. Hence, this system has a unique solution if $Z$ takes at least $L \geq K$ values and the $L \times K$ matrix, with $(l,k)$-th component $p_k(z_l)$ has rank equal to $K$. In that case $F(y \mid .)$ is exactly identified. If the rank of this matrix is strictly greater than $K$ (this requires that $L > K$), then the equation has no solution. Hence, if $L > K$ a test of the rank of the matrix, and in particular a test whether the rank is equal to $K$ is a test of the overidentifying restrictions, or in other words, a test of the exclusion restriction. If the exclusion restriction is rejected, we can allow the conditional cdf of $Y$ given $X, Z$ to depend on $Z$. For instance, if $X$ takes two values and $Z$ contains two variables, of which the first takes two values and the second four, then we obtain an exactly identified model by allowing the conditional cdf to depend on the first variable in $Z$.

If $X$ and $Z$ take two values, i.e. $K = L = 2$, the solutions to (56) is

$$\begin{aligned} F(y \mid x_1) &= \frac{p_2(z_1)F(y \mid z_2) - p_2(z_2)F(y \mid z_1)}{p_1(z_2) - p_1(z_1)} \\ F(y \mid x_2) &= \frac{p_1(z_2)F(y \mid z_1) - p_1(z_1)F(y \mid z_2)}{p_1(z_2) - p_1(z_1)} \end{aligned} \tag{57}$$

Note that this implies that

$$F(y \mid x_2) - F(y \mid x_1) = \frac{F(y \mid z_2) - F(y \mid z_1)}{p_1(z_2) - p_1(z_1)} \tag{58}$$

If conditional cdf's are replaced by conditional expectations, this is the Wald estimator (Wald, 1940), which is the Instrumental Variable (IV) estimator for a dichotomous endogenous variable with a dichotomous instrument.

Solving (56) for the case that $X$ is continuous is much harder. In effect, we have to find the components of a mixture in the case that the mixing distribution is known. The problem is that the solution is not continuous in $F(y \mid .)$ unless restrictions are imposed on these conditional distributions. For instance, if $Z$ is independent of $Y, X$ (exclusion restriction) and the joint distribution of $Y, X$ is normal, then the covariance of $Y, X$ can be recovered from

$$\mathrm{E}(Y \mid Z = z) = \mu_Y + \Sigma_{YX}\Sigma_{XX}^{-1}(\mathrm{E}(X \mid Z = z) - \mu_X) \tag{59}$$

with $\mu$ the mean and $\Sigma$ the covariance matrix of the joint normal distribution. Further details on weaker restrictions can be found in Newey and Powell (2000).

The similarity of the nonparametric two-sample estimator and the corresponding IV estimator with endogenous $X$ and $Z$ as instrumental variable, can lead (and as will be noted in section 4.4 has led) to much confusion. In particular, it does not mean that we should consider $X$ as an endogenous variable.

If $L < K$ the conditional cdf $F(y|.)$ is not identified. In that case we can use the results in Horowitz and Manski (1995) and Cross and Manski (2002) to obtain bounds (see the discussion in section 3.1). The exclusion restriction imposes additional restrictions on the conditional cdf. Figure 3 illustrates these

bounds for the case $K = 3, L = 2$. In this figure the two triangles give the sets of $F(y|x_1), F(y|x_2), F(y|x_3)$ that are consistent with sample information if $Z = z_1$ or $Z = z_2$. Because $Z$ takes both values and is excluded from the conditional distribution of $Y$ given $X = x$, $F(y|x_1), F(y|x_2), F(y|x_3)$ has to be in the intersection of these triangles. Note that the extreme points are the Wald estimators of $F(y|x_1), F(y \mid x_3)$ and $F(y|x_2), F(y|x_3)$ for the case that $F(y|x_2)$ and $F(y|x_1)$ are set to 0, respectively. In general the extreme points are Wald estimators for conditional cdf's that are obtained by imposing identifying restrictions. Figure 3 is drawn for $p_k(z_l) \leq \frac{1}{2},\ k = 1, 2, 3,\ l = 1, 2$ and $y < \min\{F^{-1}(p_k(z_l)),\ k = 1, 2, 3,\ l = 1, 2\}$. The other bounds can be obtained in the same way. Note that the exclusion restriction gives a narrower bound. To see this, compare the bound on $F(y|x_1)$ in the figure to those for $Z = z_1$ or $Z = z_2$, which are 0 (lower bound) and $\frac{F(y|z_1))}{p_1(z_1)}$ and $\frac{F(y|z_2))}{p_1(z_2)}$ (upper bound), respectively.

## 4.3 Parametric inference

### 4.3.1 Conditional independence

Often two samples are merged to estimate a parametric relation between a dependent variable $Y$, present in one sample, and a vector of independent variables $X$ some of which may be only present in an independent sample. We assume that sample A contains $X, Z$, sample B contains $Y, Z$ and that we estimate a relation between $Y$ and $X, Z_0$ with $Z_0$ a subvector of $Z$. This relation has a vector of parameters $\theta$ and we assume that the population parameter vector $\theta_0$

is the unique solution to the population moment conditions

$$\mathrm{E}(m(Y, X, Z_0; \theta)) = 0 \tag{60}$$

This framework covers Maximum Likelihood (ML) and Generalized Method of Moments (GMM). Initially, we assume that $X$ and $Y$ are conditionally independent given $Z$.

Under conditional independence we have

$$\mathrm{E}(m(Y, X, Z_0; \theta)) = \mathrm{E}_{YZ}(\mathrm{E}_X(m(Y, X, Z_0; \theta) \mid Y, Z)) = \mathrm{E}_{YZ}(\mathrm{E}_X(m(Y, X, Z_0; \theta) \mid Z)) \tag{61}$$

If we we have an estimate of the conditional distribution of $X$ given $Z$, identified in sample A, we can estimate $\mathrm{E}(m(y, X, z_0; \theta) \mid Z = z)$ for fixed values $Y = y$ and $Z = z$ using the data from sample A. The sample moment conditions corresponding to (61) are

$$\frac{1}{N_2} \sum_{j=1}^{N_2} \widehat{\mathrm{E}_{X|Z}}(m(Y_j, X, Z_{02j}; \theta) \mid Z_{2j}) = 0 \tag{62}$$

where the hat indicates that the conditional expectation is estimated using the data from sample A.

As an example consider the regression model

$$Y = \beta_1 X + \beta_2 Z_0 + \varepsilon \tag{63}$$

The scalar dependent variable $Y$ and a vector of common variables $Z_1$ are ob-

served in sample A. The (scalar) independent variable $X$ and a vector of the common variables $Z_2$ are observed in sample B (the subscript on $Z$ indicates the sample). We assume that $Z_1$ and $Z_2$ are independently and identically distributed. The scalar variable $Z_0$ is a component of $Z$. The parameters $\beta_1, \beta_2$ are identified by

$$\mathrm{E}(\varepsilon \mid X, Z) = 0 \tag{64}$$

In general this assumption is too strong, because it generates more moment conditions than are needed to identify the regression parameters. These parameters are identified, even if (scalar) $X$ is correlated with $\varepsilon$, provided that $Z$ has two variables that are not correlated with $\varepsilon$. In general, $Z$ is chosen to ensure that the variables in the relation that are in sample A and those that are in sample B are conditionally independent given $Z$, and $Z$ may contain many variables. It is not even necessary to assume that all the variables in $Z$ are exogenous, as suggested by (64). If $X$ is exogenous, only $Z_0$ (or one other variable in $Z$) has to be exogenous.

We first consider the case that both $X$ and $Z_0$ are exogenous. The population moment conditions are

$$\begin{aligned} \mathrm{E}[(Y - \beta_1 X - \beta_2 Z_0)X] &= 0 \\ \mathrm{E}[(Y - \beta_1 X - \beta_2 Z_0)Z_0] &= 0 \end{aligned} \tag{65}$$

Under conditional independence these can be written as

$$\mathrm{E}_{YZ_2}[Y\mathrm{E}_{X|Z_1}(X \mid Z_2) - \beta_1\mathrm{E}_{X|Z_1}(X^2 \mid Z_2) - \beta_2 Z_{02}\mathrm{E}_{X|Z_1}(X \mid Z_2)] = 0 \quad (66)$$

$$\mathrm{E}_{YZ_2}[(Y - \beta_1\mathrm{E}_{X|Z_1}(X \mid Z_2) - \beta_2 Z_{02})Z_{02}] = 0 \quad (67)$$

In these expressions $\mathrm{E}_{X|Z_1}(X \mid Z_2)$ is the conditional expectation of $X$ given $Z_1$ that can be estimated from sample A and that is a function of $Z_1$, with $Z_2$ substituted for $Z_1$. In other words, it is the imputed $X$ in sample B based on $Z_2$ observed in sample B and using the conditional expectation of $X$ given $Z_1$ in sample A.

If we substitute the sample moments for $\mathrm{E}_{YZ_2}[Y\mathrm{E}_{X|Z_1}(X \mid Z_2)]$, $\mathrm{E}_{YZ_2}[\mathrm{E}_{X|Z_1}(X \mid Z_2)]$, $\mathrm{E}_{YZ_2}[\mathrm{E}_{X|Z_1}(X^2 \mid Z_2)]$, and $\mathrm{E}_{YZ_2}[Z_{02}\mathrm{E}_{X|Z_1}(X \mid Z_2)]$, we obtain the sample moment conditions that can be solved to obtain the estimator of the regression coefficients. From GMM theory (Hansen, 1982) it follows that this estimator is consistent and asymptotically normal . If the number of moment conditions exceeds the number of parameters, we obtain an efficient estimator by minimizing a quadratic form in the sample moment conditions with the inverse of the variance matrix of these conditions as weighting matrix.

It is interesting to note that the GMM estimator obtained from (66)-(67) is not the imputation estimator obtained by replacing the unobserved $X$ in sample B by its imputed value. The imputation estimator is not even available, if $X$ and $Z_0$ are both exogenous and $Z = Z_0$.

If $Z$ contains at least one additional exogenous variable, $Z_0^c$, we can choose to use the moment condition corresponding to $Z_0^c$, instead of the moment condition corresponding to $X$, even if $X$ is exogenous. In that case we can replace the moment conditions (65) by

$$\begin{aligned} \mathrm{E}[(Y - \beta_1 X - \beta_2 Z_0) Z_0^c] &= 0 \\ \mathrm{E}[(Y - \beta_1 X - \beta_2 Z_0) Z_0] &= 0 \end{aligned} \tag{68}$$

Because the $Z$'s are in both samples, all expected values in these population moment conditions can be obtained from sample A ($\mathrm{E}(XZ_0), \mathrm{E}(XZ_0^c)$), sample B ($\mathrm{E}(YZ_0), \mathrm{E}(YZ_0^c)$) or both ($\mathrm{E}(Z_0^2), \mathrm{E}(Z_0 Z_0^c)$). Hence, in this case we need not make the assumption of conditional independence of $X$ and $Y$ given $Z$. Note that this is true, irrespective of whether $X$ is endogenous or not. Key are the availability of additional common variables that can replace $X$ in the moment conditions and the additive separability of variables that are in different samples in the residual $Y - \beta_1 X - \beta_2 Z_0$. We shall explore this below.

In the example the distribution of $X$ given $Z$ was not needed to obtain the GMM estimator, because the moment conditions were quadratic in $X$ and only $\mathrm{E}(X \mid Z)$ and $\mathrm{E}(X^2 \mid Z)$ had to be estimated. In general, this will not be the case, and an assumption on this conditional distribution is needed. Econometricians are usually reluctant to specify the distribution of exogenous variables, and for that reason we may consider a semi-parametric alternative in

which $\mathrm{E}_{X|Z_1}(m(y, X, z_0; \theta) \mid Z = z)$ is estimated by a nonparametric regression (series or kernel estimator) of $m(y, X_i, z_0; \theta)$ on $Z_{1i}$ in sample A. This gives $\hat{\mathrm{E}}_{X|Z}(m(y, X, z_0; \theta))$ which is substituted to obtain the sample moment conditions as an average in sample B. This estimator is similar to the estimator considered in Chen, Hong and Tarrozi (2004) and Imbens, Newey, and Ridder (2004), and their results can be used to analyze this estimator.

### 4.3.2 Exclusion restrictions

In section 4.2.2 we discussed conditions under which exclusion restrictions are sufficient for the nonparametric identification of the conditional distribution of $Y$ given $X, Z_0$. In this section we consider parametric inference. The assumptions we impose are convenient, but stronger than needed. In particular, we restrict the discussion to additively separable moment conditions. The existing literature only considers this case. If the exclusion restrictions identify the joint distribution as explained in section 4.2.2, the separability assumption can be relaxed. This has not been studied, and developing procedures for this case is beyond the scope of this chapter.

The setup and notation is as in section 4.2.2 with $Z_0^c$ the components of $Z$ that are not in the relation and satisfy (69), i.e. that are exogenous for the relation between $Y$ and $X, Z_0$. We consider moment conditions that can be written as

$$\mathrm{E}((f(Y; \theta) - g(X, Z_0; \theta))h(Z_0, Z_0^c)) = 0 \tag{69}$$

with $f, g, h$ known functions and $\theta$ a vector of parameters. If $Y$ is scalar, then so is $g$. The dimension of $h$ is not smaller than that of $\theta$. In general, this implies that the dimension of $Z_0^c$ has to exceed that of $X$[7], i.e. the number of common exogenous variables that is excluded from the relation can not be smaller than the number of variables in $X$. If we assume that some variables in either $X$ or $Z_0$ are endogenous we need as many additional variables in $Z_0^c$ as there are endogenous variables among $X, Z_0$.

The estimator based on the population moment conditions (69) is called the Two-sample Instrumental Variable (2SIV) estimator. In the case that all variables are observed in a single sample, the estimator based on the moment conditions in (69) is related to Amemiya's nonlinear simultaneous equations estimator (see e.g. Amemiya, 1985, Chapter 8).

We discuss three examples of models that give moment conditions as in (69): the linear regression model, the probability model for discrete dependent variables, and the mixed proportional hazard model for duration data. In all models we take $h(Z_0, Z_0^c) = (Z_0' \;\; Z_0^{c'})'$. For the linear regression model the moment conditions are

$$\mathrm{E}(Y - \beta_0 - \beta_1' X - \beta_2' Z_0) = 0 \tag{70}$$

$$\mathrm{E}((Y - \beta_0 - \beta_1' X - \beta_2' Z_0) Z_0) = 0 \tag{71}$$

[7]If $Z_0$ is exogenous, then functions, e.g. powers, of $Z_0$ are also exogenous. To avoid identification by functional form, we need the additional exogenous variables in $Z_0^c$.

$$\mathrm{E}((Y - \beta_0 - \beta_1' X - \beta_2' Z_0) Z_0^c) = 0 \tag{72}$$

Note that we can replace $X$ by $\mathrm{E}(X \mid Z_0, Z_0^c)$[8]. We can even replace $X$ by the linear approximation to this conditional expectation, i.e. by $\pi_0 + \pi_1' Z_0 + \pi_2' Z_0^c$ where the vector $\pi$ minimizes $E[(X - \pi_0 - \pi_1' Z_0 - \pi_2' Z_0^c)^2]$. This gives the estimating equations of the two-stage linear imputation estimator first suggested by Klevmarken (1982). In the first stage, the vector of independent variables $X$ is regressed on the common exogenous variables $Z_0, Z_0^c$ using data from sample A. This estimated relation is used to compute the predicted value of $X$ in sample B, using the common variables as observed in sample B. These predicted values are substituted in the estimating equations that now only contain variables observed in sample B.

The second example is the probability model for discrete dependent variables. If we consider a dummy dependent variable then we specify

$$\Pr(Y = 1 \mid X, Z_0) = G(\beta_0 + \beta_1' X + \beta_2' Z_0) \tag{73}$$

with $G$ a cdf of some continuous distribution, eg. the standard normal (Probit) or logistic cdf (Logit). The moment conditions are

$$\mathrm{E}(Y - G(\beta_0 + \beta_1' X + \beta_2' Z_0)) = 0 \tag{74}$$

$$\mathrm{E}((Y - G(\beta_0 + \beta_1' X + \beta_2' Z_0) Z_0) = 0 \tag{75}$$

[8]This is a consequence of the equivalence of 2SLS and IV in this type of models

$$\mathrm{E}((Y - G(\beta_0 + \beta_1' X + \beta_2' Z_0) Z_0^c) = 0 \tag{76}$$

Except for the logit model, these moment conditions do not give the efficient estimator of $\beta$. To obtain the efficient estimator we must multiply the residual by

$$\frac{g(\beta_0 + \beta_1' X + \beta_2' Z_0)}{G(\beta_0 + \beta_1' X + \beta_2' Z_0)(1 - G(\beta_0 + \beta_1' X + \beta_2' Z_0))} \tag{77}$$

The resulting moment equation can not be computed from the separate samples. Ichimura and Martinez-Sanchis (2005) discuss this case and also derive bounds on the parameters if there is no point identification.

The last example is the Mixed Proportional Hazard (MPH) model for duration data. In that model the hazard rate $h$ of the duration Y is specified as

$$h(y \mid x, V; \theta) = \lambda(y; \theta_1) \exp\{\theta_2' X + \theta_3' Z_0\} V \tag{78}$$

with $\lambda$ the baseline hazard and $V$ a random variable that is independent of $Z_0, Z_1$ and that captures the effect of omitted variables. By (78) we have that

$$\ln \Lambda(Y; \theta_1) + \theta_2' X + \theta_3' Z_0 = U \tag{79}$$

with $U$ independent of $Z_0, Z_1$ and $\Lambda$ the integral of $\lambda$. This gives the moment conditions

$$\mathrm{E}((\ln \Lambda(Y; \theta_1) + \theta_2' X + \theta_3' Z_0) Z_0) = 0 \tag{80}$$

$$\mathrm{E}((\ln \Lambda(Y;\theta_1) + \theta_2' X + \theta_3' Z_0) Z_0^c) = 0 \tag{81}$$

The number of variables in $Z_0^c$ must at least be equal to the number of parameters in $(\theta_1' \; \theta_2')'$[9]. Alternatively, we can identify $\theta_1$ by making assumptions on the functional form of the regression function. For instance, if we maintain the hypothesis that the regression function is linear, we can use powers of the variables in $Z_0^c$ in the moment conditions. In that case no additional common variables are needed[10]. Besides the MPH model, we can estimate other transformation models from two independent samples. Examples are the Box-Cox transform (Box and Cox, 1964) and the transform suggested by Burbidge, Magee, and Robb (1988)[11].

These three examples correspond to linear regression, nonlinear regression and transformation models. Other models, as the Tobit model, can also be estimated with this type of data. For the Tobit model we can employ the two-part estimation method that yields moment conditions as in (69). Only in the linear regression model is the GMM estimator equivalent to a (linear) imputation estimator. In the other examples, imputation yields biased estimates.

The additional common variables $Z_0^c$ must be exogenous. They also have

[9] If we the baseline hazard is Weibull we can identify the regression parameters up to scale. These parameters can be identified, if we choose a functional form for the baseline hazard that is not closed under a power transformation.

[10] Provided that the identification condition (A3) below is satisfied.

[11] The latter transform is used by Carroll, Dynan, and Krane (1999) who use two independent samples to estimate their regression model. Because their model has a 'missing parameter' and not a missing regressor, they do not use 2SIV.

to be correlated with the variables in $X$. In other words, they must satisfy the requirements for valid instruments for $X$, irrespective whether the variables in $X$ are exogenous or endogenous. As noted before, the separability of the moment conditions is a sufficient, but not necessary condition for identification.

The asymptotic distribution theory of the 2SIV estimator based on (69) raises some new issues. First, we introduce some notation. Let

$$m(\theta) = (f(Y;\theta) - g(X, Z_0;\theta))h(Z_0, Z_0^c) \tag{82}$$

and for $i = 1, \ldots N_1$, $j = 1, \ldots N_2$

$$\begin{aligned} m_{2j}(\theta) &= f(Y_j;\theta)h(Z_{02j}, Z_{02j}^c) \\ m_{1i}(\theta) &= g(X_i, Z_{01i};\theta)h(Z_{01i}, Z_{01i}^c) \end{aligned} \tag{83}$$

with the second subscript in e.g. $Z_{01i}$ indicating that the common included exogenous variable $Z_0$ is observed in sample A etc. Using this notation, the sample moment conditions are

$$m_N(\theta) = \frac{1}{N_2}\sum_{j=1}^{N_2} m_{2j}(\theta) - \frac{1}{N_1}\sum_{i=1}^{N_1} m_{1i}(\theta) \tag{84}$$

We make the following assumptions (the derivatives in the assumptions are assumed to exist and to be continuous in $\theta$)

(A1) The common variables in samples A and B, the random vectors $Z_{01}, Z_{01}^c$ and $Z_{02}, Z_{02}^c$ are independently but identically distributed.

(A2) If $N_1, N_2 \to \infty$

$$\frac{\partial m_N}{\partial \theta'}(\theta) \xrightarrow{p} \mathrm{E}\left(\frac{\partial m}{\partial \theta'}(\theta)\right)$$

uniformly for $\theta \in \Theta$ with $\Theta$ the parameter space.

(A3) The rank of the matrix $\mathrm{E}\left(\frac{\partial m}{\partial \theta'}(\theta_0)\right)$ is equal to the dimension of $\theta$.

Assumption (A1) ensures that the limit in (A2) holds pointwise for every $\theta \in \Theta$. Assumption (A3) is the identification condition. The probability limit of the derivative of the moment conditions is

$$\mathrm{E}\left(\frac{\partial m}{\partial \theta'}(\theta)\right) = \mathrm{E}\left(\frac{\partial f(Y;\theta)}{\partial \theta'} h(Z_{02}, Z_{02}^c)\right) - \mathrm{E}\left(\frac{\partial g(X, Z_{01};\theta)}{\partial \theta'} h(Z_{01}, Z_{01}^c)\right) \quad (85)$$

This matrix can be estimated consistently from the samples A and B, because the expectations only involve variables that are observed in the same sample.

The 2SIV is formally defined by

$$\hat{\theta}_N = \arg\min_{\theta \in \Theta} m_N(\theta)' W_N m_N(\theta) \quad (86)$$

with $W_N$ a weighting matrix that satisfies

$$W_N \xrightarrow{p} W \quad (87)$$

with $W$ a positive definite matrix and $N \to \infty$ if $N_1, N_2 \to \infty$. In the appendix we show that assumptions (A1)-(A3) are sufficient for weak consistency of the 2SIV.

If (A1) does not hold, the 2SIV is biased. The probability limit is the minimizer of

$$(\theta-\theta_0)'\mathrm{E}\left[\frac{\partial m'}{\partial\theta}(\theta_*)\right]W\mathrm{E}\left[\frac{\partial m}{\partial\theta'}(\theta_*)\right](\theta-\theta_0)+ \tag{88}$$

$$2\mathrm{E}[m(\theta_0)]'W\mathrm{E}\left[\frac{\partial m}{\partial\theta'}(\theta_*)\right](\theta-\theta_0)+\mathrm{E}[m(\theta_0)]'W\mathrm{E}[m(\theta_0)]$$

but the last two terms do not vanish. We can use this expression to find the asymptotic bias of the 2SIV estimator.

The optimal weight matrix $W$ is the inverse of the variance matrix of $m_N(\theta_0)$. To derive the asymptotic variance matrix we have to make an assumption on the rate at which the sample sizes increase. Such an assumption was not needed to establish weak consistency of the 2SIV estimator. We assume

(A4) $\lim_{N_1\to\infty,N_2\to\infty}\frac{N_2}{N_1}=\lambda$ with $0<\lambda<\infty$.

Consider, using the fact that $E(m(\theta_0))=0$ if (A1) is true,

$$\sqrt{N_2}m_N(\theta_0)= \tag{89}$$

$$=\frac{1}{\sqrt{N_2}}\sum_{j=1}^{N_2}(m_{2j}(\theta_0)-\mathrm{E}(m_{2j}(\theta_0))-\sqrt{\frac{N_2}{N_1}}\frac{1}{\sqrt{N_1}}\sum_{i=1}^{N_1}(m_{1i}(\theta_0)-\mathrm{E}(m_{1i}(\theta_0))$$

Hence, the asymptotic variance matrix of the moment conditions is

$$M(\theta_0)=\lim_{N_2\to\infty}\mathrm{E}[N_2m_N(\theta_0)m_N(\theta_0)']=\lambda\mathrm{Var}(m_{2j}(\theta_0))+\mathrm{Var}(m_{1i}(\theta_0)) \tag{90}$$

and the inverse of this matrix is the optimal choice for $W(\theta_0)$. This matrix can be easily estimated if we have an initial consistent estimator. Note that by the

central limit theorem for i.i.d. random variables (if the asymptotic variance is finite) $\sqrt{N_2}m_N(\theta_0)$ converges to a normal distribution with mean 0. However, if (A1) does not hold and as a consequence $E(m(\theta_0)) \neq 0$, the mean diverges. This will affect the interpretation of the test of overidentifying restrictions that will be discussed below.

Under (A1)-(A4)

$$\sqrt{N_2}(\hat{\theta}_N - \theta_0) \xrightarrow{d} N(0, V(\theta_0)) \tag{91}$$

with

$$\begin{aligned} V(\theta_0) = &\left[\mathrm{E}\left(\frac{\partial m'}{\partial \theta}(\theta_0)\right) W(\theta_0)\mathrm{E}\left(\frac{\partial m}{\partial \theta'}(\theta_0)\right)\right]^{-1} \cdot \\ &.\mathrm{E}\left(\frac{\partial m'}{\partial \theta}(\theta_0)\right) W(\theta_0)(\lambda \mathrm{Var}(m_{2j}(\theta_0)) + \mathrm{Var}(m_{1i}(\theta_0)))W(\theta_0)\mathrm{E}\left(\frac{\partial m}{\partial \theta'}(\theta_0)\right) \cdot \\ &.\left[\mathrm{E}\left(\frac{\partial m'}{\partial \theta}(\theta_0)\right) W(\theta_0)\mathrm{E}\left(\frac{\partial m}{\partial \theta'}(\theta_0)\right)\right]^{-1} \end{aligned} \tag{92}$$

See the appendix for a proof.

The preceding discussion suggest a two-step procedure. In the first step we use a known weight matrix, e.g. $W_N = I$. The resulting 2SIV estimator is consistent, but not efficient. In the second step, we first estimate the optimal weight matrix, the inverse of (90). This matrix only depends on the first-step consistent estimator and moments that can be computed from the two independent samples A and B (for $\lambda$ we substitute $\frac{N_2}{N_1}$). Next, we compute the efficient 2SIV estimator (86) with this weight matrix. This estimator has asymptotic variance

$$\left[\mathrm{E}\left(\frac{\partial m'}{\partial \theta}(\theta_0)\right)\left(\lambda \mathrm{Var}(m_{2j}(\theta_0)) + \mathrm{Var}(m_{1i}(\theta_0))\right)\mathrm{E}\left(\frac{\partial m}{\partial \theta'}(\theta_0)\right)\right]^{-1} \tag{93}$$

which can be estimated from the independent samples.

In general, the efficient 2SIV estimator is less efficient than efficient estimators based on a sample that contains all the variables. In the case that the information matrix only depends on variables in sample A, we can estimate the variance of the efficient estimator, even if the estimator itself can not be computed from the independent samples. The inverse of the information matrix gives an indication of the efficiency loss, due to the fact that we do not have a sample that has all variables.

If the number of moment conditions is larger than the number of parameters, we can test the overidentifying restrictions. The test statistic is

$$T_N = N_2 m_N(\hat{\theta}_N)' \left[\frac{N_2}{N_1}\widehat{\mathrm{Var}}(m_{2j}(\hat{\theta}_N)) + \widehat{\mathrm{Var}}(m_{1i}(\hat{\theta}_N))\right]^{-1} m_N(\hat{\theta}_N) \tag{94}$$

where $\widehat{\mathrm{Var}}$ denotes the sample variance. If (A1)-(A4) hold, then $T_N \xrightarrow{d} \chi^2(\dim(m) - \dim(\theta))$. The appendix contains a proof.

As noted before, rejection of the overidentifying restrictions indicates that either some of the common variables that are used as instruments are not exogenous or that they are not identically distributed in the samples A and B.

Although the technique of choice for estimating relations from combined samples has been GMM, Maximum Likelihood can be used as well. A reason for the preference for GMM (or IV) may be that in that framework it is easier

to obtain consistent estimates of structural parameters if some of the regressors are endogenous. Orthogonality conditions for equation errors and instrumental variables are more natural in GMM. To define the Two-Sample Maximum Likelihood (2SML) estimator we start with a parametric model for the conditional distribution of $Y$ given $X, Z_0$, $f(y \mid x, z_0; \theta)$. Because $X$ is not observed in sample A, we use sample B to estimate the conditional density of $X$ given $Z_0, Z_1$. We can use a parametric or a non-parametric estimator for the latter conditional density. The likelihood contributions are obtained from the conditional density of $Y$ given $Z_0, Z_1$

$$f(y \mid z_0, z_1; \theta) = \int f(y \mid x, z_0; \theta) g(x \mid z_0, z_1) \mathrm{d}x \tag{95}$$

With a parametric estimator for $g(x \mid z_0, z_1)$ the 2SML estimators is a conventional MLE with all the usual properties. The properties of the 2SML with a non-parametric estimator of this conditional density have not been studied. In section 4.2.2 we considered nonparametric identification of $f(y|x_1, z_0)$, and nonparametric identification is sufficient for parametric identification. Again Chen, Hong, and Tarrozi (2004) and Imbens, Newey, and Ridder )(2004) provide the framework in which the 2sML can be analyzed.

2SIV or 2SML are used if some of the explanatory variables in a relation are not measured in the same sample as the dependent variable. Another situation occurs in models with generated regressors, in which the parameters of the generated regressor cannot be estimated from the same sample. An important example of a generated regressor is the sample selection correction function. An

example is the estimation of a wage equation on a sample of working individuals. This yields biased estimates of the regression coefficients if a positive fraction of the population under consideration does not work. A method to reduce this bias is to include a sample selection correction function (Heckman, 1979). The parameters of this function cannot be estimated from the sample of working individuals. However, if an independent sample is available that contains both working and non-working individuals but no information on wages, then the parameters can be estimated from this sample. This allows us to compute the sample selection correction for the working individuals.

Another example of a generated regressor is Carroll, Dynan, and Krane (1999) who estimate the effect of the probability of becoming unemployed on the wealth to income ratio. They estimate the wealth equation with data from the Survey of Consumer Finances (SCF). However, the SCF has no information on unemployment. The probability of becoming unemployed is estimated from the Current Population Survey (CPS) and because the variables that enter this probability are also observed in the SCF, this probability can be imputed in the SCF. Note that in these examples there are no missing variables. Only the parameters that enter the generated regressor are estimated from an independent sample. This type of data combination can be treated as any estimation problem with a generated regressor (Pagan, 1984). The fact that the parameter is estimated from an independent sample even simplifies the distribution theory.

## 4.4 The origin of two-sample estimation and applications

Of the methods discussed in this section only the 2SIV estimator is prominent in econometrics. The first author who suggested this estimator was Klevmarken (1982). Since then it was rediscovered independently by Angrist and Krueger (1992) and Arellano and Meghir (1992)[12]. Klevmarken derives the 2SIV estimator for a single equation that is part of a system of linear simultaneous equations. In our notation he considers

$$Y = \beta_0 + \beta_1' X + \beta_2' Z_0 + \varepsilon \tag{96}$$

with $X$ observed in sample A and $Y$ in sample B, while $Z_0$ is a subvector of the common variables $Z$. He also assumes that all the variables in $X$ are endogenous[13], that all the common variables $Z$ are exogenous and that $Z$ contains all exogenous variables[14]. If we compare these assumptions with ours, we see that Klevmarken's assumptions are far too strong and limit the application of 2SIV to rather special cases. In particular, the assumption that $Z$ contains all exogenous variables seems to be inspired by a desire to give a structural interpretation to the first-stage imputation regression, in which $X$ is regressed on the exogenous variables in $Z$. Such an interpretation is not needed, and hence the only requirement is the order condition discussed in the previous subsection. Moreover, not all common variables need to be exogenous, as long as this order condition is satisfied. Finally, some of the variables in $X$ may be exogenous.

[12] These authors do not cite Klevmarken's contribution

[13] Klevmarken (1982), p. 160

[14] Klevmarken (1982), p. 159

Klevmarken states that we can only allow for exogenous variables if the joint distribution of $X$ and $Z$ is multivariate normal, which ensures that the conditional mean of $X$ given $Z$ is linear in $Z$. As the derivation in the previous subsection shows, a linear conditional mean is not essential for the 2SIV estimator. In the linear regression model replacing the conditional expectation by the linear population projection on $Z$ will not affect the moment conditions[15] and hence the assumption of multivariate normality is not needed. Carroll and Weil (1994) start from the same model as Klevmarken. They claim[16] that to compute the variance matrix of the 2SIV estimator it is required that in one of the datasets we observe $Y, X, Z$. The discussion in the previous subsection shows that this is not necessary. The problem with their approach is that their estimator of the variance matrix requires the residuals of the regression and these cannot be recovered from the independent samples.

At this point, we should clarify the role of endogenous and exogenous regressors in 2SIV estimation. The natural solution to missing variables in a statistical relation is imputation of these variables. Indeed, the 2SIV estimator in the linear regression model can be seen as an imputation estimator. Econometricians are used to imputation if the regression contains some endogenous variables. In the Two-stage Least Squares (2SLS) estimator the endogenous variables are replaced by a predicted or imputed value. Hence, it is not surprising that 2SIV was originally developed for linear regression models with endogenous regressors. Our derivation shows that such a restriction is not necessary, and in particular,

[15] Provided that the distribution of the common variables in the two samples is the same.

[16] See the Technical Appendix t their paper.

that the 2SIV only imputes missing variables, if the model is a linear regression. In the general case specified in (69), there is no imputation of missing variables.

After Klevmarken (1982) the 2SIV estimator was reinvented independently by Arellano and Meghir (1992) and Angrist and Krueger (1992). Arellano and Meghir consider moment restrictions of the form (we use our earlier notation with $Z_1, Z_2$ the common variables $Z$ as observed in sample A and B, respectively)

$$\begin{aligned} \mathrm{E}(m((X, Z_1; \theta)) &= 0 \\ \mathrm{E}(m((Y, Z_2; \theta)) &= 0 \end{aligned} \tag{97}$$

i.e. the moment restrictions are defined for the samples A and B separately. These separate moment restrictions are obtained if we consider the linear regression model (96). If we relate the $X$ to the exogenous common variables $Z$

$$X = \Pi Z + \eta \tag{98}$$

we can substitute this in (96) to obtain

$$Y = \beta_0 + \beta_1' \Pi Z + \beta_2' Z_0 + \varepsilon + \beta_1' \eta \tag{99}$$

If the order condition is satisfied, we can estimate $\beta$ from the linear regression in (99). In particular, (98) can be estimated from sample A and (99) from sample

B. The corresponding moment conditions are

$$\begin{aligned} \mathrm{E}((X_1 - \Pi Z_1)' Z_1) &= 0 \\ \mathrm{E}((Y - \beta_0 - \beta_1' \Pi Z_2 - \beta_2' Z_{02}) Z_2) &= 0 \end{aligned} \tag{100}$$

and this has the form (97). Note again that the linear first step can be seen as a linear population projection and is valid even if the conditional expectation of $X_1$ given $Z$ is not linear (provided that $Z_1$ and $Z_2$ have the same distribution). Also the moment restrictions are nonlinear in the structural parameters $\beta$. Arellano and Meghir (1992) propose to estimate $\beta_0$, $\pi = \Pi' \beta_1$ and $\beta_2$, and to use Chamberlain's (1982) minimum distance estimator in a second stage to obtain an estimate of the structural parameters. Their estimator is equivalent to the imputation estimator. In particular, it can only be used if the $X$ enters linearly in the moment conditions, and it can not be used if we estimate a model with a nonlinear (in $X$) moment condition.

Arellano and Meghir apply their estimator to a female labor supply equation. In this equation the dependent variable, hours, is observed in the UK Labor Force Survey (LFS), the European counterpart of the US Current Population Survey. Two of the independent variables, the wage rate and other income, are obtained from a budget survey, the Family Expenditure Survey (FES). This situation is common: budget surveys contain detailed information on the sources of income, while labor market surveys contain information on labor supply and

job search. An indicator whether the woman is searching for (another) job is one of the explanatory variables. Arellano and Meghir estimate the labor supply equation using the LFS data after imputing the wage rate and other income, using a relation that is estimated with the FES data. The common variables (or instruments) that are used in the imputation, but are excluded in the labor supply equation are education and age of husband and regional labor market conditions.

Angrist and Krueger (1992) consider the linear regression model

$$Y = \beta_0 + \beta_1' X + \varepsilon \tag{101}$$

with $X, Z_1$ observed in sample A and $Y, Z_2$ in sample B with A and B independent samples from a common population. They assume that all common variables are exogenous, and they implicitly assume that the number of (exogenous) common variables exceeds the number of variables in $X$, i.e. that the order condition is satisfied. Under these conditions the 2SIV estimator is based on a special case of the moment conditions in (70)-(72).

Angrist and Krueger apply the 2SIV estimator to study the effect of the age at school entry on completed years of schooling. Children usually go to school in the year in which they turn 6. If this rule were followed without exceptions, then the age at school entry would be determined by the birthdate. However, exceptions occur and there is some parental control over the age at school entry which makes this variable potentially endogenous. Angrist and Krueger assume that the the date of birth is not correlated with any characteristic of the child

and hence has no direct effect on completed years of schooling. Because there is no dataset that contains both age at school entry and completed years of schooling, Angrist and Krueger combine information in two US censuses, the 1960 and the 1980 census. Because they use 1% (1960) and 5% (1980) samples they assume that the number of children who are in both samples is negligible. They compute the age at school entry from the 1960 census and the completed years of schooling from the 1980 census. The common variable (and instrument) is the quarter in which the child is born.

Other applications of 2SIV are Carroll and Weil (1994), Lusardi (1996), Dee and Evans (2003), and Currie and Yelowitz (2000). Carroll and Weil (1994) combine data from the 1983 Survey of Consumer Finances (SCF) that contains data on savings and wealth and the Panel Study of Income Dynamics (PSID) that contains data on income growth to study the relation between the wealth income ratio and income growth. The common variables are education, occupation, and age of the head of the household. Lusardi (1996) estimates an Euler equation that relates the relative change in consumption to the predictable component of income growth. Because the consumption data in the PSID are unreliable, she uses the Consumer Expenditure Survey (CEX) to obtain the dependent variable. She also shows that the income data in the CEX are measured with error (and that number of observations with missing income is substantial) and for that reason she uses the PSID to measure income growth. She experiments with different sets of common exogenous variables that contain household characteristics (marital status, gender, ethnicity, presence of children,

number of earners), education and occupation (interacted with age), education (interacted with age). Dee and Evans (2003) study the effect of teen drinking on educational attainment. The problem they face is that there is no dataset that has both information on teen drinking and on later educational outcomes. Moreover, drinking may be an endogenous variable, because teenagers who do poorly in school may be more likely to drink. Data on teen drinking are obtained from the 1977-1992 Monitoring the Future (MTF) surveys, while data on educational outcomes are obtained from the 5% public use sample from the 1990 US census. The common exogenous variables are the minimum legal drinking age that differs between states, but more importantly increased over the observation period, state beer taxes, ethnicity, age and gender. The indicator of teen age drinking is considered to be endogenous. Currie and Yelowitz (2000) consider the effect of living in public housing on outcomes for children. The outcome variables, living in high density housing, overcrowding in the house, being held back at school, are from the 1990 census. The indicator of living in public housing is from the pooled 1990-1995 March supplements to the Current Population Survey (CPS). This indicator is assumed to be endogenous in the relation with the outcome variables. The common exogenous variable is the sex composition of the household where households with two children of different gender are more likely to live in public housing because they qualify for larger units.

## 4.5 Combining samples to correct for measurement error

One of the reasons to merge datasets is that the variables in one of the sets is measured more accurately. An example is the study by Okner (1972) who merged the 1967 Survey of Economic Opportunity with the 1966 Tax File using file matching, because the income measures reported in the SEO were thought to be inaccurate. In this section we show that even for this purpose the datafiles need not be merged, and that we can correct for measurement error in one (or more) of the explanatory variables with only marginal error free information.

The procedure that we describe works even if there are no common variables in the two datasets. If there are common variables and if these are exogenous and not correlated with the measurement error, we can use the 2SIV estimator to obtain consistent estimates of the coefficients in a linear relation where some independent variables are measured with error.

There is a larger literature on the use of validation samples to correct for measurement error. In a validation sample both $X_1$ and the true value $X_1^*$ (and $X_2$) are observed. This allows for weaker assumptions on the measurement error process. In particular, the measurement error can be correlated with $X_1^*$ and with $X_2$. This type of sample combination is beyond the scope of the present chapter. Validation samples are rare, because they require the matching of survey and administrative data. Chen, Hong, and Tamer (2003) propose a method for the use of validation samples if variables are measured with error.

We consider a simple example of a conditional distribution with pdf $f(y \mid x_1^*, x_2; \theta)$. There are two explanatory variables $X_1^*, X_2$ where $X_2$ is observed

without error and the error-free $X_1^*$ is not observed. Instead, we observe $X_1$ that is related to $X_1^*$ as specified below. The observed conditional distribution of $Y$ given $X_1, X_2$ is

$$f(y \mid x_1, x_2; \theta) = \int f(y \mid x_1^*, x_2; \theta) g(x_1^* \mid x_1, x_2) dx_1^* \tag{102}$$

if $X_1^*$ is continuous and the integral is replaced by a sum if $X_1^*$ is discrete. To determine the observed conditional distribution we need to specify or identify $g(x_1^* \mid x_1, x_2)$. We show that this conditional density can be identified from a separate dataset that only contains observations from the distribution of $X_1^*$, i.e. observations from the marginal distribution of the error-free explanatory variable. Hence we have a sample A that contains $Y, X_1, X_2$ and an independent sample B that contains only $X_1^*$.

We consider a special case that allows for a closed-form solution. In particular, we assume that both $X_1^*$ and $X_1$ are 0-1 dichotomous variables. The relation between these variables, the measurement error model, can be specified in a number of ways. We only allow for measurement error models that are identified from observations from the marginal distribution of $X_1$ observed in sample A and the marginal distribution of $X_1^*$, observed in the independent sample B. An example of such a measurement error model is classical measurement error which assumes

$$\Pr(X_1 = 1 \mid X_1^* = 1, X_2) = \Pr(X_1 = 0 \mid X_1^* = 0, X_2) = \lambda \tag{103}$$

i.e. the probability of misclassification is independent of $X_1^*$. Moreover, (103)

implies that $X_1$ is independent of $X_2$ given $X_1^*$. Solving for $\lambda$ we find

$$\lambda = \frac{\Pr(X_1 = 1) + \Pr(X_1^* = 1) - 1}{2\Pr(X_1^* = 1) - 1} \tag{104}$$

Hence, $\lambda$ is indeed identified from the marginal distributions of $X_1$ and $X_1^*$.

Note that (104) only gives solutions between 0 and 1 if

$$\Pr(X_1 = 1) < \Pr(X_1^* = 1) \tag{105}$$

if $\Pr(X_1^* = 1) > 1/2$, or if

$$\Pr(X_1 = 1) > \Pr(X_1^* = 1) \tag{106}$$

if $\Pr(X_1^* = 1) > 1/2$. This is equivalent to

$$\Pr(X_1 = 1)(1-\Pr(X_1 = 1)) = \mathrm{Var}(X_1) > \mathrm{Var}(X_1^*) = \Pr(X_1^* = 1)(1-\Pr(X_1^* = 1)) \tag{107}$$

In other words, the observed $X$ has a larger variance than the true $X_1^*$, as is generally true for classical measurement error models. This restriction on the observable marginal distributions must be satisfied, if we want to consider the classical measurement error model.

The second measurement error model assumes that misclassification only occurs if $X_1^*$ is equal to 1[17], maintaining the assumption that $X_1$ is independent of $X_2$ given $X_1^*$. Hence

[17]The misclassification can also only occur if $X_1^*$ is 0.

$$\begin{aligned} \Pr(X_1 = 0 \mid X_1^* = 0, X_2) &= 1 \\ \Pr(X_1 = 1 \mid X_1^* = 1, X_2) &= \lambda \end{aligned} \tag{108}$$

With this assumption we find

$$\lambda = \frac{\Pr(X_1 = 1)}{\Pr(X_1^* = 1)} \tag{109}$$

As in the case of classical measurement error, this measurement error model implies an observable restriction on the two observed marginal distributions, in the case $\Pr(X_1 = 1) \leq \Pr(X_1^* = 1)$.

Both measurement error models are special cases of the general misclassification error model

$$\begin{aligned} \Pr(X_1 = 0 \mid X_1^* = 0, X_2) &= \lambda_0 \\ \Pr(X_1 = 1 \mid X_1^* = 1, X_2) &= \lambda_1 \end{aligned} \tag{110}$$

Again we assume that $X_1$ is independent of $X_2$ given $X_1^*$. In this general model the parameters $\lambda_0, \lambda_1$ are not identified from the marginal distributions of $X_1$ and $X_1^*$. Hence we must fix one of these parameters or their ratio, as is done in the measurement error models that we introduced in this section. We also assume that the misclassification is independent of $X_2$.

Of course, it is not sufficient to identify the measurement error distribution. The conditional density of $Y$ given $X_1, X_2$, which is the basis for likelihood inference, is obtained from the density of $Y$ given $X_1^*, X_2$, which contains the parameters of interest, if we integrate the unobserved $X_1^*$ with respect to the density of $X_1^*$ given the observed $X_1, X_2$ ( see (102)). Hence, the key is the identification of the distribution of $X_1^*$ given $X_1, X_2$.

This conditional distribution is identified from the measurement error model that in turn is identified from the marginal distributions of $X_1$ and $X_1^*$ and the joint distribution of $X_1, X_2$. The solution depends on the measurement error model. Here we give the solution, if we assume that the measurement error is classical, but the solution for other (identified) measurement error models is analogous. In the sequel we use subscripts to indicate the variables in the distribution.

Consider

$$g_{x_1,x_1^*,x_2}(x_1, x_1^*, x_2) = g_{x_1}(x_1 \mid x_1^*, x_2) g_{x_1^*,x_2}(x_1^*, x_2) = \tag{111}$$

$$= g_{x_1}(x_1 \mid x_1^*) g_{x_1^*,x_2}(x_1^*, x_2)$$

because $X_1$ is independent of $X_2$ given $X_1^*$. After substitution of (103) we obtain

$$\begin{aligned} g_{x_1,x_1^*,x_2}(x_1, x_1^*, x_2) &= \lambda g_{x_1^*,x_2}(x_1^*, x_2), && x_1 = x_1^* \\ &= (1-\lambda) g_{x_1^*,x_2}(x_1^*, x_2) && x_1 \neq x_1^* \end{aligned} \tag{112}$$

The marginal distribution of $X_1, X_2$, which can be observed, is

$$g_{x_1,x_2}(x_1,x_2) = \lambda g_{x_1^*,x_2}(x_1,x_2) + (1-\lambda) g_{x_1^*,x_2}(1-x_1,x_2) \tag{113}$$

Solving for $g_{x_1^*,x_2}(x_1^*,x_2)$ we find

$$g_{x_1^*,x_2}(x_1^*,x_2) = \frac{(1-\lambda) g_{x_1,x_2}(1-x_1^*,x_2) - \lambda g_{x_1,x_2}(x_1^*,x_2)}{1-2\lambda} \tag{114}$$

Substitution in (112) gives the joint density of $X_1, X_1^*, X_2$. The conditional density of $X_1^*$ given $X_1, X_2$ is obtained if we divide the result by $g_{x_1,x_2}(x_1,x_2)$.

With a dichotomous $X_1$ we obtain a simple closed form solution. If $X_1$ is continuous, we can still identify the distribution of $X_1^*$ given $X_1, X_2$ if the measurement error model is identified from the marginal distributions of $X_1$ and $X_1^*$, as is the case if we assume classical measurement error. Hu and Ridder (2003) show that the identification involves two sequential deconvolution problems. They also develop the distribution theory of the resulting estimator.

# 5 Repeated cross sections

## 5.1 General principles

Repeated cross sections consist of independent samples drawn from a population at multiple points in time $t = 1, \ldots, T$. There are many such data sets in the U.S. and other countries, and more than true panel data sets in some. In the U.S., the

Current Population Survey (CPS) is a leading example, as is the General Social Survey, and even the Survey of Income and Program Participation, if data from different cohorts are employed. There are also examples of firm-level data sets of this kind. In the U.K., the Family Expenditure Survey (FES) is a prominent example. In continental Europe, CPS-like cross sections are often used, as are repeated cross sectional labor force surveys. In developing countries, such labor force surveys are often available as well as several of the World Bank LSMS surveys which have multiple waves.

Although repeated cross section (RCS) data have the obvious disadvantage relative to panel data of not following the same individuals over time, they have certain advantages over panel data. Attrition and nonresponse problems are generally much less severe, for example, and often RCS data have much larger sample sizes than available panels. In many cases RCS data are available farther back in calendar time than longitudinal data because governments began collecting repeated cross sections prior to collecting panel data. In some cases, RCS data are available for a broader and more representative sample of the population than true panel data, at least in cases where the latter only sample certain groups (e.g., certain cohorts as in the U.S. NLS panels).

Although the cross sections can be pooled and cross-sectional models can be estimated on them, the more interesting question is whether they can be used to estimate models of the type estimable with true panel data. To consider this question, assume that in each cross section $t$ we observe a sample from the distribution $W_t, Z_t$ where $W_t$ is a vector of variables that are only measured in

each cross section and $Z_t$ is a vector of variables which are measured in all cross sections, and hence can be used to match the individuals across the different waves (individual subscripts $i = 1, \ldots, N$ are omitted for now). Both $W_t$ and $Z_t$ may contain variables which are identical at all $t$ (i.e., time invariant variables) although in most applications all time invariant variables will be measured at all $t$ and hence will be in $Z_t$. We assume that the population is sufficiently large and the sample sufficiently small that there are no common individuals in the cross sections. Further, we assume that the population from which the samples are drawn is closed[18], and thus we ignore out in- and out-migration, births, and mortality.

At issue is what distributions and what types of models can be identified from the set of cross sections. The unconditional joint distribution of $W_1, \ldots, W_T$ is not identified except in the trivial case in which the elements are independent. Models which require for identification only moments from each cross-section, and which therefore do not require knowledge of the joint distribution, are identified but do not make particular use of the repeated cross section (RCS) nature of the data except perhaps for investigations of time-varying parameters. The models of interest and under discussion here are those which require identification of the joint distribution or of some aspect of it.

Identification necessarily requires restrictions. Nonparametric identification of conditional distributions $f(W_t|W_\tau), t \neq \tau$ follows from the general principles

[18]This ensures that the relation between a dependent and independent variables does not change over time due to in- and outflow from the population, and we can make this assumption, instead of that of a closed population.

and restrictions elucidated in section 4.2.2 above, with the change of notation from $Y$ to $W_t$ and from $X$ to $W_\tau$. With the common variable $Z_t$ available in each cross section and used for matching, bounds on those conditional distributions can be established. If $Z_t$ or some elements of it are excluded from the relation between $W_t$ and $W_\tau$, and $Z_t$ is discrete, the conditional distributions are exactly identified provided a rank condition is met which relates the number of points in the support of $Z_t$ to the number of conditional distributions to be estimated.

We shall focus in this section primarily on parametric models for which independence of $W_1, \ldots, W_T$ is not assumed but which contain exclusion restrictions. While there are in general many models which can be identified under different restrictions, we will work with a model similar to that in section 4.3.2 above:

$$f(Y_t; \theta) = g_1(X_t, Z_0; \theta) + g_2(Y_{t-1}, Z_0; \theta) + \varepsilon_t \tag{115}$$

and with associated GMM moment condition, following on (69), of:

$$E[(f(Y_t; \theta) - g_1(X_t, Z_0; \theta) - g_2(Y_{t-1}, Z_0; \theta))h(Z_0, Z_{1t})] = 0 \tag{116}$$

where $f, g_1, g_2$, and $h$ are known (possibly up to parameters) functions and $\theta$ a vector of parameters. The vector $Z_0$ is a vector of common time-invariant variables in the cross sections which are included in the $g_1$ and $g_2$ relations[19]. In most applications, $f(Y_t; \theta) = Y_t$. The function $g_1$ contains only $X_t$ and $Z_0$ and hence appears to be estimable from a single cross-section, but, as will be

[19]These variables can be time-varying but this is rare in applications so we consider only the case where they are time-constant. None of the results we discuss below are substantially changed by this restriction.

shown below, this is problematic in fixed effects models because $X_t$ is correlated with the error in that case. The functions $g_1$ and $g_2$ must be separable because $X_t$ and $Y_{t-1}$ do not appear in the same cross-section.

Individuals across cross sections are identified by variables $Z_0$ and $Z_{1t}$, with the latter excluded from the relation between $Y_t$ and $X_t, Y_{t-1}, Z_0$. In most applications to date, $Z_{1t} = t$ or a set of time dummies [20]. The critical exclusion restriction in all RCS models is that $Z_{1t}$ and its interactions with $Z_0$ do not enter in $g_1$ and $g_2$, and yet these variables are correlated with those functions. For the $Z_{1t} = t$ case, this implies that variables that change predictably with time, as individual age, year, unemployment duration, or firm lifetimes (depending on the application) cannot enter $g_1$ and $g_2$. Thus the essential restriction in RCS estimation is that intertemporal stability exist in the true relationship. Such a restriction is not needed when true panel data are available. Note as well that the number of independent components in $h$ must not be smaller than the dimension of $\theta$ and, in most models, must be larger than the dimension of $X_t$, $Y_{t-1}$, and $Z_0$. This also can be a fairly limiting condition in practice if the number of cross-sections available is small relative to the number of parameters whose identification requires instrumenting with functions of $t$.

In linear models the GMM estimator can be implemented as a two-step estimator. First, project $X_t$ and $Y_{t-1}$ on $h(Z_0, Z_{1t})$, i.e obtain $E(X_t|h(Z_0, Z_{1t}))$

[20] However, it is possible that some history information is available in each cross-section which means that these time-varying variables (e.g., employment or marital status histories in the case of household survey data; ages of children are another) are potential additional instruments.

and $E(Y_{t-1}|h(Z_0, Z_{1t}))$[21]. Second, regress $Y_t$ on these projections and on $Z_0$. If there are no $Z_0$ in the data and $h(Z_{1t})$ is a set of time dummies, this is equivalent to an aggregate time-series regression where the time means of $Y_t$ are regressed upon the time means of $X_t$ and $Y_{t-1}$. In this case the number of cross-sections has to be at least 3. Most interesting cases arise however when $Z_0$ variables are available; in household survey data, these may be birth year (=cohort), education, race, sex, and so on. If these variables are all discrete and $h(Z_{1t}, Z_0)$ is assumed to be a vector of indicators for a complete cross-classification $Z_0$ and time, estimation using (116) is equivalent to a regression of the cell means of $Y_t$ on the cell means of $X_t$, $Y_{t-1}$, and the dummy variables $Z_0$. Note that in that case we need fewer cross-sections. However, if a parametric form of $h$ is assumed, this aggregation approach is not necessary, and if the model is nonlinear (including the binary choice and related models), the two-step aggregation approach is not possible in the first place. In that case the estimator is the possibly overidentified GMM estimator defined by the moment conditions in (116).

Two leading examples fit into this framework. One is the linear first-order autoregression (with individual $i$ subscripts now added)

$$Y_{it} = \alpha + \beta Y_{i,t-1} + \gamma X_{it} + \delta Z_{0i} + \varepsilon_{it} \tag{117}$$

With time dummies as excluded variables the number of observations is equal to the number of cross-sections and this imposes restrictions on the time-variation of the parameters of (117). The restriction that the instrument must be relevant

[21] Projections onto $Z_0$ and $Z_{1t}$ directly are an alternative.

implies that the mean of $E(Y_{t-1}|Z_0, t)$ must vary with $t$. If $Y_{t-1}$ is correlated with $\varepsilon_t$, an instrument $Z_{1t}$ must be found which is orthogonal to $\varepsilon_t$.

A second example is the linear individual effects model

$$Y_{it} = \gamma X_{it} + \delta Z_{0i} + f_i + \varepsilon_{it} \tag{118}$$

where $f$ is an individual effect which is potentially correlated with $X_t$ and $Z_0$. The within-estimator commonly used with true panel data cannot be implemented with RCS data because it requires knowledge of $Y_t$ at multiple $t$. RCS IV estimation using (116) proceeds by using the elements of $h$ as instruments for $X_t$, which again requires some minimal time-invariance of the parameters of (118). Consistency (see below) is based on the presumption that time-varying variables like those in $Z_{1t}$ must be orthogonal to time-invariant variables like $f$. For instrument relevance, $E(X_t|Z_0, t)$ must vary with $t$.

Estimation of the model in (118) by the aggregation method mentioned previously was proposed by Deaton (1985). He considers cohort data, so that time in his case is age. Deaton considered $Z_0$ to contain only birth year (=cohort) indicators and $h$ to be a set of all cohort-age indicators. He then proposed constructing a data set of cohort profiles of mean $Y$ and $X$ (a 'pseudo' panel data set) and estimating (118) by regressing the age-cohort means of $Y$ on those of $X$ and on cohort dummies (or by the within-estimator for fixed effects models applied to these aggregate observations).

## 5.2 Consistency and Related Issues

The conditions for consistency of moment estimators in the form (116) are well-known in general (Hansen, 1982). The special form they take in the two sample case were considered in section 4.3.2 above, where weak consistency was proven. For the RCS case, aside from the usual rank conditions and conditions on convergence of matrices to positive definite forms, we have the condition that the instruments are not correlated with the random error

$$\mathrm{E}\left[\eta_{it} h(Z_{oi}, Z_{it})\right] = 0$$

where $\eta_{it} = f(Y_{it};\theta) - g_1(X_{it}, Z_{0i};\theta) - g_2(Y_{i,t-1}, Z_{0i};\theta)$. If there is an individual effect, we have that $\eta_{it} = f_i + \varepsilon_{it}$ and hence we require that $\mathrm{E}\left[\varepsilon_{it} h(Z_{oi}, Z_{it})\right] = 0$, and $\mathrm{E}\left[f_i h(Z_{oi}, Z_{it})\right] = 0$. The first assumption must hold even with the presence of $Y_{t-1}$ in the equation and represents an IV solution familiar to panel data models with dynamics and lagged endogenous variables. However, with a lagged dependent variable in the equation the errors in successive periods have a MA(1) structure because the errors in not observing the same individuals in each cross section are correlated (McKenzie, 2004).

The assumption on the individual effect $f_i$ that may be correlated with $X_{it}$ is the more problematic assumption. If $h$ is a set of time dummies, then a sufficient condition is that the (population) mean of $f_i$ does not change over time. If we have repeated cross sections of size $N_t$ in period $t = 1, \ldots, T$, then

this implies that[22]

$$\overline{f}_t = \frac{1}{N_t}\sum_{i=1}^{N_t} f_i \xrightarrow{p} 0$$

Hence, if $\min\{N_1, \ldots, N_T\} \to \infty$, then the limit of the time averaged regression without a lagged dependent variable is

$$Y_t^* = \alpha + \gamma X_t^* + \varepsilon_t^*$$

with $\varepsilon_t^*$ a common time shock in the $\varepsilon_{it}$ and * indicating population averages of the variables. OLS applied to this equation gives consistent estimators of the regression parameters, and this establishes that the GMM estimator that uses moment condition (116) is consistent if $\min\{N_1, \ldots, N_T\} \to \infty$, i.e. for large $N$ asymptotics.

For the same model and assumptions on the random error, time dummies are not valid instruments if $N_t$ is fixed and $T$ becomes large. Note that in this case the number of instruments is equal to $T$ and hence goes to infinity. The problem is obvious if we consider the second stage regression that involves the projections on the instruments, i.e. the averages in the repeated cross sections

$$\overline{Y}_t = \alpha + \gamma \overline{X}_t + \overline{f}_t + \overline{\varepsilon}_t$$

Hence

$$\mathrm{E}\left[\overline{X}_t \overline{f}_t\right] = \frac{1}{N_t}\mathrm{E}\left[X_{it} f_i\right] \neq 0$$

for finite $N_t$.

There is another asymptotic that can be considered as well, which is an asymptotic in the number of cohorts (Deaton, 1985; Verbeek, 1993). Up to this

[22] Without loss of generality we can take the common time constant limit equal to 0.

point we have assumed that a single population of $N$ individuals is followed over time for $T$ periods, which is equivalent to a single cohort (or a fixed set of birth years). Now let us consider increasing the number of such cohort groups ($c$) by moving over calendar time, or possibly space, and increasing the number of pseudo-panels in the data. Each new panel has $N$ individuals and is observed for $T$ periods. Once again, with fixed $N$, the average individual effect will be correlated with the average covariate, so that the GMM estimator is biased.

Deaton (1985) has proposed a modification of the estimator for the linear fixed effects model which contains a bias adjustment for the finite, fixed $N$ case and which is consistent for the large $T$ case, an estimator that has been much discussed in the literature. Deaton notes that estimation of the aggregated estimation equation

$$\overline{Y}_{ct} = \gamma \overline{X}_{ct} + \delta_c + \overline{\varepsilon}_{ct} \tag{119}$$

where means are taking over observations within each cohort ($c$) and year ($t$) cell yields biased estimates for finite $N$ because $\overline{f}_{ct}$ is correlated with $\overline{X}_{ct}$. Deaton instead considers the "population" equation

$$Y^*_{ct} = \gamma X^*_{ct} + \delta_c + \varepsilon^*_{ct} \tag{120}$$

where variables with asterisks represent population values, i.e. values that would obtain if the cohort would be infinitely large. Note that $\delta_c$ absorbs a non-zero mean of the $f_i$ in cohort $c$.

For the estimation of (120) $\overline{X}_{ct}$ and $\overline{Y}_{ct}$ must be inserted to proxy their population counterparts but they do so with error. Deaton suggests that the

measurement error for each be estimated by the within-cell variances of $X$ and $Y$ using the individual data and that a finite-sample adjustment be made when estimating the coefficient vector.

Deaton does not set up his model in the GMM framework but it can be done so. Although he discusses his estimator as an errors-in-variables estimator, it is more in line with our discussion to consider it as a finite $N$ bias-corrected version of the GMM estimator. To focus on the key issues, assume that only one cohort of $N$ individuals is observed for $T$ periods. The individual model is

$$y_{it} = \delta + \beta x_{it} + f_{it} + \varepsilon_{it} \tag{121}$$

The second stage equation when using time dummies as instruments is

$$\overline{y}_t = \beta \overline{x}_t + \overline{f}_t + \overline{\varepsilon}_t \tag{122}$$

Consequently,

$$\text{Cov}(\overline{y}_t, \overline{x}_t) = \beta \text{Var}(\overline{x}_t) + \text{Cov}(\overline{f}_t, \overline{x}_t) \tag{123}$$

The bias term in (123) is

$$\text{Cov}(\overline{f}_t, \overline{x}_t) = \frac{\text{Cov}(f_i, x_{it})}{N} \tag{124}$$

This bias term is small if $N$ is large or if the correlation between the regressor and the individual effect is small. The Deaton finite sample adjustment can be derived by noting that $f_{it} = y_{it} - \beta x_{it} - \varepsilon_{it}$ and that, therefore, $\text{Cov}(f_{it}, x_{it}) = \text{Cov}(y_{it}, x_{it}) - \beta \text{Var}(x_{it})$. Hence $\text{Cov}(\overline{f}_t, \overline{x}_t) = \frac{\sigma_{yx} - \beta \sigma_x^2}{N}$ where $\sigma_{yx}$ and $\sigma_x^2$ are the covariance of $x$ and $y$ and the variance of $x$ for the individual observations in a time period. Inserting this into (123) and solving for $\beta$, we obtain the Deaton

estimator if we replace the population variances and covariances with sample variances and covariances.

$$\hat{\beta} = \frac{\text{Cov}(\overline{y}_t, \overline{x}_t) - \frac{\sigma_{yx}}{N}}{\text{Var}(\overline{x}_t) - \frac{\sigma_x^2}{N}} \tag{125}$$

As $N \to \infty$ the bias and the bias correction terms go to 0 and the least squares estimate of the aggregate model is consistent. Deaton noted that the estimator is also consistent as $T \to \infty$ and Verbeek and Nijman (1992, 1993) show that this estimator is consistent as $C \to \infty$ provided a minor change is made in the bias correction. Verbeek and Nijman also note that the Deaton estimator increases variance at the same time that it reduces bias, giving rise to a mean-squared error tradeoff that can be addressed by not subtracting off the "full" bias correction in (125). Devereux (2003) shows that the Deaton estimator is closely related to estimators which adjust for finite sample bias in IV estimation and that, in fact, the estimator is equivalent to the Jacknife Instrumental Variables estimator and is closely related to k-class estimators. Devereux also proposes a modification of the Deaton estimator which is approximately unbiased but has a smaller finite sample variance.

There have been some explorations in the literature seeking to determine how large $N$ must be for the finite sample adjustments to be avoided by Monte Carlo simulations. Verbeek and Nijman (1992) suggest that cell sizes of 100 to 200 are sufficient, while Devereux (2003) suggests that should be higher, possibly 2000 or more. The necessary $N$ is sensitive to the specification of the model. Devereux also conducts an exercise which subsamples the available $N$ in a model to gauge the degree of bias.

There has also been a discussion in the literature of how to divide the available data into cohort groups, given that most data sets do not have sufficient samples to divide the data completely by discrete values of birth year (Verbeek and Nijman, 1992, 1993). Dividing the sample into more birth cohorts increases $C$ while decreasing the sample size per cohort. In the applied literature, groupings of birth cohorts and formation of cells for the aggregated estimation has been, by and large, ad hoc. Moffitt (1993) suggests that aggregation not be conducted at all but rather that the individual data be employed and a parametric function of birth year and $t$ be estimated to smooth the instrument to achieve efficiency, but he does not present any formal criteria for how to do so. A better framework for analyzing these issues is that which considers alternative specifications of the instrument which trade off bias and variance. Donald and Newey (2001) present one such analysis.

The literature has also addressed dynamic fixed effects models. In this case we are interested in the individual model

$$Y_{it} = \alpha + \beta Y_{i,t-1} + \delta Z_{0i} + f_i + \varepsilon_{it} \tag{126}$$

which is a combination of (117) and (118). The desirability of different instrument sets $Z_{1i}$ depends once again on the asymptotics involved. But when asymptotics are taken in $N$ (the number of observations per cohort), the consistency properties of different instrument sets are almost identical to those for true panel data (Sevestre and Trognon, 1996; Arellano and Honoré, 2001). Using simple functions of $t$ as instruments, for example, will yield inconsistent estimates for the same reasons that conventional fixed effects methods in true

panel data yield inconsistent estimates in the presence of both fixed effects and lagged regressors. As in the case of true panel data, additional instruments which generate first-differenced estimators and which use lagged values of the dependent variable can yield consistent estimates.

Collado (1997) and McKenzie (2004) consider this model and discuss various applications of IV to the model, using the same principles in the literature on true panel data, using lagged values of the dependent variable as instruments and possibly using the larger instrument set implied by the Arellano-Bond estimator. Collado and McKenzie also propose Deaton-style bias-correction terms to correct for the finite $N$ problem discussed above. Collado shows that her estimator is consistent in $C$ and, for a different bias-correction, consistent in $T$. McKenzie considers a sequential asymptotic in which N is first allowed to go to infinity conditional on fixed T and then limits are taken w.r.t. $T$.

## 5.3 Binary Choice Models

In the binary choice model we return to (115) and let $f(Y_t;\theta) = Y_t^*$, $Y_t = I(Y_t^* \geq 0)$, and $F$ be the c.d.f. of $-\varepsilon_t$. Then $\Pr(Y_t = 1|X_t, Z_0, Y_{t-1};\theta) = F(g_1(X_t, Z_0;\theta) + g_2(Y_{t-1}, Z_0;\theta))$ so that

$$Y_{it} = F(g_1(X_{it}, Z_{0i};\theta) + g_2(Y_{i,t-1}, Z_{0i};\theta)) + \nu_{it} \tag{127}$$

which does not fit into the framework of the moment condition in (117) because $X_t$ and $Y_{t-1}$ are not separable. Let us therefore initially assume $g_2 = 0$ and consider lagged indicators below. Now (117) applies directly assuming the availability of a suitable exclusion restriction, as before. The moment conditions are

a simple extension of those shown in equations (74)-(76). The method is applicable to the individual effects binary choice model or to a binary choice model with endogenous $X_t$ with the restrictions that hold in the cross-section case. For instance, in parametric estimation where the $F$ distribution is assumed to be known, a distributional assumption is needed for the individual effect in order to derive $F$, e.g., if $f$ is the individual effect component of $\varepsilon_t$,

$$f_i = v(Z_{0i}; \phi) + \eta_i \tag{128}$$

where $v$ is assumed to be of known form and where $\eta_i$ has a known parametric distribution from which the c.d.f. of the composite error $\eta_i - \varepsilon_{it}$ can be derived.

If $X_t$ is endogenous and if the instrument is a set of time dummies, possibly interacted with $Z_0$, the nonlinearity of the conditional expectation function means that GMM is not equivalent to any type of aggregate regression of cell means of $Y$ on cell means of $X$ and $Z$. However, with a stronger assumption, a version of such an approach is possible (Moffitt, 1993). The necessary assumption, in addition to (128), is

$$X_{it} = w(Z_{0i}, Z_{1it}; \psi) + \omega_{it} \tag{129}$$

where $w$ is a function of known parametric form and $\omega_{it}$ is an error term with a parametric distributional form that may be correlated with $\varepsilon_{it}$. The assumption that the exact form of dependence of the endogenous variable on the instruments is known and that the conditional distribution of the regressor follows a specific parametric form are very strong. In the simplest case, $g_1$ is linear in $X_t$ and $Z_0$ and $w$ is linear in $Z_0$ and $Z_{1t}$, and $\varepsilon_t$ and $\omega_t$ are assumed to be bivariate normal.

Then a variety of estimating techniques are possible, drawing on the literature on endogenous regressors in limited dependent variable models (Amemiya, 1978; Heckman, 1978; Nelson and Olsen, 1978; Rivers and Vuong, 1988; Smith and Blundell, 1986; see Blundell and Smith, 1993 for a review). Options include replacing $X_t$ in $g_1$ with its predicted value from (129); inserting an estimated residual from (129) into (127); and estimating (153) and (155) in reduced from by inserting (129) into (127). In this approach, the parameters of (127) are estimated by maximum likelihood, which implies that the instrument vector $h$ in (116) is the binary choice instrument vector that is equal to $\frac{F'}{(1-F)F}$ times the derivative of the argument of $F$ w.r.t. $\theta$.

To consider the model with $Y_{t-1}$ let us first consider the case in which $X_t = X$ is time invariant, in which case it can be folded into $Z_0$ and we can let $g_1 = 0$ without loss of generality. Then we have

$$\mathrm{E}(Y_{it}|Z_{0i}, Y_{i,t-1}) = F(g_2(Y_{i,t-1}, Z_{0i}; \theta)) \tag{130}$$

where we have assumed that $\varepsilon_{it}$ is distributed independently of $Y_{i,t-1}$ i.e., there is no serial correlation. Instrumental variable estimation of (130) conducted by replacing $Y_{t-1}$ by a predicted value and applying maximum likelihood to the resulting model is known to be inconsistent because $Y_{t-1}$ is binary and hence its prediction error follows a non-normal, two-point discrete distribution. An alternative procedure is to integrate $Y_{t-1}$ out of the equation. Letting $p_t(Z_0)$ be the marginal probability $\Pr(Y_t = 1|Z_0)$, we have

$$E(Y_t|Z_0) = p_t(Z_0) = \tag{131}$$

$$= p_{t-1}(Z_0)\Pr(Y_t = 1|Z_0, Y_{t-1} = 1) + (1 - p_{t-1}(Z_0))\Pr(Y_t = 1|Z_0, Y_{t-1} = 0) =$$

$$= p_{t-1}(Z_0)F(g_2(1, Z_0; \theta) + (1 - p_{t-1}(Z_0))F(g_2(0, Z_0; \theta) =$$

$$= p_{t-1}(Z_0)(1 - \lambda(Z_0; \theta)) + (1 - p_{t-1}(Z_0))\mu(Z_0; \theta) =$$

$$= \mu(Z_0; \theta) + \eta(Z_0; \theta)p_{t-1}(Z_0)$$

where $\lambda(Z_0; \theta) = \Pr(Y_t = 0|Z_0, Y_{t-1} = 1) = F(g_2(1, Z_0; \theta)$ is the exit rate from $Y_{t-1} = 1$ to $Y_t = 0$, $\mu(Z_0; \theta) = \Pr(Y_t = 1|Z_0, Y_{t-1} = 0) = F(g_2(0, Z_0; \theta)$ is the exit rate from $Y_{t-1} = 0$ to $Y_t = 1$, and $\eta(Z_0; \theta) = 1 - \lambda(Z_0; \theta)\mu(Z_0; \theta)$. Equation (131) is a familiar flow identity from renewal theory showing how the marginal probability at $t - 1$ is transformed by the two transition probabilities into the marginal probability at $t$. It suggests a procedure by which the reduced form model $Y_t = \mu(Z_0; \theta) + \eta(Z_0; \theta)p_{t-1}(Z_0) + \nu_t$ is estimated by nonlinear least squares (given the nonlinearity of the two transition probabilities in $\theta$) or GMM using a first-stage estimate of $p_{t-1}(Z_0)$ similar to the case of a generated regressor. Because the marginals at every $t$ are estimable from the RCS data, such a first-stage estimate is obtainable. Identification of the transition probabilities is achieved by restricting their temporal dependence (indeed, in (131) they are assumed to be time invariant); identification is lost if the two transition probabilities vary arbitrarily with $t$ (Moffitt, 1993). The model is equivalent to a two-way contingency table where the marginals are known; the data furnish a sample of tables and the restrictions on how the joint distribution varies across the sample yields identification.

The first-stage estimation of $p_{t-1}(Z_0)$ can be obtained from an approximation of the function or the structure of the model can be used to recursively

solve for $p_{t-1}(Z_0)$ back to the start of the process. Assuming that $p_0 = 0$ and that the process begins with $t = 1$, and continuing to assume time-invariant hazards,

$$p_{t-1}(Z_0) = \mu(Z_0;\theta)\left[1 + \sum_{\tau=1}^{t-2} \eta(Z_0;\theta)^{t-1-\tau}\right] = \tag{132}$$
$$= \mu(Z_0;\theta)\frac{1-\eta(Z_0;\theta)^{t-1}}{1-\eta(Z_0;\theta)}$$

which can be jointly estimated with (131) imposing the commonality of the functions[23]. Alternatively, (131) can be expressed in fully solved back form and estimated as well.

Equation (131) has been used as the basis of RCS estimation at the aggregate level. Miller (1952) considered estimation of (131) with time-series data on the proportions of a variable, $p_t$ which is special case of RCS data. Without data on individual regressors $Z_0$, he suggested simple least squares estimation of

$$p_t = \mu + \eta p_{t-1} + \nu_t \tag{133}$$

Madansky (1959) proved that the least squares estimators of the two hazards are consistent for fixed $N$ as $T \to \infty$ and for fixed $T$ as $N \to \infty$. Lee, Judge, and Zellner (1970) and MacRae (1977) proposed various types of restricted least squares estimators to ensure that the estimated hazards do not fall outside the unit interval. This problem would not arise in the approach here, which specifies the hazards in proper probability form.

Estimation of the Markov model with RCS data is considerably complicated if there is serial correlation in the errors or if time-varying $X_t$ are allowed. With

[23] Alternatively an initial conditions can be specified as a marginal $p$ in the first period.

serial correlation of the errors, the two transition probabilities require knowledge of the functional dependence of $\varepsilon_t$ on $Y_{t-1}$. The most straightforward approach would require replacing the simple transition probabilities we have shown here with joint probabilities of the entire sequences of states $Y_{t-1}, Y_{t-2}, \ldots, Y_1$ which in turn would be a nonlinear function of $Z_0$ and the parameters of the assumed joint distribution of $\varepsilon_{t-1}, \varepsilon_{t-2}, \ldots, \varepsilon_1$. This treatment would be parallel to maximum likelihood estimation with true panel data in random effects and similar models where the joint distribution is likewise integrated out. With time-varying $X_t$, the approach in (131) is problematic because

$$E(Y_t|X_t, Z_0) = \mu(X_t, Z_0; \theta) + \eta(X_t, Z_0; \theta)p_{t-1}(X_t, Z_0) \tag{134}$$

where $\mu(X_t, Z_0; \theta) = F(g_1(X_t, Z_0; \theta) + g_2(0, Z_0; \theta))$ and $\lambda(X_t, Z_0; \theta) = 1 - F(g_1(X_t, Z_0; \theta) + g_2(1, Z_0; \theta))$. The difficulty is that $p_{t-1}(X_t, Z_0)$ is not identified from the data. Estimation would require the assumption of a Markov or other process for $X_t$ which could be used to formulate a function $p_{t-1}(X_t, Z_0)$ which could be identified from the data.

## 5.4 Applications

Despite the large number of RCS data sets in the U.S. and abroad, the methods described in this section have been applied relatively infrequently. The vast majority of uses of RCS data simply estimate pooled cross-sectional parameters without matching individuals across waves by birth cohort, education, or other individual time-invariant covariates. A rather large literature on program evaluation in the U.S. uses RCS data with area fixed effects in a period where

policies differ across areas and over time and policy effects are estimated from the cross-area covariation in the change in policies and in the outcome (migration is ignored). This literature likewise does not make use of the techniques discussed here.

Of the applications that have been conducted, virtually all have used the Deaton linear fixed effects aggregation approach rather than the more general GMM-IV approach described here. Most of the applications have been to life cycle models, which is a natural area of application because age profiles are central to the theory and the Deaton approach is explicit in formulating aggregate cohort profiles of that type. Browning, Deaton, and Irish (1985) estimated a life cycle model of labor supply and consumption using seven waves of the FES and were the first to demonstrate the estimation of the fixed effects model, which arises naturally from the first order conditions of separable lifetime utility functions, by aggregation into cohort profiles. Subsequent FES analyses include Blundell, Browning, and Meghir (1994), who estimated Euler equations under uncertainty for aggregate cohort profiles of consumption, using instrumental variables with lags to control for the endogeneity of lagged consumption; Attanasio and Weber (1994), who estimated life cycle consumption profiles with aggregate cohort means but allowed calendar-time varying effects in an attempt to explain macro trends in UK consumption; and Alessie et al. (1997), who added borrowing constraints to the model. Analyses using RCS methods to other data sets are small in number. Attanasio (1998) used the U.S. Consumer Expenditure Survey to construct aggregate cohort profiles of

saving rates in an attempt to explain the decline in saving rates in the U.S. Blow and Preston (2002) used a UK tax data set that did not contain information on age to estimate the effect of taxes on earnings of the self-employed, and followed the aggregation approach grouping on region of residence and occupation. Paxson and Waldfogel (2002) used the Deaton method but applied to state-specific means over time in the U.S., regressing state-specific measures of measures of child mistreatment on a number of state-level variables and mean socioeconomic characteristics obtained from the CPS as well as state and year fixed effects. The authors applied the Deaton finite-sample correction to the regressor matrix containing the moments for the aggregate CPS regressors and reported large increases in estimated coefficients as a result. Finally, Heckman and Robb (1985) showed that treatment effects models can be estimated with RCS data even if information on who has been trained and who has not is not available in post-training cross-sections if the fraction who are trained is known.

There have been a few applications of the Markov model described above. Pelzer, Eisinga and Franses (2002, 2004) have implemented the maximum likelihood estimator suggested in Moffitt (1993) and discussed above, adding unobserved heterogeneity, for two applications. The papers also discuss alternative computational methods and algorithms. In the first application, the authors used a true panel data set with five waves to estimate a Markov model for changes in voter intentions (Democrat vs Republican), treating the panel as a set of repeated cross sections. They then validated the model by estimating model on the true panel, and found that the coefficients on the regressor variables were

quite similar in both methods but that the intercept was quite different. In the second application, the authors examined transition rates in personal computer ownership in the Netherlands over a 16-year period, but again using a true panel data set which was initially treated as a set of repeated cross sections. The authors again found the regressor coefficients to be quite close in both cases. The authors also note that the RCS Markov model is formally identical to problem of ecological inference, or the problem of how to infer individual relationships from grouped data (Goodman, 1953; King, 1997). In the ecological inference problem, a set of grouped observations furnishes data on the marginals of binary dependent and independent variables (the "aggregate" data) and restrictions on how the joint distribution (the "individual data") varies across groups is used for identification.

Güell and Hu (2003) studied the estimation of hazard functions for leaving unemployment using RCS data containing information on the duration of the spell, allowing matching across cross-sections on that variable. The authors used a GMM procedure very similar to that proposed here. The similarity to the RCS Markov model discussed here is superficial, however, for the matching on duration permits direct identification of transition rates. The authors apply the method to quarterly Spanish labor force survey data, which recorded spell durations, over a 16 year period, and estimate how exit rates from unemployment have changed with calendar time and what that implies for the distribution of unemployment. A simpler but similar exercise by Peracchi and Welch (1994) used matched CPS files in adjacent years over the period 1968-1990 to measure

labor force transitions between full-time, part-time, and no work, and then assemble the transition rates into an RCS data set which they use to estimate transition rates by cohort as a function of age, year, and other variables.

# References

Alessie, R., M. Devereux, and G. Weber (1997), "Intertemporal consumption, durables and liquidity constraints: A cohort analyis." *European Economic Review* 41: 37-59.

Alter, H. A.(1974), "Creation of a synthetic data set linking records of the Canadian survey of consumer finances with the family expenditure survey 1970." *Annals of Economic and Social Measurement* 3: 395-97.

Amemiya, T. (1978), "The estimation of a simultaneous-equation generalized probit model." *Econometrica* 46, no. 5: 1193-1205.

Amemiya, T. (1985), *Advanced econometrics.* Cambridge, Massachusetts: Harvard University Press.

Angrist, J. D., and A. B. Krueger (1992), "The effect of age at school entry on educational attainment: An application of instrumental variables with moments from two samples." *Journal of the American Statistical Association* 87: 328-36.

Arellano, M., and B. Honoré (2001), "Panel data models: Some recent developments." In *Handbook of Econometrics*, edited by J. Heckman and E. Leamer. Vol. 5. Amsterdam and New York: Elsevier.

Arellano, M., and C. Meghir (1992), "Female labour supply and on-the-job-search: An empirical model estimated using complementary data sets." *Review of Economic Studies* 59: 537-57.

Attanasio, O. (1998), "A cohort analysis of saving behavior by U.S. households." *Journal of Human Resources* 33: 575-609.

Attanasio, O., and G. Weber (1994), "The UK consumption boom of the late 1980's: Aggregate implications of microeconometric evidence." *Economic Journal* 104: 1269-1302.

Attansio, O., and G. Weber (1994), "The UK consumption boom of the Late 1980s: Aggregate implications of microeconometric evidence." *Economic Journal* 104: 1269-1302.

Barr, R. S., and J. S. Turner (1978), "A new, linear programming approach to microdata file merging." In *Compendium of Tax Research*. 1978 ed., 131- 55. Washington, DC: Office of Tax Analysis, Department of the Treasury.

Belin, T. R., and D. B. Rubin (1995), "A method for calibrating false-match rates in record linkage." *Journal of the American Statistical Association* 90: 694-707.

Blow, L., and I. Preston (2002), Deadweight loss and taxation of earned income: Evidence from tax records of the UK self-employed. London: IFS Working Paper 15.

Blundell, R., M. Browning, and C. Meghir (1994), "Consumer demand and the life- cycle allocation of household expenditures." *Review of Economic Studies* 61: 57-80.

Blundell, R., and R. Smith (1993), "Simultaneous microeconometric models with censored and qualitative dependent variables." In *Handbook of Statistics: Econometrics*, edited by G. S. Maddala, C. R. Rao, and H. D. Vinod. Vol. 11. Amsterdam and New York: Elsevier.

Box, G. E. P., and D. R. Cox (1964), "An analysis of transformations." *Journal of the Royal Statistical Society, B* 26: 211-52.

Browning, M., A. Deaton, and M. Irish (1985), "A profitable approach to labor supply and commodity demands over the life cycle." *Econometrica* 53: 503-44.

Buehler, J. W.,K. Prager, C.J. Hogue, K. Shamsuddin, E. Lieberman, T.K. Young, E. Kliewer, J. Blanchard, and T. Mayer (2000), "The role of linked birth and and infant death certificates in maternal and child health epidemi-

ology in the United States." *American Journal of Preventive Medicine* 19: 3-11.

Burbidge, J. B., L. Magee, and A. L. Robb (1988), "Alternative transformations to handle extreme values of the dependent variable." *Journal of the American Statistical Association* 83: 123-27.

Card, D., A. K. G. Hildreth, and L. D. Shore-Sheppard (2001), The measurement of Medicaid coverage in the SIPP: Evidence from California, 1990-1996. Working paper, NBER 8514.

Carroll, C. D., K. E. Dynan, and S. D. Krane (1999), Unemployment risk and precautionary wealth: Evidence from household's balance sheetd. Finance and economics dicussion series, 1999-15, Federal Reserve Board, Washington, DC.

Carroll, C. D., and D. N. Weil (1994), "Saving and growth: A reinterpretation." *Carnegie-Rochester Conference Series on Public Policy* 40: 133-191.

Chamberlain, G. (1982), "Multivariate regression models for panel data." *Journal of Econometrics* 18: 5-46.

Chen, X., H. Hong, and E. Tamer (2003), "Measurement Error Models with Auxiliary Data." forthcoming *Review of Economic Studies*.

Chen, X., Hong, H., and Tarozzi, A. (2004), Semiparametric efficiency in GMM models of nonclassical measurement error, missing data and treatment effects.

Cohen, M. J. (1991) "Statistical matching and microsimulation models." In *Improving Information for Social Policy Decisions; The Uses of Microsimulation Modeling*, edited by C. F. Citro and E. A. Hanushek, 62-85. Vol. II, Technical papers. Washington, DC: National Academy Press.

Collado, L. (1997), "Estimating dynamic models from time series of independent cross-sections." *Journal of the Econometrics* 82: 37-62.

Copas, J. B., and F. J. Hilton (1990), "Record linkage: Statistical methods for matching computer records." *Journal of the Royal Statistical Society, A* 153: 287-320.

Cramer, J. S., and A. H. Paape (1990), Synthetische koppeling van microdata (Synthetic linkage of microdata). Report SEO, Amsterdam, The Netherlands.

Cross, P. J., and C. F. Manski (2002), "Regressions, short and long." *Econometrica* 70: 357-68.

Currie, J., and A. Yelowitz (2000), "Are public housing projects good for kids?" *Journal of Public Economics* 75: 99-124.

Deaton, A. (1985), "Panel data from time series of cross sections." *Journal of Econometrics* 30: 109-26.

Dee, T. S., and W. N. Evans (2003), "Teen drinking and educational attainment: Evidence from Two-Sample Instrumental Variables (TSIV) estimates." *Journal of Labor Economics* 21: 178-209.

DeGroot, M. H., P. I. Feder, and P. K. Goel (1971), "Matchmaking." *Annals of Mathematical Statistics* 42: 578-93.

DeGroot, M. H., and P. K. Goel (1976), "The matching problem for multivariate normal data." *Sankhya* 38: 14-29.

DeGroot, M. H., and P. K. Goel (1980), "Estimation of the correlation coefficient from a broken random sample." *Annals of Statistics* 8: 264-78.

Devereux, P. J. (2003), Small sample bias in synthetic cohort models of labor supply, mimeo.

Donald, S., and W. Newey (2001), "Choosing the number of instruments." *Econometrica* 69: 1161-91.

Fair, M., M. Cyr, S.W. Wen, G. Guyon, R.C. MacDonald, J.W. Buehler, K. Prager, C.J. Hogue, K. Shamsuddin, E. Lieberman, T.K. Young, E. Kliewer, J. Blanchard, and T. Mayer (2000), "An assessment of the validity of a computer system for probabilistic record linkage of birth and death records in Canada. The fetal and infant health study group." *Chronic Diseases in Canada* 21: 8-13.

Fellegi, I. P. (1999), "Record linkage and public policy." In *Record linkage techniques-1997*, 3-12. Washington, DC: National academy press.

Fellegi, I. P., and A. B. Sunter (1969), "A theory of record linkage." *Journal of the American Statistical Association* 64: 1183-1210.

Fréchet, M (1951), "Sur les tableaux de correlation dont les marges sont données." *Annales de Université,Lyons Sect A* 14: 53-77.

Goodman, L. (1953), "Ecological regressions and behavior of individuals." *American Sociological Review* 18: 663-64.

Güell, M., and L. Hu (2003), Estimating the probability of leaving unemployment using uncompleted spells from repeated cross-section data. Working Paper 473, Industrial Relations Section, Princeton.

Härdle, W. (1990), *Applied nonparametric regression.* New York: Cambridge

University Press.

Hajek, J., and Z. Sidak (1967), *Theory of Rank Tests.* New York: Academic Press.

Hansen, L. P. (1982), "Large sample properties of generalized method of moments estimators." *Econometrica* 50: 1029-54.

Hartley, H. O., and R. R. Hocking (1971), "The analysis of incomplete data." *Biometrics* 27: 783-823.

Heckman, J. J. (1978), "Dummy endogenous variables in a simultaneous equation system." *Econometrica* 46, no. 6: 931-59.

Heckman, J. J., and R. Robb (1985), "Alternative methods for evaluating the impact of interventions." In *Longitudinal Analysis of the Labor Market*, edited by J. Heckman and B. Singer. Cambridge: Cambridge University Press.

Horowitz, J., and C. F. Manski (1995), "Identification and robustness with contaminated and corrupted data." *Econometrica* 63: 281-302.

Horvitz, D. G., and D. J. Thompson (1952), "A generalization of sampling without replacement from a finite universe." *Journal of the American Statisti-*

*cal Association* 47: 663-85.

Hu, Y., and G. Ridder (2003), Estimation of nonlinear models with measurement errors using marginal information. Working Paper, CLEO, University of Southern California.

Ichimura, H., and E. Martinez-Sanchis (2005), "Identification and estimation of GMM models by a combination of two data sets", mimeo, University College London.

Imbens, G. W., W. K. Newey, and G. Ridder (2004), Mean-squared-error calculations for Average Treatment Effects, mimeo.

Kadane, J. B. (1978), "Some problems in merging data files." In *Compendium of Tax Research*. 1978 ed., 159-79. Washington, DC: Office of Tax Analysis, Department of the Treasury.

King, G. (1997), *A Solution to the Ecological Inference Problem: Reconstructing Individual Behavior from Aggregate Data.* Princeton: Princeton University Press.

Klevmarken, W. A. (1982), Missing variables and two-stage least-squares estimation from more than one data set. 1981 Proceedings of the American Statis-

tical Association, Business and Economic Statistics Section, 156-161.

Lee, T., G. Judge, and A. Zellner (1970), *Estimating the Parameters of the Markov Probability Model from Aggregate Time Series Data.* Amsterdam and London: North-Holland.

Lusardi, A. (1996), "Permanent income, current income, and consumption: Evidence from two panel data sets." *Journal of Business and Economic Statistics* 14: 81-90.

MacRae, E. C. (1977), "Estimation of time-varying Markov processes with aggregate data." *Econometrica* 45: 183-98.

Madansky, A. (1959), "Least squares estimation in finite Markov processes." *Psychometrika* 17: 149-67.

McKenzie, D. (2004), "Asymptotic Theory for Heterogeneous Dynamic Pseudo-Panels." *Journal of Econometrics* / 120: 235-262.

Miller, G. (1952), "Finite Markov processes in psychology." *Psychometrika* 24: 137-44.

Moffitt, R. (1993), "Identification and estimation of dynamic models with a time

series of repeated cross sections." *Journal of Econometrics* 59: 99-123.

Nelson, F., and L. Olsen (1978), "Specification and estimation of a simultaneous equation model with limited dependent variables." *International Economic Review* 19: 695-705.

Neter, J., E. S. Maynes, and R. Ramanathan (1965), "The effect of mismatching on the measurement of response errors." *Journal of the American Statistical Association* 60: 1005-27.

Newcombe, H. B. (1988) *Handbook of Record Linkage: Methods for Health and Statistical Studies, Administration and Business.* Oxford: Oxford University Press.

Newcombe, H. B., M. E. Fair, and P. Lalonde (1992), "The use of names for linking personal records." *Journal of the American Statistical Association* 87: 1193-1208.

Newcombe, H. B., J. M. Kennedy, S. J. Axford, and A. P. James (1959), "Automatic linkage of vital records." *Science* 130: 954-59.

Newey, W. K. (1986), "Linear instrumental variable estimation of limited dependent variable models with endogenous explanatory variables." *Journal of*

*Econometrics* 32: 127-41.

Newey, W. K., and J. L. Powell (2003), "Instrumental variables estimation for nonparametric models", *Econometrica* 71: 1565-1578.

Okner, B. A. (1972), "Constructing a new data base from existing microdata sets: The 1966 merge file." *Annals of Economic and Social Measurement* 1: 325-62.

Okner, B. A. (1974), "Data matching and merging: An overview." *Annals of Economic and Social Measurement* 3: 347-52.

Pagan, A. (1984), "Econometric issues in the analysis of regressions with generated regressors." *International Economic Review* 25: 221- 47.

Paxson, C., and J. Waldfogel (2002), "Work, welfare and child maltreatment." *Journal of Labor Economics* 20: 435-74.

Pelzer, B., R. Eisinga, and P. H. Franses (2004), "Ecological panel inference from repeated cross sections.", In *Ecological Inference. New Methodological Strategies*, edited by G. King, O. Rosen, and M. Tanner. Cambridge: Cambridge University Press.

Pelzer, B., R. Eisinga, and P. H. Franses (2002), "Inferring transition probabilities from repeated cross sections." *Political Analysis* 18: 113-33.

Perrachi, F., and F. Welch (1994), "Trends in labor force transitions of older men and women." *Journal of Labor Economics* 12: 210-42.

Radner, D. B. (1974), "The statistical matching of microdata sets: The Bureau of Economic Analysis 1964 Current Population Survey-Tax model match." Ph.D. Thesis, Department of Economics, Yale University.

Radner, D. B. (1978), "The development of statistical matching in economics." *Proceedings of the American Statistical Association, Social Statistics Section* 503-8.

Radner, D. B. , R. Allen, M.E. Gonzalez, T.B. Jabine, and Muller, H.J. (1980), Report on exact and statistical matching techniques. Statistical Policy Working Paper no. 5, US Department of Commerce, Washington DC.

Raessler, S. (2002), *Statistical Matching: A Frequentist Theory, Practical Applications, and Alternative Bayesian Approaches.* New York: Springer.

Rivers, D., and Q. Vuong (1988), "Limited information estimators and exogeneity tests for simultaneous probit models." *Journal of Econometrics* 39: 347-66.

Rodgers, W. L. (1984), "An evaluation of statistical matching." *Journal of Business and Economic Statistics* 2: 91-102.

Rodgers, W., and E. DeVol (1982), An evaluation of statistical matching. 1981 Proceedings of the American Statistical Association, Section on Survey Research Methods, 128-132.

Rubin, D. B. (1986), "Statistical matching using file concatenation with adjusted weights and multiple imputations." *Journal of Business and Economic Statistics* 4: 87-94.

Rubin, D. B., and D. Thayer (1978), "Relating tests given to different samples." *Psychometrika* 43: 3-10.

Ruggles, N., and R. Ruggles (1974), "A strategy for merging and matching microdata sets." *Annals of Economic and Social Measurement* 3: 353-71.

Ruggles, N., R. Ruggles, and E. Wolff (1977), "Merging microdata: Rationale, practice, and testing." *Annals of Economic and Social Measurement* 6: 407-29.

Scheuren, F., and W. E. Winkler (1993), "Regression analysis of data files that are computer matched." *Survey Methodology* 19: 39-58.

Sevestre, P., and A. Trognon (1996), "Dynamic linear models." In *The Econometrics of Panel Data: A Handbook of the Theory with Applications*. 2nd ed., edited by L. Mátyás and P. Sevestre. Dordrecht: Kluwer.

Sims, C. A. (1972), "Comments." *Annals of Economic and Social Measurement* 1: 343-45.

Smith, R., and R. Blundell (1986), "An exogeneity test for a simultaneous equation Tobit model with an application to labor supply." *Econometrica* 54: 679-85.

Tepping, B. J. (1968), "A model for optimal linkage of records." *Journal of the American Statistical Association* 63: 1321-32.

Van der Vaart, A. (1998), *Asymptotic statistics.* Cambridge, UK: Cambridge University Press.

Verbeek, M. (1996), "Pseudo panel data." In *The Econometrics of Panel Data*, edited by L. Matyas and P. Sevestre.

Verbeek, M., and T. Nijman (1992), "Can cohort data be treated as genuine panel data?" *Empirical Economics* 17: 9-23.

Verbeek, M., and T. Nijman (1993), "Minimum MSE estimation of a regression model with fixed effects from a series of cross sections." *Journal of Econometrics* 59: 125-36.

Wald, A. (1940), "The fitting of straight lines if both variables are subject to error." *Annals of Mathematical Statistics* 11: 284-300.

Woodbury, M. (1983) "Statistical record matching for files." In *Incomplete data in sample surveys*, edited by W. G. Madow and I. Olkin, 173- 81. Vol. 3. New York: Aacdemic Press.

Wooldridge, J. (1999), "Asymptotic properties of weighted M-estimators for variable probability samples." *Econometrica* 67: 1385-1406.

**Appendix**

**Theorem 1**

If assumptions (A1)-(A3) hold, then the 2SIV estimator is weakly consistent.

**Proof**

We have by adding and subtracting $m_N(\theta_0)$

$$m_N(\theta)'W_N m_N(\theta) = (m_N(\theta) - m_N(\theta_0))'W_N(m_N(\theta) - m_N(\theta_0)) + \tag{135}$$

$$+2m_N(\theta_0)'W_N(m_N(\theta) - m_N(\theta_0)) + m_N(\theta_0)'W_N m_N(\theta_0)$$

By the mean value theorem

$$m_N(\theta) = m_N(\theta_0) + \frac{\partial m_N}{\partial \theta'}(\theta_*)(\theta - \theta_0) \tag{136}$$

with $\theta_*$ between $\theta$ and $\theta_0$. Substitution in (135) and taking the limit $N_1, N_2 \to \infty$ gives

$$(\theta - \theta_0)'\mathrm{E}\left[\frac{\partial m'}{\partial \theta}(\theta_*)\right] W \mathrm{E}\left[\frac{\partial m}{\partial \theta'}(\theta_*)\right](\theta - \theta_0) + \tag{137}$$

$$2\mathrm{E}[m(\theta_0)]'W\mathrm{E}\left[\frac{\partial m}{\partial \theta'}(\theta_*)\right](\theta - \theta_0) + \mathrm{E}[m(\theta_0)]'W\mathrm{E}[m(\theta_0)]$$

and this limit is attained uniformly in $\theta$. If (A1) holds, then $E(m(\theta_0)) = 0$, so that the last two terms on the right-hand side are equal to 0. Because $\mathrm{E}\left[\frac{\partial m'}{\partial \theta}(\theta)\right]$ is continuous in $\theta$ this matrix has full rank in a neighborhood of $\theta_0$. In that neighborhood $\theta_0$ is the unique minimizer. By Van der Vaart (1998), Theorem

5.7, this implies that the 2SIV estimator converges in probability to $\theta_0$.

**Theorem 2**

If assumptions (A1)-(A4) hold, then

$$\sqrt{N_2}(\hat{\theta}_N - \theta_0) \xrightarrow{d} N(0, V(\theta_0)) \tag{138}$$

with

$$\begin{aligned} V(\theta_0) = &\left[ \mathrm{E}\left( \frac{\partial m'}{\partial \theta}(\theta_0) \right) W(\theta_0) \mathrm{E}\left( \frac{\partial m}{\partial \theta'}(\theta_0) \right) \right]^{-1} . \\ &. \mathrm{E}\left( \frac{\partial m'}{\partial \theta}(\theta_0) \right) W(\theta_0) (\lambda \mathrm{Var}(m_{1j}(\theta_0)) + \mathrm{Var}(m_{2i}(\theta_0))) W(\theta_0) \mathrm{E}\left( \frac{\partial m}{\partial \theta'}(\theta_0) \right) . \\ &. \left[ \mathrm{E}\left( \frac{\partial m'}{\partial \theta}(\theta_0) \right) W(\theta_0) \mathrm{E}\left( \frac{\partial m}{\partial \theta'}(\theta_0) \right) \right]^{-1} \end{aligned} \tag{139}$$

**Proof**

The first order conditions give

$$0 = \frac{\partial m'_N}{\partial \theta}(\hat{\theta}_N) W_N \sqrt{N_2} m_N(\hat{\theta}_N) \tag{140}$$

By the mean value theorem we have for some $\overline{\theta}_N$ between $\theta_0$ and $\hat{\theta}_N$

$$\sqrt{N_2} m_N(\hat{\theta}_N) = \sqrt{N_2} m_N(\theta_0) + \frac{\partial m_N}{\partial \theta'}(\overline{\theta}_N) \sqrt{N_2}(\hat{\theta}_N - \theta_0) \tag{141}$$

Substitution in (140) and solving for $\sqrt{N_2}(\hat{\theta}_N - \theta_0)$ gives

$$\sqrt{N_2}(\hat{\theta}_N - \theta_0) = - \left[ \frac{\partial m'_N}{\partial \theta}(\hat{\theta}_N) W_N \frac{\partial m_N}{\partial \theta'}(\overline{\theta}_N) \right]^{-1} \frac{\partial m'_N}{\partial \theta}(\hat{\theta}_N) W_N \sqrt{N_2} m_N(\theta_0) \tag{142}$$

The proof is completed by noting that $\frac{\partial m_N}{\partial \theta}(\theta)$ is continuous in $\theta$, and by using the central limit theorem for i.i.d. random variables to obtain the asymptotic distribution of $\sqrt{N_2} m_N(\theta_0)$.

**Theorem 3**

If (A1)-(A4) hold, then $T_N \xrightarrow{d} \chi^2(\dim(m) - \dim(\theta))$.

**Proof**

Substitution of (142) in (141) gives

$$\sqrt{N_2} m_N(\hat{\theta}_N) = \tag{143}$$

$$= \left[ I - \frac{\partial m_N}{\partial \theta'}(\overline{\theta}_N) \left[ \frac{\partial m_N'}{\partial \theta}(\hat{\theta}_N) W_N \frac{\partial m_N}{\partial \theta'}(\overline{\theta}_N) \right]^{-1} \frac{\partial m_N'}{\partial \theta}(\hat{\theta}_N) W_N \right] \sqrt{N_2} m_N(\theta_0)$$

Using the notation $A(\theta) = \frac{\partial m_N'}{\partial \theta}(\theta)$ and the assumption that this matrix is continuous in $\theta$, we have

$$\sqrt{N_2} m_N(\hat{\theta}_N) = \left[ I - A(\theta_0)'(A(\theta_0) W A(\theta_0)')^{-1} A(\theta_0) W \right] \sqrt{N_2} m_N(\theta_0) + o_p(1) \tag{144}$$

Upon substitution of (144) in (94)

$$T_N = \sqrt{N_2} m_N(\theta_0)' \left[ I - W' A(\theta_0)'(A(\theta_0) W A(\theta_0)')^{-1} A(\theta_0) \right] W. \tag{145}$$

$$. \left[I - A(\theta_0)'(A(\theta_0)WA(\theta_0)')^{-1}A(\theta_0)W\right]\sqrt{N_2}m_N(\theta_0) + o_p(1) =$$

$$= \sqrt{N_2}m_N(\theta_0)'\left[W - W'A(\theta_0)'(A(\theta_0)WA(\theta_0)')^{-1}A(\theta_0)W\right]\sqrt{N_2}m_N(\theta_0) +$$

$$+o_p(1)$$

If $W = M(\theta_0)^{-1}$, we can find a matrix $M(\theta_0)^{-\frac{1}{2}}$ with $M(\theta_0)^{-1} = M(\theta_0)^{-\frac{1}{2}}M(\theta_0)^{-\frac{1}{2}}$. Then

$$T_N = \sqrt{N_2}m_N(\theta_0)'M(\theta_0)^{-\frac{1}{2}}. \tag{146}$$

$$. \left[I - M(\theta_0)^{-\frac{1}{2}}A(\theta_0)'(A(\theta_0)M(\theta_0)^{-1}A(\theta_0)')^{-1}A(\theta_0)M(\theta_0)^{-\frac{1}{2}}\right].$$

$$.M(\theta_0)^{-\frac{1}{2}}\sqrt{N_2}m_N(\theta_0) + o_p(1)$$

Because $\sqrt{N_2}m_N(\theta_0)'M(\theta_0)^{-\frac{1}{2}} \xrightarrow{d} N(0, I)$ and the matrix between [.] is idempotent with rank equal to $\dim(m_N) - \dim(\theta)$, the result follows.

Figure 2: Bounds on $F(y|x_1)$ and $F(y|x_2)$).

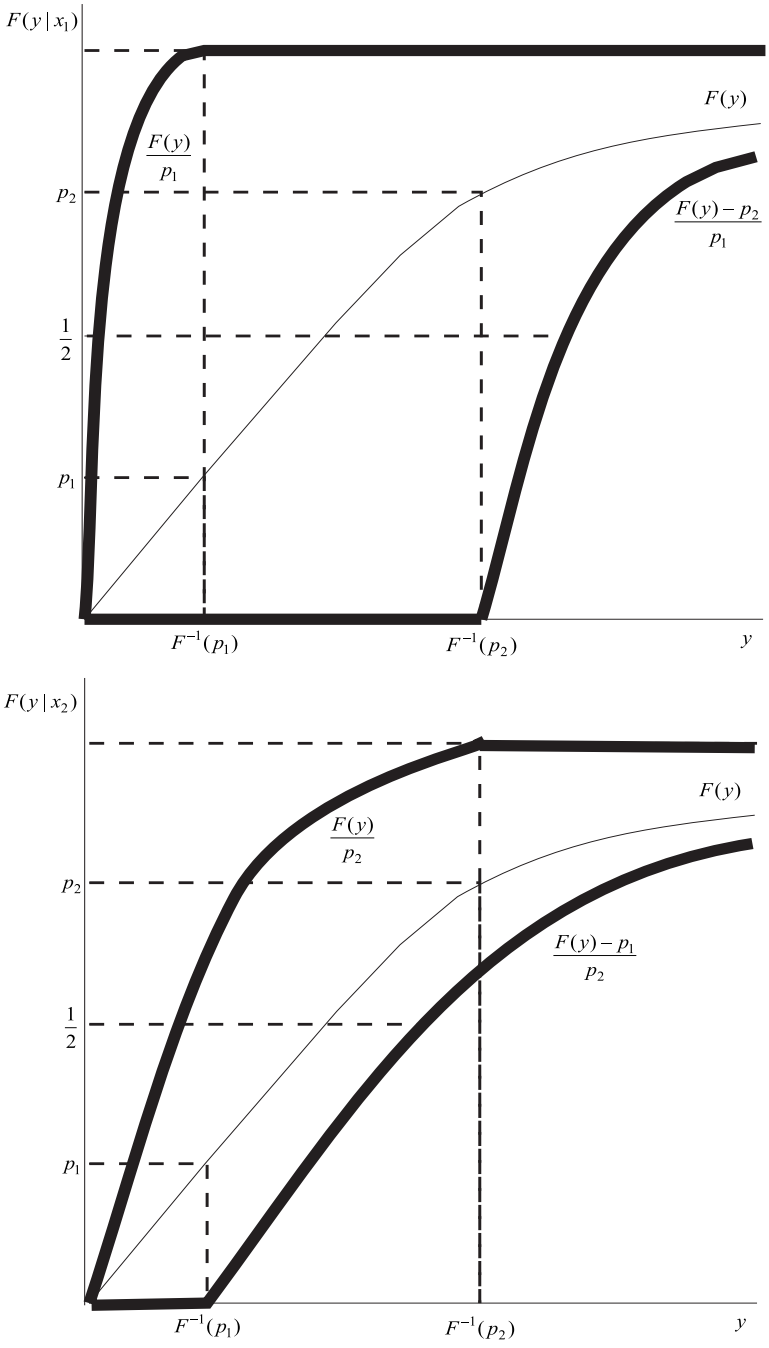


Figure 3: Bounds on $(F(y \mid x_1), F(y \mid x_2), F(y \mid x_3)$ in underidentified case; $p_k(z_l) \leq \frac{1}{2},\ k = 1,2,3,\ l = 1,2$ and $y < \min\{F^{-1}(p_k(z_l)),\ k = 1,2,3,\ l = 1,2\}$.