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
Comments invited
November 7, 2003
Contents

1. Introduction

2. Combining samples with some common variables from a single population
   2.1. Reasons for sample combination
   2.2. Merging samples
      2.2.1. Broken random samples
      2.2.2. Probabilistic record linkage
      2.2.3. Matching errors and estimation
   2.3. Statistical matching
      2.3.1. Fréchet bounds and conditional Fréchet bounds
      2.3.2. Matching of independent samples
   2.4. Estimation without merging
      2.4.1. Types of inference
      2.4.2. Non-parametric inference
      2.4.3. Parametric inference
      2.4.4. A review of the literature
      2.4.5. Combining samples to correct for measurement error
3. Combining biased samples and marginal information

3.1. Biased samples and marginal information

3.2. Identification in biased samples

3.3. Non-parametric and efficient estimation in biased samples

3.3.1. Efficient non-parametric estimation in biased samples

3.3.2. Efficient parametric estimation in endogenously stratified samples

3.3.3. Efficient parametric estimation with marginal information

4. Repeated cross-sections

4.1. General principles

4.2. Consistency and related issues

4.3. Binary choice models

4.4. 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 (i) there is no single sample that contains all relevant variables, (ii) one of the samples has all relevant variables, but the sample size is too small, (iii) the survey uses a stratified design. A special case of (i) occurs if longitudinal or panel data are needed, while only repeated cross-sections are available.

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. Samples A and B give a random sample from the joint distribution of $Y, Z, X$, if the two random samples have a substantial fraction of common units and these common units can be identified from the common variable(s) $Z$. This is the case if $Z$ is a unique identifier, for instance the social security number of an individual (reported without error). In section 2.2 we consider the merger of samples, if there is a substantial fraction of common
units, but the identification of these common units is subject to error. 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 2.3). As shown in section 2.4 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 2.4 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.
In section 3 we consider the combination of samples with common variables that are drawn from possibly overlapping subpopulations of some target population. We distinguish between (i) all samples have the same set of variables, but they are drawn from distinct subpopulations, (ii) there is one sample that has all variables of interest and at least one other sample that is drawn from the same population, but contains a subset of the variables of interest. Case (i) occurs if the sample design is stratified. Often, a simple random sample from a population is not the most efficient sample design. If subpopulations are identifiable from the sampling frame, a design that oversamples heterogeneous subpopulations and undersamples homogeneous ones, requires fewer observations to achieve the same accuracy. Such a sample design is called a stratified design (with unequal probabilities of selection). It may even be that in a simple random sample certain subpopulations that are of particular interest will not be represented at all. For instance, if the dependent variable is the indicator of a rare event, there may be insufficient observations to study factors that affect the occurrence of the event. With a stratified sample design the samples from the strata must be combined. The procedure depends on the type of inference, in particular on whether the inference is on the conditional distribution of a (vector of) dependent variable(s) given a set of conditioning variables or not. If the strata are subsets of the support of the conditioning variables or of variables that are independent of the dependent variables given the conditioning variables, then the stratified sample can be treated as a random sample. If the inference is unconditional or if the strata are subsets of the support of the
dependent variables, then we cannot treat the stratified sample as if it were a random sample. The correct procedure is to use some weighting scheme that uses the inverse probability of selection as the sampling weights (Horvitz and Thompson (1952)).

In case (ii) a small sample with all relevant variables is typically combined with a larger sample with fewer variables. The goals is to increase the precision of the estimates obtained from the small sample. The main difference between the sample combination considered in sections 2 and 3 is that in section 2 the issue is whether the distribution, moments or parameters of interest are identified from the combined samples. In section 3 identification is usually ensured (see section 3.2 for a discussion of the conditions) and the focus is on efficient inference.

There are instances in which the two types of sample combination occur simultaneously. In a simple case the goal is to estimate the conditional probability \( \Pr(Y = 1 \mid X = x) \) with \( Y \) a dichotomous 0-1 variable. Sample A is a sample from the distribution of \( X \) for the stratum defined by \( Y = 1 \). Sample B is a sample from the marginal distribution of \( X \), and we do not observe \( Y \). If we combine the samples A and B, then \( \Pr(Y = 1 \mid X = x) \) is non-parametrically identified if \( \Pr(Y = 1) \) is known. This example that is referred to as estimation from a contaminated sample (Lancaster and Imbens (1996)) is close to the type of sample combination considered in section 3, be it that in this case the strata overlap.

A special case of sample combination with some distinct variables are syn-
thetic 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 4.

2 Combining samples from a single population that have some common variables

2.1 Reasons for sample combination

The research questions that can be addressed by a sample survey are limited by the variables that are included in the survey. 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 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, economists often confront the situation that there is no single data set that contains all variables of interest. However, it may be that 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 the sequel we only consider the combination of two samples. Sample \( A \) contains the variables \( Y \) and \( Z \) and sample \( B \) the variables \( Z \) and \( X \). The variables \( Z \) are implicitly defined as the variables that are in both \( A \) and \( B \). Some-
times 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. We distinguish between three goals that one may want to achieve by combining the two samples

(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 2.3. 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 sections 2.4 and 4, 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 (Angrist and Krueger (1992)) combining data from the US censuses in 1960 and 1980.
(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 2.3.

In section 2.4.5. we show that sample merger is not necessary to avoid measurement error bias.

Although this list of research questions is neither exhaustive, nor exclusive, it is useful in organizing the discussion of methods and applications.

For problems of type (i) there are a number of methods that try to merge the samples A and B into one sample that can be treated as a random sample from the joint distribution of $X, Y, Z$. Because the common variables 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 subsections 2.2.1 and 2.2.2. 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 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. In subsection 2.2.1 we consider the case that $Z$ is a (vector of) continuous real variable(s). Subsection 2.2.2 discusses the more realistic case that $Z$ is a (vector of) numeric variable(s). We also 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.

In subsection 2.3 we consider statistical matching of samples that have no common population members. Uncritical application of techniques that have been derived for samples with overlap in membership creates the impression that the merged sample can be treated as a random sample from the joint distribution of $X, Y$. The problem with this is that the only relation between variables in separate samples is through the common variables $Z$. These common variables induce informative bounds on the joint distribution of $X, Y$. We will argue that it is preferable not to merge the independent samples, but to consider inference that combines the information in these samples. This is the subject of section 2.4. We discuss estimation from the separate samples under the assumption of conditional independence of $X, Y$ given $Z$. Often the assumption of conditional independence is too strong. Therefore we also consider identification by exclusion restrictions. We discuss nonparametric identification
by sample combination under exclusion restrictions. Parametric inference is easier and we show that a large class of relations can be estimated from independent samples if we are prepared to make exclusion restrictions. Finally, in subsection 2.4.5 we consider combining samples to reduce measurement error bias.

2.2 Merging samples

An obvious way to combine information in two samples is to merge them. We distinguish between two situations

(i) A substantial fraction of the observations in the two samples is for the same units.

(ii) The fraction of observations for the same units in the two samples is negligible.

In case (i) 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, and if the two surveys are drawn using a sample design with unequal probabilities of selection, the probability that a unit is in the linked survey is the product of the probabilities that the unit is in one of them.

This is important if the probability of selection, or the sampling weight which is the inverse of this probability, is needed for inference, e.g. if the probability of selection depends on the dependent variable of a regression model.

2.2.1 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

\[
Z_{ki} = Z_i + \varepsilon_{ki}, \quad k = 1, 2
\]  

(1)

The identifier \(Z\) is observed with error and unit \(i\) is erroneously linked to unit \(j\).

We ignore for the moment \(Y, X\). We also assume that \(Z, \varepsilon_1, \varepsilon_2\) are jointly normally distributed\(^2\), and as a consequence the observed \(Z_1, Z_2\) have a bivariate

---

\(^1\)If \(Y, X\) are correlated (given \(Z_1, Z_2\)) they could be helpful in reconstructing the correctly linked sample.

\(^2\)This assumption can be relaxed, see DeGroot, Feder, and Goel (1976)
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

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

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 $\phi$

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

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 for $N = 2$ and it can be shown for all $N$ (Hájek and Šidák (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 $\rho_{\text{max}}$ if $\rho^2_{\text{max}} > \rho^2_{\text{min}}$ and $\rho_{\text{min}}$ 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 $\rho^2_{\text{max}} > \rho^2_{\text{min}}$ 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). Note that we did not make any assumption on the process that linked $i$ with $\phi(i)$ in the data. Hence, it is natural to take the average over all permutations. DeGroot and Goel (1980) show that this average likelihood indeed contains information on $\rho$, although the estimates are not very precise.

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

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

(4)

and this is in turn equivalent to minimization of
\[ \sum_{i=1}^{N} z_{1i}^2 + \sum_{i=1}^{N} z_{2i}^2 - 2 \sum_{i=1}^{N} z_{1i} z_{2\phi(i)} = \sum_{i=1}^{N} (z_{1i} - z_{2\phi(i)})^2 \]  \hspace{1cm} (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

\[ \ln L(\mu, \Sigma_{12}) = \exp \left\{ -\frac{1}{2} \sum_{i=1}^{N} z_{1i}^{'} \Sigma_{12}^{-1} z_{2\phi(i)} \right\} \]  \hspace{1cm} (6)

In this expression

\[ \Sigma_{12} = -\Sigma_{11}^{-1} \Sigma_{12} (\Sigma_{22} - \Sigma_{21} \Sigma_{11}^{-1} \Sigma_{12})^{-1} \]  \hspace{1cm} (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

\[ \sum_{i=1}^{N} z_{1i} C_{12} z_{2\phi(i)} \]  \hspace{1cm} (8)
with $C_{12} = -\Sigma^{12}$. This is equivalent to the minimization of

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

(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 2.3).

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

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

(10)

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

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

(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.
2.2.2 Probabilistic record linkage

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, e.g. 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 2.2.1) and statistical matching (section 2.3).

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 com-
mon, 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, even if they are for the same unit, 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 $k$,

$$h(w_{ij} \mid M_{ij} = 1) = f(w_{ij})$$

(12)
\[ h(w_{ij} \mid M_{ij} = 0) = \int \int f(w_{ij} - z' + z) dG_1(z) dG_2(z') \]

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),

\[
\frac{h(w_{ij} \mid M_{ij} = 1)}{h(w_{ij} \mid M_{ij} = 0)}
\]

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.

The threshold can be found empirically by ordering the \( N_1 \times N_2 \) comparison vectors according to a decreasing density ratio and choosing the cutoff value such that the sum of \( h(w_{ij} \mid M_{ij} = 1) \) over the pairs below the cutoff value is equal to a specified probability. This cutoff value controls the non-match error of the procedure. Alternatively, we can control the match error. Instead of the empirical distribution we can use the distribution of the comparison vector over the unmatched pairs to compute the threshold value. Fellegi and Sunter (1969) suggest to control both errors simultaneously, and this is feasible if we are willing to accept a region of density ratios that are inconclusive.

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

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

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

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

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

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

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

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. This sample can be a random sample from the full sample. In the training sample

\[ \text{cov}(Z_1, Z_2) = \Sigma \]  \( \text{(18)} \)

and

\[ \text{var}(Z_1) + \text{var}(Z_2) = 2\Sigma + \Sigma_1 + \Sigma_2 \]  \( \text{(19)} \)

and this suggests obvious method-of-moments estimates of \( \Sigma, \Sigma_1, \Sigma_2 \). Alternatively, Belin and Rubin use the Box-Cox transform to make the distributions of the logarithms of the density ratios (13) (they refer to these quantities as
the weights) normal both for the matched and unmatched pairs. They then use
normal discriminant analysis to estimate the parameters of the discrete normal
mixture distribution of the log density ratios in the population of all compar-
isons under consideration. In particular, they obtain an estimate of the fraction
of comparisons that is matched. Note that the training sample is only used to
obtain the transformation to normality. With this procedure the false-match
probability can be controlled without reliance on the record generating model.
Fellegi and Sunter (1969) suggest a method-of-moments estimator that assumes
that there are at least three variables in the comparison vector and that these
variables are independently distributed both among the matched and unmatched
pairs. This method does not require a training sample either.

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.
2.2.3 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. The bias depends on the data generating process. 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 2.2.2. 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

\[
\Pr(M_{i\phi(i)} = 1) = \Pr(Z_i = Z_{\phi(i)} | |Z_{2\phi(i)} - Z_{1i}| \leq C) = \frac{\Pr(Z_i = Z_{\phi(i)}) \Pr(|\varepsilon_{2\phi(i)} - \varepsilon_{1i}| < C)}{\Pr(Z_i = Z_{\phi(i)}) \Pr(|\varepsilon_{2\phi(i)} - \varepsilon_{1i}| < C) + \Pr(Z_i \neq Z_{\phi(i)}) \Pr(|Z_{\phi(i)} + \varepsilon_{2\phi(i)} - Z_i - \varepsilon_{1i}| < C)} = \\
\frac{\frac{1}{N_2} \phi \left( \frac{C}{\sqrt{\sigma_1^2 + \sigma_2^2}} \right)}{\frac{1}{N_2} \phi \left( \frac{C}{\sqrt{\sigma_1^2 + \sigma_2^2}} \right) + \frac{N_2 - 1}{N_2} \Phi \left( \frac{C}{\sqrt{\sigma_1^2 + \sigma_2^2 + 2\sigma^2}} \right)}
\]

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

\[
\frac{1}{N_2} + \frac{N_2 - 1}{N_2} \sqrt{\frac{\sigma_1^2 + \sigma_2^2}{\sigma_1^2 + \sigma_2^2 + 2\sigma^2}}
\]

and this probability approaches 1 if the reporting error in the identifier is relatively 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.

The conditional probability of a correct match given the identifiers \( Z_1, Z_2 \) is

\[
\Pr(M_{i\phi(i)} = 1 \mid Z_{1i}, Z_{2\phi(i)}) = \frac{\Pr(M_{i\phi(i)} = 1)\phi(Z_{2\phi(i)} - Z_{1i}; \Sigma_1 + \Sigma_2)}{\Pr(M_{i\phi(i)} = 1)\phi(Z_{2\phi(i)} - Z_{1i}; \Sigma_1 + \Sigma_2) + \Pr(M_{i\phi(i)} = 0)\phi(Z_{2\phi(i)} - Z_{1i}; 2\Sigma + \Sigma_1 + \Sigma_2)}
\]

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

\[
g(x_i, z_{1i}, y_{\phi(i)}, z_{2\phi(i)}) = g(x_i, z_{1i}, y_{\phi(i)}, z_{2\phi(i)}; M_{i\phi(i)} = 1) + \\
+ g(x_i, z_{1i}, y_{\phi(i)}, z_{2\phi(i)}; M_{i\phi(i)} = 0)
\]
If the joint density of $X,Y$ is $f(x,y)$, then

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

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

and

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

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

Because we assume that $X,Y$ and $Z,Z_1,Z_2$ are independent, the expression simplifies, and solving for the joint density of $X,Y$ we find

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

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

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

$E[m(X_i, Y_j; \theta)] = 0 \quad (28)$

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 (26) suggests the sample moment conditions that combine information from the sample and the register

$$\frac{1}{N_1} \sum_{i=1}^{N_1} \frac{m(x_i, y_{\phi(i)}; \theta)}{\Pr(M_{i\phi(i)} = 1 \mid z_{1i}, z_{2\phi(i)})} - \frac{1}{N_1} \sum_{j=1}^{N_1} \sum_{k=1}^{N_1} \frac{1 - \Pr(M_{j\phi(k)} = 1 \mid z_{1j}, z_{2\phi(k)})}{\Pr(M_{j\phi(k)} = 1 \mid z_{1j}, z_{2\phi(k)})} m(x_j, y_{\phi(k)}; \theta) = 0 \quad (29)$$

and the weighted GMM estimator of $\theta$ either makes (29) equal to 0 or is the minimizer of a quadratic form in these sample moment conditions. In this expression (but not in (26)) 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 (23) we start with the joint distribution of $X_i, Y_{\phi(i)}$. This will again produce 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

\[
\text{cov}(X_i, Y_i) = \frac{\text{cov}(X_i, Y_{\phi(i)})}{\lambda} \tag{30}
\]

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

\[
\begin{align*}
\mathbb{E}[(Y_i - \alpha - \beta X_i)X_i] &= \\
&= \mathbb{E}[(Y_{\phi(i)} - \alpha - \beta X_i)X_i] - (1 - \lambda) \left[ \mathbb{E}(Y_{\phi(i)}) \mathbb{E}(X_i) - \alpha \mathbb{E}(X_i) - \beta \mathbb{E}(X_i^2) \right] \\
\end{align*}
\]

\[
\begin{align*}
\mathbb{E}[Y_i - \alpha - \beta X_i] &= \\
&= \mathbb{E}[Y_{\phi(i)} - \alpha - \beta X_i] - (1 - \lambda) \left[ \mathbb{E}(Y_{\phi(i)}) - \alpha - \beta \mathbb{E}(X_i) \right] \\
&= \mathbb{E}[Y_{\phi(i)} - \alpha - \beta X_i] \\
\end{align*}
\]

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

\[
\begin{align*}
\beta &= \frac{\text{cov}(X_i, Y_{\phi(i)})}{\lambda \text{var}(X_i)} \\
\alpha &= \mathbb{E}(Y_{\phi(i)}) - \beta \mathbb{E}(X_i) \\
\end{align*}
\]

(33)
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

$$\text{var}(\hat{\beta}) = \frac{\sigma^2}{N_1 N_2 \lambda^2 \text{var}(X)}$$

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

### 2.3 Statistical matching

#### 2.3.1 Fréchet bounds and conditional Fréchet bounds

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\textsuperscript{3}. 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 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

\[
\max\{F(x \mid z) + F(y \mid z) - 1, 0\} \leq F(x, y \mid z) \leq \min\{F(x \mid z), F(y \mid z)\} \quad (35)
\]

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

\[
E[\max\{F(x \mid Z) + F(y \mid Z) - 1, 0\}] \leq F(x, y) \leq \quad (36)
\]

\textsuperscript{3}Fellegi (1990) 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.
The bounds can not be improved, because the upper and lower bound are joint cdf’s with marginal cdf’s equal to \( F(x \mid z) \) and \( F(y \mid z) \). The bounds are obtained if, given \( Z = z \), \( X \equiv Y \) (upper bound) or \( X \equiv -Y \) (lower bound). 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 more informative 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 (36) gives that the lower and upper bound coincide and that

\[
F(x, y) = E[F(y \mid Z) \mid Z \in A(x)] \Pr(Z \in A(x))
\]

(37)

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

\[
\rho_{XY} = \rho_{XY \mid Z} \sqrt{1 - \rho_{XZ}^2} \sqrt{1 - \rho_{YZ}^2 + \rho_{XZ} \rho_{YZ}}
\]

(38)

This gives the bounds
\[
\rho_{XZ} \rho_{YZ} - \sqrt{1 - \rho_{XZ}^2} \sqrt{1 - \rho_{YZ}^2} \leq \rho_{XY} \leq \rho_{XZ} \rho_{YZ} + \sqrt{1 - \rho_{XZ}^2} \sqrt{1 - \rho_{YZ}^2}
\]

(39)

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 bound is more informative 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 informative bounds. Consider a linear regression model

\[
Y = \alpha + \beta X + U
\]

(40)

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

\[
\frac{\sigma_Y}{\sigma_X} = \frac{|\beta|}{\sqrt{R^2}}
\]

(41)

with \( R^2 \) the coefficient of determination of the regression. If we multiply the bounds in (40) 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

\[
\rho_{XZ} = \sqrt{1 - \frac{p \sqrt{R^2}}{2}}
\]

(42)
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.

2.3.2 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 (20) (for probabilistic record linkage) we specify the distance function as a quadratic function of the difference, but other choices are possible\textsuperscript{4}. 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 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

\textsuperscript{4}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.
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 $d_{ij}, i = 1, \ldots, M, \ j = 1, \ldots, M$ with $d_{ij} = 1$ if $i$ and $j$ are matched

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

subject to

$$
\sum_{k=1}^{M} d_{ik} = 1
$$

$$
\sum_{k=1}^{M} d_{kj} = 1
$$

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 (43) 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 \( Z, Y \) in the matched file may differ from that in the original sample B. Note that if we use the distance function (20), 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\textsuperscript{5}. 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 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

\textsuperscript{5}Even better: also include the variability due to parameter uncertainty.
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.

2.4 Estimation from independent samples without sample merging

2.4.1 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 informative, 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 4 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.

2.4.2 Nonparametric inference

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

$$f(x, y) = E(f(x|Z)f(y|Z))$$

(45)

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

$$E(g(X, Y)) = E_{Y|Z}(E(g(X, Y) | Y, Z)) = E_{Y,Z}(E(g(X, Y) | Z))$$

(46)

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. We mention it, because in the literature it is 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)).

Note that a similar method can be used to estimate $E(g(X,Y,Z_0))$ with $Z_0$ a subvector of $Z$.

**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) for the special case $E(Y \mid X = x, Z_0 = z_0)$, i.e. the population regression of $Y$ on $X$ and $Z_0$. If $Y$ is unbounded, the expectation is a discontinuous function of the cdf $F(y \mid X = x, Z_0 = z_0)$. If $Y$ is bounded, then the (conditional) expectation is a continuous function of the (conditional) cdf. Assume that this function is also increasing in the cdf, i.e. if $F_1$ first-order stochastically dominates $F_2$, then the expected value is not smaller for $F_1$. If we combine continuity and monotonicity we can derive bounds on the (conditional) expectation from bounds on the (conditional) cdf. We show that it is relatively easy to derive sharp bounds on the conditional cdf, and even on the vector of conditional cdf’s given the values taken by $X$. The bounds on the
conditional cdf’s are the Fréchet bounds of section 2.3.1 and if we consider the vector of conditional cdf’s, then they must also satisfy the law of total probability. The sharp bounds can be used to obtain sharp bounds on any population parameter that is a monotone continuous function of the (conditional) cdf. Examples of such functions are conditional quantiles and conditional expectations $E(g(h(Y, Z_0), X, Z_0) \mid X = x, Z_0 = z_0)$ with $g$ bounded and monotone in $h$ for (almost all) $x, z_0$. For example, if $Y$ is bounded we can obtain sharp bounds on all conditional moments of $Y$.

The bounds derived by Horowitz and Manski and Cross and Manski assume that $X$ is a discrete variable. They also assume that $Y$ is a continuous variable, but this is not essential. As we show below, if the probability $Pr(X = x \mid Z_0 = z_0)$ is small for some value of $x$ the bounds on the conditional cdf are wide, so that the discreteness of $X$ is essential.

The Fréchet bounds on the conditional cdf can be derived directly. To simplify the notation we omit the common variables $Z$ in the derivation, so that we have two independent samples without common variables. We already know that we should not expect much in this setup, but the simple case is helpful for understanding the general idea.

If $X$ is a discrete random variable with distribution

$$Pr(X = x_k) = p_k \quad k = 1, \ldots, K$$

(47)

then the upper bound on $F(y; x_k) = Pr(Y \leq y, X = x_k)$ is by the same argument that yields the Fréchet bounds on the joint cdf.
\[ F(y; x_k) \leq \min\{F(y), p_k\} \quad (48) \]

and the lower bound is

\[ F(y; x_k) \geq \max\{F(y) - (1 - p_k), 0\} \quad (49) \]

Note that we use a semicolon to distinguish between the joint and the conditional cdf. If \( p_k \leq \frac{1}{2} \)

\[
\begin{align*}
0 \leq F(y; x_k) &\leq F(y), \quad y < F^{-1}(p_k) \\
0 \leq F(y; x_k) &\leq p_k, \quad F^{-1}(p_k) \leq y < F^{-1}(1 - p_k) \\
F(y) - (1 - p_k) \leq F(y; x_k) &\leq p_k, \quad y \geq F^{-1}(1 - p_k) 
\end{align*}
\]

(50)

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

\[
\begin{align*}
0 \leq F(y \mid x_k) &\leq \frac{F(y)}{p_k}, \quad y < F^{-1}(p_k) \\
0 \leq F(y \mid x_k) &\leq 1, \quad F^{-1}(p_k) \leq y < F^{-1}(1 - p_k) \\
\frac{F(y) - (1 - p_k)}{p_k} \leq F(y \mid x_k) &\leq 1, \quad y \geq F^{-1}(1 - p_k) 
\end{align*}
\]

(51)

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 \( Y \) are given by
Horowitz and Manski (1995) and Cross and Manski (2002) without reference to Fréchet bounds. The argument above shows that these bounds are Fréchet bounds and hence are sharp, i.e. they can not be improved without further assumptions.

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

$$0 \leq F(y; x_1) \leq F(y), \ y < F^{-1}(p_1)$$

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

$$F(y) - p_2 \leq F(y; x_1) \leq p_1, \quad y \geq F^{-1}(p_2)$$

$$0 \leq F(y; x_2) \leq F(y), \ y < F^{-1}(p_1)$$

$$F(y) - p_1 \leq F(y; x_2) \leq F(y), \quad F^{-1}(p_1) \leq y < F^{-1}(p_2)$$

$$F(y) - p_1 \leq F(y; x_2) \leq p_2, \quad y \geq F^{-1}(p_2)$$

(52)

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

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

(53)

Hence, the vector of conditional cdf’s is in a set that is the intersection of the Fréchet bounds in (52) and the hyperplane in (53). 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 (53). 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.

The bounds on the conditional cdf’s \(F(y \mid x_1)\) and \(F(y \mid x_2)\) are also given in figure 2. By the law of total probability, the lower bound of \(F(y \mid x_1)\) corresponds with upper bound of \(F(y \mid x_2)\) and the other way around. Note that the bounds are more informative for \(F(y \mid x_2)\) 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

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

(54)

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 confirms our earlier conclusion that the relation between $Y$ and $X$ cannot be inferred from two independent samples 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

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

(55)

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 (55). We find

$$\frac{\tilde{p}_1}{p_1} F(y) \leq \tilde{F}(y) \leq \min \left\{ \frac{\tilde{p}_2}{p_2} F(y), 1 \right\}, \quad y < F^{-1}(p_1)$$

$$1 - \frac{\tilde{p}_2}{p_2} (1 - F(y)) \leq \tilde{F}(y) \leq \min \left\{ \frac{\tilde{p}_2}{p_2} F(y), 1 \right\}, \quad F^{-1}(p_1) \leq y < F^{-1}(p_2)$$

$$1 - \frac{\tilde{p}_2}{p_2} (1 - F(y)) \leq \tilde{F}(y) \leq 1 - \frac{\tilde{p}_1}{p_1} (1 - F(y)), \quad y \geq F^{-1}(p_2)$$

(56)

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

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

(57)

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

48
As noted, a key assumption in the derivation of the bounds is that $X$ is a discrete variable. From (51) 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.

Now that the notation and bounds for the case with discrete $X$ are clear we reintroduce the common variables $Z$. 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 $Z_{c0}$ 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 $F(y; \cdot \mid Z = z)$ or $F(y \mid \cdot, Z = z))$. If we are interested in $F(y; \cdot)$ or $F(y \mid \cdot)$, 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 2.3.1 this averaging results in more informative 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 $Z_{c0}$ 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. As noted before, 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 $Z = (Z_0^c Z_0^c)'$, 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 (2000)). 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
\[ \Pr(X = x_k \mid Z = z) = p_k(z) \tag{58} \]

With the exclusion restriction we have that for all \( z \)

\[ F(y \mid z) = \sum_{k=1}^{K} F(y \mid x_k)p_k(z) \tag{59} \]

If \( Z \) is also discrete, (59) 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 (59) is

\[
\begin{align*}
F(y \mid x_1) &= \frac{p_2(z_1)F(y \mid z_2) - p_2(z_2)F(y \mid z_1)}{p_1(z_2) - p_1(z_1)} \\
F(y \mid x_2) &= \frac{p_1(z_2)F(y \mid z_1) - p_1(z_1)F(y \mid z_2)}{p_1(z_2) - p_1(z_1)}
\end{align*}
\tag{60}
\]
Note that this implies that

$$F(y | x_2) - F(y | x_1) = \frac{F(y | z_2) - F(y | z_1)}{p_1(z_2) - p_1(z_1)}$$  \hspace{1cm} (61)$$

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 (59) 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 \(F(y | \cdot)\) 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

$$E(Y | Z = z) = \mu_Y + \Sigma_{YX} \Sigma_{XX}^{-1} (E(X | Z = z) - \mu_X)$$ \hspace{1cm} (62)$$

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 noted in section 2.4.3 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 \(F(y | \cdot)\) is not identified. However, 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 2 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 more informative 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 $\frac{F(y|z_1)}{p_l(z_1)}$ and $\frac{F(y|z_2)}{p_l(z_2)}$ (upper bound), respectively.

2.4.3 Parametric inference

**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
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

\[
E(m(Y, X, Z_0; \theta)) = 0
\]  

(63)

If we are prepared to make a parametric assumption on the conditional distribution of \( X \) given \( Z \), identified in sample A, we can estimate \( E(m(y, X, z_0; \theta) \mid Z = z) \) for fixed values \( Y = y \) and \( Z = z \) using the data from sample A. The sample moment conditions corresponding to (64) are

\[
\frac{1}{N_2} \sum_{j=1}^{N_2} \hat{E}_{X|Z}(m(Y_j, X, Z_{02j}; \theta) \mid Z_{2j}) = 0
\]

(65)

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

As an example consider the regression model

\[
Y = \beta_1 X + \beta_2 Z_0 + \varepsilon
\]

(66)

The scalar dependent variable \( Y \) and a vector of common variables \( Z_1 \) are observed in sample A. The (scalar) independent variable \( X \) and a vector of common
variables $Z_2$ are observed in sample B. 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

$$E(\varepsilon \mid X, Z) = 0 \quad (67)$$

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 may contain many variables. It is not even necessary to assume that all the variables in $Z$ are exogenous, as suggested by (67). If $X$ is exogenous, only $Z_0$ (or one other variable in $Z$) has to be exogenous. If $X$ is correlated with $\varepsilon$ two variables in $Z$ need to be exogenous (one of them may be $Z_0$).

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

$$E[(Y - \beta_1 X - \beta_2 Z_0)X] = 0$$

$$E[(Y - \beta_1 X - \beta_2 Z_0)Z_0] = 0 \quad (68)$$

Under conditional independence these can be written as
In these expressions $E_{X|Z_1}(X \mid Z_2)$ 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 $E_{YZ_2}[Y|E_{X|Z_1}(X \mid Z_2)]$, $E_{YZ_2}[E_{X|Z_1}(X \mid Z_2)]$, $E_{YZ_2}[E_{X|Z_1}(X^2 \mid Z_2)]$, and $E_{YZ_2}[Z_{02}E_{X|Z_1}(X \mid Z_2)]$, 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 variance matrix of these conditions as weighting matrix.

It is interesting to note that the GMM estimator obtained from (69)-(70) 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 (68) by

\begin{align}
E[(Y - \beta_1 X - \beta_2 Z_0)Z_0^c] &= 0 \\
E[(Y - \beta_1 X - \beta_2 Z_0)Z_0] &= 0
\end{align}

(71)

Because the $Z$’s are in both samples, all expected values in these population moment conditions can be obtained from sample A ($E(XZ_0)$, $E(XZ_0^c)$), sample B ($E(YZ_0)$, $E(YZ_0^c)$) or both ($E(Z_0^2)$, $E(Z_0Z_0^c)$). 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 $E(X \mid Z)$ and $E(X^2 \mid Z)$ 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 $E_{X \mid Z}(m(y,X,z_0;\theta) \mid Z = z)$ 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 \( \hat{E}_{X|Z}(m(y, X, z_0; \theta)) \) which is substituted to obtain the sample moment conditions as an average in sample B. Analysis of this estimator is beyond the scope of this chapter.

**Exclusion restrictions** In section 2.4.2, 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. Inference under weaker conditions that ensure nonparametric identification 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 2.4.2, with \( Z_0^c \) the components of \( Z \) that are not in the relation and satisfy (72), i.e. that are exogenous for the relation between \( Y \) and \( X, Z_0 \). We consider moment conditions that can be written as

\[
E((f(Y; \theta) - g(X, Z_0; \theta))h(Z_0, Z_0^c)) = 0 \tag{72}
\]

with \( f, g, h \) known, possibly up to parameters, 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 \( X_6 \), i.e. the number of common exogenous variables that is excluded from the

\[\text{6} \text{If } Z_0 \text{ is exogenous, then functions, e.g. powers, of } Z_0 \text{ are also exogenous. To avoid}\]
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 (72) 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 (72) 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 (72): 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

$$E(Y - \beta_0 - \beta_1' X - \beta_2' Z_0) = 0$$ (73)

$$E((Y - \beta_0 - \beta_1' X - \beta_2' Z_0) Z_0) = 0$$ (74)

$$E((Y - \beta_0 - \beta_1' X - \beta_2' Z_0) Z_0^c) = 0$$ (75)

Note that we can replace $X$ by $E(X \mid Z_0, Z_0^c)^7$. We can even replace $X$ by the linear approximation to this conditional expectation, i.e. by $\pi_0 + \pi_1' Z_0 + \pi_2' Z_0^c$. 

\footnote{identification by functional form, we need the additional exogenous variables in $Z_1$.}

\footnote{This is a consequence of the equivalence of 2SLS and IV in this type of models}
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

$$\Pr(Y = 1 \mid X, Z_0) = G(\beta_0 + \beta'_1 X + \beta'_2 Z_0)$$

(76)

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

$$E(Y - G(\beta_0 + \beta'_1 X + \beta'_2 Z_0)) = 0$$

(77)

$$E((Y - G(\beta_0 + \beta'_1 X + \beta'_2 Z_0)Z_0) = 0$$

(78)

$$E((Y - G(\beta_0 + \beta'_1 X + \beta'_2 Z_0)Z_0^c) = 0$$

(79)

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

60
The resulting moment equation can not be computed from the separate samples.

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

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

(81)

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 (81) we have that

\[
\ln \Lambda(Y; \theta_1) + \theta_2' X + \theta_3' Z_0 = U
\]  

(82)

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

\[
E((\ln \Lambda(Y; \theta_1) + \theta_2' X + \theta_3' Z_0) Z_0) = 0
\]  

(83)

\[
E((\ln \Lambda(Y; \theta_1) + \theta_2' X + \theta_3' Z_0) Z_0^c) = 0
\]  

(84)

The number of variables in \( Z_0^c \) must at least be equal to the number of parameters in \((\theta_1' \theta_2')'\). Alternatively, we can identify \( \theta_1 \) by making assumptions on the \(8^{*}\) 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.
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 $Z_1$ in the moment conditions. In that case no additional common variables are needed. 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).

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 (72). 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 $Z_{c0}$ must be exogenous. They also have 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 (72) raises some new issues. First, we introduce some notation. Let

---

9 Provided that the identification condition (A3) below is satisfied.
10 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.
\[ m(\theta) = (f(Y; \theta) - g(X, Z_0; \theta))h(Z_0, Z_0^c) \] (85)

and for \( i = 1, \ldots N_1, j = 1, \ldots N_2 \)

\[ m_{2j}(\theta) = f(Y_j; \theta)h(Z_{02j}, Z_{02j}^c) \] (86)

\[ m_{1i}(\theta) = g(X_i, Z_{01i}; \theta)h(Z_{01i}, Z_{01i}^c) \]

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

\[ m_N(\theta) = \frac{1}{N_2} \sum_{j=1}^{N_2} m_{2j}(\theta) - \frac{1}{N_1} \sum_{i=1}^{N_1} m_{1i}(\theta) \] (87)

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 \( N_1, N_2 \to \infty \)

\[ \frac{\partial m_N}{\partial \theta'}(\theta) \overset{p}{\rightarrow} \text{E} \left( \frac{\partial m}{\partial \theta'}(\theta) \right) \]

uniformly for \( \theta \in \Theta \) with \( \Theta \) the parameter space.

(A3) The rank of the matrix \( \text{E} \left( \frac{\partial m}{\partial \theta}(\theta_0) \right) \) 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

\[
E \left( \frac{\partial m}{\partial \theta'}(\theta) \right) = E \left( \frac{\partial f(Y; \theta)}{\partial \theta'} h(Z_{02}, Z_{02}^c) \right) - E \left( \frac{\partial g(X, Z_{01}; \theta)}{\partial \theta'} h(Z_{01}, Z_{01}^c) \right)
\] (88)

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

\[
\hat{\theta}_N = \arg \min_{\theta \in \Theta} m_N(\theta)' W_N m_N(\theta)
\] (89)

with \( W_N \) a weighting matrix that satisfies

\[
W_N \overset{p}{\to} W
\] (90)

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

\[
(\theta - \theta_0)' E \left[ \frac{\partial m}{\partial \theta'}(\theta_*) \right] W E \left[ \frac{\partial m}{\partial \theta'}(\theta_*) \right] (\theta - \theta_0) + 2E[m(\theta_0)]' W E \left[ \frac{\partial m}{\partial \theta'}(\theta_*) \right] (\theta - \theta_0) + E[m(\theta_0)]' W E[m(\theta_0)]
\] (91)

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,

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

Hence, the asymptotic variance matrix of the moment conditions is

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

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)

\[
\sqrt{N_2}(\hat{\theta}_N - \theta_0) \xrightarrow{d} N(0, V(\theta_0))
\]
with

\[ V(\theta_0) = \left[ E \left( \frac{\partial m'}{\partial \theta} (\theta_0) \right) W(\theta_0) E \left( \frac{\partial m}{\partial \theta'} (\theta_0) \right) \right]^{-1}. \]  \hspace{1cm} (95)

\[\cdot E \left( \frac{\partial m'}{\partial \theta} (\theta_0) \right) W(\theta_0) (\lambda \text{Var}(m_{2j}(\theta_0)) + \text{Var}(m_{1i}(\theta_0))) W(\theta_0) E \left( \frac{\partial m}{\partial \theta'} (\theta_0) \right).\]

\[\cdot \left[ E \left( \frac{\partial m'}{\partial \theta} (\theta_0) \right) W(\theta_0) E \left( \frac{\partial m}{\partial \theta'} (\theta_0) \right) \right]^{-1}\]

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 (93). 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 (89) with this weight matrix. This estimator has asymptotic variance

\[\left[ E \left( \frac{\partial m'}{\partial \theta} (\theta_0) \right) (\lambda \text{Var}(m_{2j}(\theta_0)) + \text{Var}(m_{1i}(\theta_0))) E \left( \frac{\partial m}{\partial \theta'} (\theta_0) \right) \right]^{-1}. \]  \hspace{1cm} (96)

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 this estimator cannot be computed
from the independent samples. The inverse of the information matrix gives an indication of the efficiency loss, because 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

$$T_N = N_2 m_N(\hat{\theta}_N)' \left[ \frac{N_2}{N_1} \hat{\text{Var}}(m_{2j}(\hat{\theta}_N)) + \hat{\text{Var}}(m_{1i}(\hat{\theta}_N)) \right]^{-1} m_N(\hat{\theta}_N)$$  \hspace{1cm} (97)

where \(\hat{\text{Var}}\) denotes the sample variance. If (A1)-(A4) hold, then \(T_N \overset{d}{\to} \chi^2(\text{dim}(m) - \text{dim}(\theta))\). 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 \(X, Z_0, f(y | x, z_0; \theta)\). 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 $Z_0, Z_1$

$$f(y \mid z_0, z_1; \theta) = \int f(y \mid x, z_0; \theta) g(x \mid z_0, z_1) dx \quad (98)$$

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 2.4.2 we considered nonparametric identification of $f(y \mid x_1, z_0)$, and nonparametric identification is sufficient for parametric identification.

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
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.

2.4.4 A review of the literature

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). Klevmarken derives the 2SIV estimator for a single equation that is part of a system of linear simultaneous equations. In our notation he considers

\[ Y = \beta_0 + \beta_1' X + \beta_2' Z_0 + \varepsilon \] (99)

with \( X \) observed in sample A and \( Y \) in sample B, while \( Z_0 \) is a subvector of

\[ 11 \] These authors do not cite Klevmarken's contribution
the common variables $Z$. He also assumes that all the variables in $X$ are endogenous\textsuperscript{12}, that all the common variables $Z$ are exogenous and that $Z$ contains all exogenous variables\textsuperscript{13}. 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.

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\textsuperscript{14} and hence the assumption of multivariate normality is not needed. Carroll and Weil (1994) start from the same model as Klevmarken. They claim\textsuperscript{15} 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

\textsuperscript{12}Klevmarken (1982), p. 160
\textsuperscript{13}Klevmarken (1982), p. 159
\textsuperscript{14}Provided that the distribution of the common variables in the two samples is the same.
\textsuperscript{15}See the Technical Appendix t their paper.
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, that the 2SIV only imputes missing variables, if the model is a linear regression. In the general case specified in (72), there is no imputation of missing variables. This restricts our ability to deal with endogenous variables if these variables enter nonlinearly. This problem also occurs, if all variables are observed in a single sample, and the solutions that have been proposed (e.g. Newey (1986)) apply if the variables are observed in two independent samples. A discussion of this issue is beyond the scope of this survey, and to avoid it we restrict our discussion of 2SIV with moment conditions in which $X$ enters nonlinearly, to the case that all variables in $X$ are exogenous.

As noted, the 2SIV estimator was reinvented independently by Arellano and Meghir (1992) and Angrist and Krueger (1992). Arellano and Meghir (1992)
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)

\[
E(m((X, Z_1; \theta)) = 0
\]

(100)

\[
E(m((Y, Z_2; \theta)) = 0
\]

(101)
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 (99). If we relate the \( X \) to the exogenous common variables \( Z \)

\[
X = \Pi Z + \eta
\]

(101)
we can substitute this in (99) to obtain

\[
Y = \beta_0 + \beta_1' \Pi Z + \beta_2' Z_0 + \varepsilon + \beta_1' Z
\]

(102)
If the order condition is satisfied, we can estimate \( \beta \) from the linear regression in (102). Now (101) can be estimated from sample A and (102) from sample B. The corresponding moment conditions are

\[
E((X_1 - \Pi Z_1)' Z_1) = 0
\]

(103)

\[
E((Y - \beta_0 - \beta_1' \Pi Z_2 - \beta_2' Z_0) Z_2) = 0
\]
and this has the form (100). 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.
Angrist and Krueger (1992) consider the linear regression model

\[ Y = \beta_0 + \beta_1'X + \varepsilon \]  

(104)

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 (73)-(75).

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 (1997), and Currie and Yelowitz (1997). 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 (1997) 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 ob-
tained 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 (1997) 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.

2.4.5 Combining samples to correct for measurement error

One of the reasons to merge datasets is that the variables in one of the sets may be 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 relation where some independent variables are measures 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

\[
    f(y \mid x_1, x_2; \theta) = \int f(y \mid x_1^*, x_2; \theta) g(x_1^* \mid x_1, x_2) dx_1^* \quad (105)
\]

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

\[
\Pr(X_1 = 1 \mid X_1^* = 1, X_2) = \Pr(X_1 = 0 \mid X_1^* = 0, X_2) = \lambda \quad (106)
\]

i.e. the probability of misclassification is independent of \( X_1^* \). Moreover, (114) implies that \( X_1 \) is independent of \( X_2 \) given \( X_1^* \). Solving for \( \lambda \) we find

\[
\lambda = \frac{\Pr(X_1 = 1) + \Pr(X_1^* = 1) - 1}{2 \Pr(X_1^* = 1) - 1} \quad (107)
\]

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

Note that (115) only gives solutions between 0 and 1 if

\[
\Pr(X_1 = 1) < \Pr(X_1^* = 1) \quad (108)
\]

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

\[
\Pr(X_1 = 1) > \Pr(X_1^* = 1) \quad (109)
\]

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

\[
\Pr(X_1 = 1)(1 - \Pr(X_1 = 1)) = \text{Var}(X_1) > \text{Var}(X_1^*) = \Pr(X_1^* = 1)(1 - \Pr(X_1^* = 1)) \quad (110)
\]
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\(^{16}\), maintaining the assumption that $X_1$ is independent of $X_2$ given $X^*_1$. Hence

$$
\Pr(X_1 = 0 \mid X^*_1 = 0, X_2) = 1
$$

(111)

$$
\Pr(X_1 = 1 \mid X^*_1 = 1, X_2) = \lambda
$$

With this assumption we find

$$
\lambda = \frac{\Pr(X_1 = 1)}{\Pr(X^*_1 = 1)}
$$

(112)

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

$$
\Pr(X_1 = 0 \mid X^*_1 = 0, X_2) = \lambda_0
$$

\(^{16}\)The misclassification can also only occur if $X^*_1$ is 0.
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 (105)). 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

\[ \Pr(X_1 = 1 \mid X_1^* = 1, X_2) = \lambda_1 \]
\[
g_{x_1, x_1^*, x_2}(x_1, x_1^*, x_2) = g_{x_1}(x_1 \mid x_1^*, x_2)g_{x_1^*, x_2}(x_1^*, x_2) = \\
= g_{z_1}(x_1 \mid x_1^*)g_{z_1^*, x_2}(x_1^*, x_2)
\]

because \( X_1 \) is independent of \( X_2 \) given \( X_1^* \). After substitution of (114) we obtain

\[
g_{x_1, x_1^*, x_2}(x_1, x_1^*, x_2) = \lambda g_{x_1^*, x_2}(x_1^*, x_2), \quad x_1 = x_1^*
\]

\[(115)\]

\[
= (1 - \lambda)g_{x_1^*, x_2}(x_1^*, x_2) \quad x_1 \neq x_1^*
\]

The marginal distribution of \( X_1, X_2 \), which can be observed, is

\[
g_{x_1, x_2}(x_1, x_2) = \lambda g_{x_1^*, x_2}(x_1, x_2) + (1 - \lambda)g_{x_1^*, x_2}(1 - x_1, x_2)
\]

(116)

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

\[
g_{x_1^*, x_2}(x_1^*, x_2) = \frac{(1 - \lambda)g_{x_1, x_2}(1 - x_1^*, x_2) - \lambda g_{x_1, x_2}(x_1^*, x_2)}{1 - 2\lambda}
\]

(117)

Substitution in (115) 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.
3 Combining biased samples and marginal information

3.1 Biased samples and marginal information

In the previous sections we combined random samples from the same population that had (some) population members and/or variables in common. In this section we study the combination of samples that are drawn from distinct, but possibly overlapping subpopulations. The most common case is that of a stratified sample. In a stratified sample the population is divided into non-overlapping subpopulations, the strata, and separate random samples, usually with different sampling fractions, are drawn from these strata. A stratified random sample usually achieves the same accuracy, as measured by the variance of estimates, with a smaller sample size.

If the sampling fraction differs between strata, the members of the population have an unequal probability of being observed. If the probability of observation depends on the variable of interest, or on variables that are correlated with the variable of interest, then the stratified sample gives a biased estimate of the distribution of the variable of interest and any parameter defined for this distribution.

A stratified sample is a special case of a biased sample. A biased sample is a sample in which the probability of observation depends on the variable(s) of interest. Let \( Y \) be the vector of variables of interest. In a biased sample the probability of observation is proportional with \( W(Y) \). The function \( W \) is called
the biasing function. The density of $Y$ in the sample is\textsuperscript{17}

$$g(y) = \frac{W(y)f(y)}{\int_{-\infty}^{\infty} W(v)f(v)dv}$$

(118)

Special cases of biasing functions are $W(y) = I_S(y)$ with $I_S$ the indicator of the subset $S$ of the support of $Y$, i.e. a stratum of the population, and $W(y) = y$, i.e. the probability of selection is proportional to $Y$. If $Y$ is a duration, the second biased sample is called a length-biased sample. A length-biased sample is a biased sample from the full distribution and not a sample from a subpopulation. Estimation from a pure length-biased sample does not involve sample combination.

For biased samples the distinction between the dependent variable(s) $Y$ and the independent variable(s) $X$ is important. In particular, it makes a difference if the distribution of interest is that of $Y$ or the conditional distribution of $Y$ given $X$. If the biasing function $W(y, x)$ is a function of $x$ only, the joint density of $Y$ given $X$ in the sample is

$$g(y, x) = \frac{W(x)f(y|x)f(x)}{\int_{-\infty}^{\infty} W(w)f(w)dw}$$

(119)

The marginal distribution of $X$ in the sample has density

$$g(x) = \frac{W(x)f(x)}{\int_{-\infty}^{\infty} W(w)f(w)dw}$$

(120)

\textsuperscript{17}Here and in the sequel $g$ and $f$ are either pdf’s or mass functions, i.e. densities with respect to the counting measure.
so that the conditional density of $Y$ given $X$ in the sample is the population conditional density $f(y|x)$. Hence, if we are interested in the conditional distribution of $Y$ given $X$ (or parameters defined for this distribution) and the biasing function is a function of $X$ only, the biased sample directly identifies the conditional distribution of $Y$ given $X$. In all other cases, we cannot ignore the fact that we have a biased sample.

In section 3.2 we consider parametric and non-parametric identification in biased samples. In leading cases parametric restrictions secure identification while there is non-parametric underidentification. This precludes tests of these parametric restrictions. Non-parametric identification requires that the biased samples are 'overlapping' (in a sense that will be made precise). Necessary and sufficient conditions for the non-parametric identification of the distribution of $Y$ or the joint distribution of $Y, X$ are given by Gill, Vardi and Wellner (1988). These conditions apply if the biased samples have the same variables. However they cannot be used if some of the subsamples only have a subset of the variables in $Y, X$. It is even possible that we do not observe the subsample itself, but only moments of the variables in the subsample. In these cases non-parametric identification has to be established on a case by case basis.

Efficient estimation from a combination of biased samples is considered in section 3.3. First, we consider efficient non-parametric estimation of the cdf of $Y$ or that of $Y, X$ from a combination of biased samples that non-parametrically identifies these distributions. Next, we consider the special case of an endogenously stratified sample and parametric inference on the conditional distribution
of \( Y \) given \( X \), if the parameters in this distribution are identified\(^{18}\). Finally, we consider the case that a (possibly biased) sample is combined with information from other samples that only specify selected moments of a subset of the variables in \( Y, X \). If the main sample is a random sample then the parameters are identified from this sample and the additional information overidentifies the parameters. The additional degrees of freedom can be used to increase the precision of the estimates or they can be used to test the (parametric) model for the conditional distribution of \( Y \) given \( X \). If the additional information just identifies the parameters there is no gain in precision. Finally, the first sample and additional information may not identify the parameters. In that case the combination may provide more informative bounds on these parameters. An alternative is to define a population that is consistent with all available information and to estimate parameters defined for this population. These parameters are equal to the population parameters in the identified case (Imbens and Hellerstein (1999)).

This final approach has all the earlier efficient parametric estimators as special cases. It also covers the combination of biased samples with samples that have marginal information on a subset of the variables in \( Y, X \). An example is the contaminated sampling problem considered by Lancaster and Imbens (1996) who consider the combination of a sample from the distribution of \( X \) given \( Y = 1 \), \( Y \) is a 0-1 variable, with a random sample from the marginal

\(^{18}\text{Parametric identification suffices, but preferably the conditional distribution of } Y \text{ given } X \text{ should be non-parametrically identified, and for this the strata need to be overlapping.}\)
distribution of $X$.

Some of the subsamples may be very large, e.g. consist of the entire population. If the additional information consist of moments from such a subsample then this information is exact, i.e. it does not have sampling uncertainty.

The theory of biased samples is now fairly complete. The general theory of identification is summarized in Gill, Vardi and Wellner (1988) who also discuss efficient nonparametric estimation of the marginal cdf of $Y$ or the joint cdf of $Y,X$. In econometrics the emphasis has been on parametric inference in the conditional distribution of $Y$ given $X$. The efficient MLE was developed by Imbens (1992). Imbens and Lancaster (1996) consider the general case. The history of this problem is interesting, because the contributions were made by researchers with different backgrounds, which reflects the prevalence of biased samples in different areas. Cox (1969) considered nonparametric inference in length-biased samples. This was followed by a number of contributions by Vardi (Vardi (1982), (1985)), culminating in Gill, Vardi, and Wellner (1988)). In econometrics the problem was first studied in discrete-choice models (Manski and Lerman (1977)). Further contributions are Manski and McFadden (1981), Cosslett (1981a), (1981b), Morgenthaler and Vardi (1986), and Imbens (1992). The case that the dependent variable $Y$ is continuous was studied by Hausman and Wise (1981) and Imbens and Lancaster (1996). Related problems that will be considered in this section are case-control studies (Prentice and Pyke (1979), Breslow and Day (1980)), contaminated samples (Hsieh, Manski, and McFadden (1985), Lancaster and Imbens (1996)) and the combination of micro and macro
3.2 Identification in biased samples

General results on nonparametric identification of the population cdf from combined biased samples are given by Vardi (1985) and Gill, Vardi, and Wellner (1988). Initially, we make no distinction between dependent and independent variables. Let the population distribution of the random vector $Y$ have cdf $F$.

Instead of a random sample from the population with cdf $F$, we have $K$ random but biased samples from distributions with cdf’s $G_k, k = 1, \ldots, K$. The relation between $G_k$ and $F$ is given by

$$G_k(y) = \frac{\int_{-\infty}^y W_k(y) dF(y)}{\int_{-\infty}^{\infty} W_k(v) dF(v)}$$  \hspace{1cm} (121)

In this expression $W_k$ is a biasing function. This function is assumed to be known and nonnegative (it may be 0 for some values of $y$). An obvious interpretation of this function is that it is proportional to the probability of selection. If $f$ is the density of $F$, then the probability of observing $y$ in the $k$-th biased sample is proportional to $W_k(y)f(y)$. Because we specify the probability of selection up to a multiplicative constant we must divide by the integral of $W_k(y)f(y)$ to obtain a proper pdf.

It is obvious that we can only recover the population cdf for values of $y$ where at least one of the weight functions is positive. The region of where $F$ is identified, $\mathcal{S}$, is defined by
\[ S = \left\{ y \mid \sum_{k=1}^{K} W_k(y) > 0 \right\} \] (122)

If \( S \) is a strict subset of the support of \( Y \) we can only recover the conditional cdf of \( Y \) given \( Y \in S \). For values of \( y \) with \( W_k(y) > 0 \), the population pdf can be found from

\[ g_k(y) = \frac{W_k(y)}{W_k} f(y) \] (123)

with

\[ W_k = \int_{-\infty}^{\infty} W_k(w) dF(w) \] (124)

If \( f \) satisfies (123), then so does \( cf \) for any positive constant \( c \). Because \( f \) is a density, the sum or integral over its support is 1, and this restriction determines the constant.

Let \( \bar{f}(y) \) be the density of a randomly selected observation from the pooled sample. If the subsample sizes are determined by a multinomial distribution with parameters \( \lambda_k, k = 1, \ldots, K \) and \( N \) (the size of the pooled sample), then we have a multinomial sampling plan. The density of a randomly selected observation from the pooled sample is \( \bar{f}(y) = \sum_{i=1}^{K} \lambda_k g_k(y) \). In the case that the subsample sizes are fixed, we substitute \( \frac{N_k}{N} \) for \( \lambda_k \) to obtain the density of a randomly selected observation in the pooled sample. This implies that the identification results for multinomial sampling and for fixed subsample sizes are identical.

From (123) we solve for \( f \) as a function of \( \bar{f} \)
This solution does not express \( f \) in terms of observable quantities, because it depends on the unknown \( W_k \)'s. The \( W_k, k = 1, \ldots, K \) are determined by the following system of equations that is obtained by multiplying (125) by \( W_k(y) \) and by integrating the resulting expression over \( y \)

\[
1 = \frac{1}{W_k} \int_{-\infty}^{\infty} \frac{W_k(y)}{\sum_{l=1}^{K} \lambda_l \frac{W_l(y)}{W_l}} g(y) dy
\]  

(126)

for \( k = 1, \ldots, K \). Note that this set of equations only determines the \( W_k \)'s up to a multiplicative factor. To obtain a solution we choose an arbitrary subsample, e.g. subsample 1, and we set \( W_1 = 1 \).

By rewriting (125) (we divide by 1), we find

\[
f(y) = \frac{1}{\sum_{k=1}^{K} \lambda_k \frac{W_k(y)}{W_k}} g(y)
\]

(127)

We see that \( f \) only depends on the ratios \( \frac{W_k}{W_1}, k = 2, \ldots, K \). Hence, we choose the normalization \( W_1 = 1 \). We can now restate the identification problem: The population pdf \( f \) (with cdf \( F \)) is nonparametrically identified from the \( K \) biased samples if and only if the equation system (126) and (127) has a unique solution for \( f \) and \( W_k, k = 2, \ldots, K \) (in the equations we set \( W_1 = 1 \)). If desired we can recover \( W_1 \) from (124) with \( k = 1 \).

We consider the solution in more detail for the case of two biased samples,
Define the set $V_{12}$ by

$$V_{12} = \{ y | W_1(y)W_2(y) > 0 \}$$

Note that if the weight functions are stratum indicators, the $V_{12}$ contains all $y$ that are common to both strata. For all $y \in V_{12}$

$$\frac{W_2}{W_1} = \frac{g_1(y)}{g_2(y)} \frac{W_2(y)}{W_1(y)}$$

Note that the functions on the right-hand side are all known or estimable from the biased samples. Hence, the ratio $\frac{W_2}{W_1}$ is (over)identified on $V_{12}$. This ratio can be substituted in (127) to obtain $f$. If the set $V_{12}$ is empty, then it is not possible to identify the ratio $\frac{W_2}{W_1}$ and $f$.

If $K \geq 3$ we look for biased samples $k, l$ for which the set $V_{kl} = \{ y | W_k(y)W_l(y) > 0 \}$ is not empty, i.e. for which

$$\int_{-\infty}^{\infty} W_k(y)W_l(y)dF(y) > 0$$

The same argument as for $K = 2$ shows that for such a pair of subsamples $k, l$ we can identify the conditional distribution of $Y$ given that $Y$ is in the set where $W_k(y) + W_l(y) > 0$. Samples for which (130) holds are called connected. Because the result holds for all pairs $k, l$ we can characterize the region of identification of the population distribution. Let $\mathcal{K}_m, m = 1, \ldots, M$ be disjoint index sets of connected subsamples. The union of these index sets is the set of all subsamples $\{1, \ldots, K\}$. The population distribution is identified on the regions $\mathcal{S}_m, m = 1, \ldots, M$ with $\mathcal{S}_m = \{ y | \sum_{k \in \mathcal{K}_m} W_k(y) > 0 \}$, i.e. we can identify the conditional distributions of $Y$ given that $Y \in \mathcal{S}_m$. If there is only one region of
identification that coincides with the support of \( Y \), the population distribution is identified on its support.

Until now we did not distinguish between the dependent variable(s) \( Y \) and independent variables \( X \). The theory developed above applies directly if biased samples from the joint distribution of \( Y, X \) are combined. The special case that the biasing function only depends on \( X \) has already been discussed. There are however other possibilities, e.g. that in some subsample only \( Y \) or only \( X \) is observed. A sample from the marginal distribution of \( X \) or \( Y \) cannot be considered as a biased sample from the joint distribution of \( X, Y \), so that the general theory cannot be used. A simple example illustrates this point\(^\text{19}\).

Assume that \( X \) and \( Y \) are both discrete with 2 and \( K \) values and assume that we have random samples from strata defined by \( Y \). The biasing functions are \( W_k(y, x) = I_{y=k}(y, x), k = 1, \ldots, K \). The subsamples are not connected and we cannot identify the joint distribution of \( X, Y \). Now assume that we have an additional random sample from the distribution of \( X \). It seems that the 'biasing' function for this sample is \( I_{x=1,2}(y, x) \) and this additional subsample is connected with each of the other subsamples. We conclude that the joint distribution is identified. This conclusion is not correct, because the marginal density of \( X \) satisfies by the law of total probability

\[
f_X(1) = \sum_{k=1}^{K} f(1|k)f_Y(k)
\]

\(^{19}\)Although Gill, Vardi, and Wellner (1988) do not claim that their identification theorem applies with marginal information, they give suggestive examples, e.g. their Example 4.4.
If \( K = 2 \) we can identify the marginal distribution of \( Y \) and therefore the joint distribution of \( Y, X \) from the biased samples and the marginal distribution of \( X \). If \( K > 2 \), there will be observationally equivalent solutions and we cannot identify the joint distribution. If the additional sample is from the marginal distribution of \( Y \) we can identify the joint distribution. Note that \( W_k = f_Y(k) \) so that this case corresponds to prior information on the \( W_k \)'s. In general, samples from marginal distributions provide prior information on the \( W_k \)'s, e.g. (131) imposes as many restrictions as the number of distinct values taken by \( X \). Currently there is no general theory of nonparametric identification with marginal information that is comparable to the Gill, Vardi, and Wellner (1988) theory.

We now consider some examples:

*Endogenous stratification* First, we consider the marginal distribution of \( Y \). Let \( S_k, k = 1, \ldots, K \) be a partition of the support of \( Y \), and let \( W_k(y) = I_{S_k} \). The population cdf of \( Y \) is not identified, because the biased samples are not connected. If we have a supplementary random sample from the distribution of \( Y \), the biased samples are connected and the cdf is identified. Next, consider the conditional distribution of \( Y \) given \( X \). If the subpopulations partition the support of the joint distribution of \( Y, X \), then the joint and conditional cdf are identified with a supplementary sample from the joint distribution. This conditional cdf is in general not identified if the supplementary sample is from the marginal distribution of \( Y \). If the subpopulations are defined as a partition of the support of \( Y \), then an additional random sample from the marginal distri-
bution of $Y$, suffices for identification of the joint and conditional cdf of $Y, X$, because the $W_k$ can be obtained from the marginal distribution of $Y$. A special case is a case-control study in which $Y$ is 0-1 and the strata are defined by $Y$.

*Case-control with contaminated controls* Consider the case that $Y$ is a 0-1 variable. We combine a random sample from the subpopulation defined by $Y = 1$, i.e. a random sample from the conditional distribution of $X$ given $Y = 1$, with random samples from the marginal distributions of $X$ and $Y$. By the law of total probability $f(x) = f(x|y = 1) \Pr(Y = 1) + f(x|y = 0)(1 - \Pr(Y = 1))$. The marginal distribution of $Y$ identifies $\Pr(Y = 1)$ and combining this with the marginal distribution of $X$ identifies $f(x|y = 0)$. Hence, the joint distribution of $X, Y$ is identified. A sample from the marginal distribution of $X$ does not identify the joint distribution of $Y, X$ nor the marginal distribution of $Y$ given $X$.

Nonparametric identification of the conditional distribution of $Y$ given $X$ is desirable, even if we assume that the conditional cdf is a member of a parametric family $F(y|x; \theta)$. Often, parametric assumptions identify $\theta$ from a single biased sample. Consider

$$f(y, x; \theta) = f(y|x; \theta)h(x) = W_k \frac{g_k(y, x)}{W_k(y, x)}$$  \hspace{1cm} (132)$$

for all $(y, x), (y', x') \in S_k$ with $S_k = \{(y, x)|W_k(y, x) > 0\}$ we have

$$\frac{f(y|x; \theta)}{f(y'|x'; \theta)} = \frac{g_k(y, x)}{g_k(y', x')} \frac{W_k(y', x')}{W_k(y, x)}$$  \hspace{1cm} (133)$$
For instance, if the model is a probit model with \( \Pr(Y = 1|x; \theta) = \Phi(\theta_0 + \theta_1 x) \)
for a dummy dependent \( Y \), and \( W_k \) is the indicator of the stratum \( Y = 1 \), then \( \theta_0, \theta_1 \) are identified from this biased sample. To see this we consider the case that \( x \) is continuous and that 0 and 1 are in the support of \( x \). Fix \( x' \) in (133) and consider the derivative with respect to \( x \) of the logarithm of the resulting expression. Evaluating the result for \( x = 0 \) and \( x = 1 \) gives a (nonlinear) system of two equations in \( \theta_0, \theta_1 \) that can be solved for these two parameters. A more comprehensive discussion of parametric identification in choice-based samples can be found in Lancaster (1992). We do not discuss this type of identification any further, because it should be avoided.

Nonresponse in sample surveys or attrition in panel data also results in biased samples from the underlying population. For conditional inference, the key question is whether the response/attrition depends on \( Y \). Note that in this case the biasing function is in general unknown. The large literature on sample selectivity goes back to Heckman (1979). Sample combination can be used to put restrictions on the biasing function, in this case the probability of response. Hirano et al. (2001) consider the combination of a panel survey with selective attrition and a refreshment sample. Manski (2002) derives bounds on the population distribution under weak assumptions on the missing data process. This type of biased samples is beyond the scope of this survey.
3.3 Nonparametric and efficient estimation in biased samples

3.3.1 Efficient nonparametric estimation in biased samples

The efficient nonparametric estimator of the population cdf from a set of biased samples was first derived by Vardi (1985). Gill et al. (1988) give a rigorous analysis of this estimator and prove that it is asymptotically efficient\(^20\).

Vardi’s estimator is the solution to the empirical counterparts of the equations (127) and (126). The estimator of the cdf is

\[
\hat{F}(y) = \frac{\int_0^y \sum_{k=1}^K \frac{1}{\lambda_k \hat{W}_k(v)} d\hat{G}(v)}{\int_{-\infty}^\infty \sum_{k=1}^K \frac{1}{\lambda_k \hat{W}_k(v)} d\hat{G}(v)} \tag{134}
\]

\[
1 = \frac{1}{W_k} \int_{-\infty}^\infty \frac{W_k(y)}{\sum_{l=1}^K \lambda_l \hat{W}_l(y)} d\hat{G}(y) \quad , k = 2, \ldots, K \tag{135}
\]

In these equations \(\lambda_k = \frac{N_k}{N}\). Integration with respect to the empirical cdf is just summation over the combined sample.

If the cdf is nonparametrically identified, then system of \(K-1\) equations in \(K-1\) unknowns (135) has a unique solution. This solution is substituted in (134) to obtain the nonparametric estimator of the cdf.

Gill et al. (1988) show that the empirical cdf is consistent (at rate \(n^{\frac{1}{2}}\)) and asymptotically normal with a covariance function that can be easily estimated.\(^21\)

\(^{20}\)In the sense that its limit process has a covariance function that reaches the lower bound for all regular estimators.

\(^{21}\)If the dimension of \(y \geq 2\) the result applies to the empirical measure that counts the number of outcomes in a set \(E \subset \mathbb{R}^M\) with \(M\) the dimension of \(y\). There are restrictions on
In the case of endogenous stratification we have \( W_k(y) = I_{S_k}(y) \) with \( S_k, k = 1, \ldots, K \) a partition of the set of values taken by \( Y \). To ensure identification we have an additional random sample and we call this stratum \( K + 1 \) with \( W_{K+1}(y) = 1 \) for all \( y \). We normalize with respect to this stratum so that in (135) we have \( K \) equations in the unknowns \( \hat{W}_1, \ldots, \hat{W}_K \). They are

\[
1 = \frac{1}{W_k} \int_{-\infty}^{\infty} \sum_{l=1}^{K+1} \frac{W_k(y) W_l(y)}{W_l} d\hat{G}(y), \quad k = 1, \ldots, K
\]

Because integration with respect to the empirical cdf \( \hat{G} \) is just averaging over the complete data we obtain

\[
1 = \frac{1}{W_k} \frac{1}{N} \sum_{i=1}^{N} \sum_{l=1}^{K+1} \frac{W_k(y_i)}{W_l} I_{S_k}(y_i) = \frac{1}{W_k} \frac{1}{N} \sum_{i=1}^{N} \frac{1}{\lambda_k \hat{W}_k + \lambda_{K+1}} I_{S_k}(y_i)
\]

(137)

If \( N_k, k = 1, \ldots, K + 1 \) is the sample size in the strata, \( N = N_1 + \ldots + N_{K+1} \), and \( \hat{N}_{K+1,k} \) is the number of observations in the random sample that is in \( S_k \), we have \( \sum_{i=1}^{N} I_{S_k}(y_i) = N_k + \hat{N}_{K+1,k} \)

\[
1 = \frac{N_k + \hat{N}_{K+1,k}}{N_k + N_{K+1} \hat{W}_k}, \quad k = 1, \ldots, K
\]

(138)

with solution

\[
\hat{W}_k = \frac{\hat{N}_{K+1,k}}{N_{K+1}}
\]

(139)

Hence the nonparametric estimator of the empirical cdf is just the sum of the empirical cdf of the random sample and the weighted empirical cdf in the strata with weights \( \frac{\hat{N}_k}{W_k} \), i.e. the ratio of the fraction of stratum \( k \) in the sample and population.

\( \text{the choice of } E, \text{ e.g. the orthants } y \leq c \text{ will do, in order to obtain uniform convergence.} \)

97
3.3.2 Efficient parametric estimation in endogenously stratified samples

We restrict the discussion to parametric models that specify the conditional density \( f(y|x; \theta) \). A special case is the discrete choice model where \( y \) is a categorical variable. The sample space \( \mathcal{Y} \times \mathcal{X} \) is divided into strata \( S_k \). These strata need not be disjoint. Indeed the analysis in section 3.2 shows that to ensure nonparametric identification of \( f(y|x) \) the strata should be overlapping. A special case occurs if \( Y \) is discrete and \( S_y = \{y\} \times \mathcal{X} \) for \( y = 1, \ldots, M \). Such a sample is called a choice-based or response-based sample. In econometrics, estimation in endogenously stratified samples was first discussed in choice-based samples (Manski and Lerman (1977), Manski and McFadden (1981), Cosslett (1981)). The surprisingly simple efficient estimator in such samples was also first discovered for choice-based samples (Imbens (1992)) and later generalized to arbitrary endogenously stratified samples (Imbens and Lancaster (1996)). We use a suggestion by Lancaster (1992) who showed that in choice-based samples the efficient estimator is the Conditional Maximum Likelihood (CML) estimator if we substitute the observed stratum fractions, even if these fractions are specified by the sample design. This is true in any endogenously stratified sample. This simple result is similar to the observation of Wooldridge (2000) and Hirano, Imbens, and Ridder (2003) who show that in stratified sampling the estimated or observed sample weights are preferred over the weight computed from the sampling probabilities that are used in the sampling. In the sequel we assume that the parameters in the conditional distribution of \( Y \) given \( X \) are
identified, preferably because this conditional distribution is non-parametrically identified.

We assume that sampling is in two stages (i) a stratum $S_k$ is selected with probability $H_k$, (ii) a random draw is obtained from $f(y, x|(Y, X) \in S_k)$ which we denote as $f(y, x|S_k)$. This is called multinomial sampling. In stratified sampling the number of observations in each stratum $S_k$ is fixed in advance. Imbens and Lancaster (1996) show that inferences for both sampling schemes are the same, because the associated likelihood functions are proportional. Let $S$ be the stratum indicator that is equal to $k$ if the observation is in $S_k$.

The joint density of $Y, X, S$ in the sample is

$$g(s, y, x) = H_s f(y, x|S_k) = H_s \frac{f(y|x; \theta) f(x)}{Q_s}$$  \hspace{1cm} (140)

where we implicitly assume that $Y$ is continuous. If not, just replace integration by summation. Now define

$$S_k(x) = \{y|(y, x) \in S_k\}$$

and

$$R(k, x, \theta) = \Pr((Y, X) \in S_k|X=x) = \Pr(Y \in S_k(x)|X=x) = \int_{S_k(x)} f(y|x; \theta) dy$$

Obviously $Q_k = E(R(k, X, \theta))$.

The marginal density of $X$ in the sample is obtained from (140) by integration with respect to $y$ over $S_k(x)$ (which may be an empty set for some $x$ and
and summation over \( k \)

\[
g(x) = f(x) \sum_{k=1}^{K} \frac{H_k}{Q_k} R(k, x, \theta)
\]

The sample density of \( X \) depends on the parameters \( \theta \). In endogenously stratified samples this distribution contains information on \( X \). The conditional density of \( S, Y \) given \( X \) in the sample is

\[
g(s, y|x) = \frac{f(y|x; \theta) \frac{H_s}{Q_s}}{\sum_{k=1}^{K} \frac{H_k}{Q_k} R(k, x, \theta)}
\]

(141)

An obvious method to obtain an efficient estimator of \( \theta \) is by maximizing the likelihood function based on (140)

\[
\ln L(\theta) = \sum_{i=1}^{N} \ln g(s_i, y_i, x_i) = \sum_{i=1}^{N} \ln f(y_i|x_i; \theta) f(x_i) + \ln \left( \frac{H_{s_i}}{Q_{s_i}} \right)
\]

This likelihood requires the evaluation of \( Q_k \) that depends on \( \theta \) and also on the marginal population density of \( X, f(x) \). This is computationally unattractive, and worse it requires the specification of the density of the independent variables.

For that reason we consider an alternative method to obtain the MLE. This method consists of three steps. First, we assume that the distribution of \( X \) is discrete with \( L \) points of support, i.e.

\[
\Pr(X = x_l) = f(x_l) = \pi_l, \quad l = 1, \ldots, L
\]

Next, we reparameterize from the discrete distribution of \( X \) in the population \( \pi_l \) to its discrete distribution in the sample \( \lambda_l \). The stratum probabilities \( Q_k \) can also be expressed in \( \lambda_l \). After this reparametrization the log likelihood is the sum of the conditional loglikelihood and the marginal loglikelihood of the
observations on $X$. The first factor depends on $\lambda_l$ only through the stratum probabilities $Q_k$.

The third step is that if we maximize the conditional loglikelihood with respect to $H_1, \ldots, H_K$ and evaluate the first-order conditions at the MLE of these 'parameters', the restrictions on the stratum probabilities $Q_k$ are satisfied. Hence maximizing the conditional loglikelihood with respect to $\theta$ and $H_1, \ldots, H_K$ is equivalent to maximization of the sample loglikelihood with respect to $\theta$. This conclusion does not depend on the assumption that $X$ has a discrete distribution. Following Chamberlain (1987) we conclude that the CMLE is efficient. Note that this is true if we replace the multinomial sampling probabilities $H_k$ in the conditional loglikelihood by their sample values $N_k/N$ with $N_k$ the number of observations in stratum $k$. The CMLE is not efficient if we use the probabilities $H_k$ that were actually used in the multinomial sampling.

The discrete distribution of $X$ in the sample is

$$g(x_l) = \lambda_l = \pi_l \left[ \sum_{k=1}^K \frac{H_k}{Q_k} R(k, x_l, \theta) \right]$$

Hence

$$Q_k = \sum_{i=1}^L R(k, x_l, \theta) \pi_l = \sum_{i=1}^L \frac{R(k, x_l, \theta)}{\sum_{m=1}^K \frac{H_m}{Q_m} R(m, x_l, \theta)} \lambda_l$$

which can be written as a sample average

$$1 = \frac{1}{N} \sum_{i=1}^N \frac{R(k, x_l, \theta) \frac{1}{Q_k}}{\sum_{m=1}^K \frac{H_m}{Q_m} R(m, x_l, \theta)} \tag{142}$$

The conditional log likelihood is

$$\ln L_c(\theta) = \sum_{i=1}^N \ln f(y_i|x_i; \theta) - \sum_{i=1}^N \ln \left( \sum_{k=1}^K \frac{H_k}{Q_k} R(k, x_i, \theta) \right) + \sum_{k=1}^K N_k \ln \frac{H_k}{Q_k}$$
The first-order condition for $H_k$ is

$$\frac{N_k}{H_k} = \sum_{i=1}^{N} \frac{R(k, x_i, \theta)}{\sum_{m=1}^{K} \frac{\partial}{\partial m} R(m, x_i, \theta)}$$

If we substitute the MLE $\hat{H}_k = \frac{N_k}{N}$ in this equation and in (142) we see that they are identical and we conclude that the restrictions for $Q_k$ are satisfied at the MLE (but not if we substitute $H_k$).

Note that if (142) holds for all $k = 1, \ldots, K$, multiplication by $H_k$ and summation over $k$ gives that $\sum_{i=1}^{L} \lambda_i = 1$. Again this condition is satisfied if the first-order conditions for maximization of the conditional loglikelihood with respect to $H_1, \ldots, H_K$ are evaluated at the MLE of these 'parameters'.

Hence the efficient estimator of $\theta$ is found by maximizing the conditional loglikelihood with respect to $\theta$ and $H_1, \ldots, H_K$. The first order conditions are evaluated at the MLE of $H_1, \ldots, H_K$ and solved for $\theta$ and $Q_1, \ldots, Q_K$. These first-order conditions set the sample average of the following functions equal to 0

$$m_1(s, y, x; \theta, Q) = \frac{\partial}{\partial \theta} f(y|x; \theta) - \frac{\sum_{k=1}^{K} \hat{H}_k \frac{\partial}{\partial k} R(k, x, \theta)}{\sum_{k=1}^{K} \hat{H}_k R(k, x, \theta)}$$

$$m_2(s, y, x; \theta, Q) = Q_k - \frac{R(k, x, \theta)}{\sum_{m=1}^{K} R(m, x, \theta)}$$

for $k = 1, \ldots, K$ and $\hat{H}_k = \frac{N_k}{N}$. Hence the efficient estimator is a GMM estimator that satisfies moment conditions based on these moment functions. An additional moment function that gives $\hat{H}_k$ can be added, but the corresponding moment condition is independent of the other moment conditions. Hence we can treat the $\hat{H}_k$ as given.

The variance of the efficient estimator can be found by the usual GMM estimation technique.
formula. The GMM formulation is convenient if we add additional sample information. This is just another moment condition.

3.3.3 Efficient parametric estimation with marginal information

Random sample with marginal information  First we consider the case that a random sample $Y_i, X_i, i = 1, \ldots, N$ is combined with marginal information. The marginal information consists of moments $E(h(Y, X)) = \overline{h}$ with $h$ a known function of dimension $K$ and $\overline{h}$ an $K$ vector of constants. The expectation is over the population distribution of $X, Y$. Hence we combine information in two random samples, one of which comprises of the whole population. Although these random samples cannot be independent, we can think of this as the combination of a relatively small random sample with a very large one. The sampling variance in the second sample is negligible. This is the setup considered by Imbens and Lancaster (1994).

Without loss of generality we set $\overline{h}$ equal to 0. The goal is to estimate the parameter vector $\theta$ in the conditional distribution of $Y$ given $X$ with conditional density $f(y|x; \theta)$. Because we have a random sample identification is not an issue. However, the additional moments overidentify the parameters, and these additional moment restrictions increase the precision of the estimation or can be used to create more powerful specification tests.

The score vector is

$$m_1(y, x; \theta) = \frac{\partial \ln f(y|x; \theta)}{\partial \theta}$$ (145)

Of course setting the sample average of the score equal to 0 gives the MLE
that is an efficient estimator without additional information. The additional information can be expressed as

\[ E(h(Y, X)) = \int \int h(y, x) f(y|x; \theta) dy g(x) dx = 0 \quad (146) \]

This gives a restriction on \( \theta \). The efficient estimator that uses this restriction is the restricted MLE that is obtained by maximizing the log likelihood subject to the constraint in (146).

The implementation of the restricted MLE requires the specification of the marginal density of \( X \). Applied researchers are usually unwilling to make parametric assumptions on this marginal distribution, and for that reason it is convenient that such a specification is not needed. Rewrite (146) as an average over the sample

\[ \frac{1}{N} \sum_{i=1}^{N} \int h(y, X_i) f(y|X_i; \theta) dy = \frac{1}{N} \sum_{i=1}^{N} m_2(Y_i, X_i; \theta) \quad (147) \]

Imbens and Lancaster (1994) show that the optimal GMM estimator with weight matrix equal to the inverse of the variance matrix of the moment restrictions has an asymptotic variance that is equal to that of the restricted MLE\textsuperscript{22}.

Their simulation study and empirical example show that the efficiency gains can be substantial. The precision of the estimator of the regression coefficient of \( X_j \) increases if the marginal information has a cross tabulation of grouped \( Y \) versus grouped \( X_k \). For instance, if \( Y \) is the employment indicator and \( X_j \) is age, a cross tabulation of employment status by age category (but no other variable) is highly informative on the age coefficient in an employment probit or

\textsuperscript{22}An alternative definition is the restricted MLE with (eq103c) as the restriction.
logit. If the model has no interactions the pairwise population cross tabulations of the dependent against the independent variables reduces the variances of the regression coefficients. Also in the case of a dummy dependent variable the marginal information is very useful if one of the outcomes is rare.

The additional moments (147) involve an integral over \( y \) (if \( Y \) is continuous). If one wants to avoid this integral one would be tempted to use the additional moment

\[
\frac{1}{N} \sum_{i=1}^{N} h(Y_i, X_i) = \frac{1}{N} \sum_{i=1}^{N} m_3(Y_i, X_i; \theta)
\]

(148)

instead of (147). The resulting GMM estimator is less efficient than the restricted MLE. This can be seen if one considers the case without covariates \( X \) and a scalar \( h \) and \( \theta \). In that case the moment condition in (147) restricts the parameter to its population value, while the moment condition in (148) does not remove the sampling variation in the restricted MLE. To achieve efficiency one should use (147) as the second set of moment conditions.

In the case that the conditional density is not specified, the moment conditions in (147) are not available and one if forced to use (148) together with the moment conditions based on \( m(y, x; \theta) \) that is a vector of moment conditions that identifies \( \theta \) and could be used to estimate the parameters if one only had the random sample from the population. The moment conditions (147) do not depend on \( \theta \), but because they are correlated with the moment conditions in (147). Hence imposing them along with (147) improves the precision of the estimators.

As noted the additional moments can be used for an often powerful test of
the parametric model $f(y|x; \theta)$. The obvious test is the GMM overidentification test based on the moment conditions (145) and (147). The test statistic is the minimal value of the optimal GMM minimand and it has under the null hypothesis of correct specification, a chi-squared distribution with $K$ (dimension of $h$) degrees of freedom. It should be noted that the test also rejects if the random sample is not from the same population that is used to compute $E(h(Y,X))$. To deal with this one could consider a joint test based on the moment conditions (145), (147) and (148) that tests both for the compatibility of the information and the specification. This test statistic has $2K$ degrees of freedom.

**Biased samples with marginal information**  Imbens and Hellerstein (1999) show\textsuperscript{23} that the optimal GMM estimator, based on (145) and (148), i.e. we consider the case that the conditional density of $Y$ given $X$ is not specified, but $\theta$ is estimated from a set of moment conditions, is equivalent to a weighted GMM estimator that solves

$$\sum_{i=1}^{N} w_i m_1(Y_i, X_i; \theta) = 0$$

(149)

with weights $w_i, i = 1, \ldots, N$ defined as the solution to

$$\max \sum_{i=1}^{N} \ln w_i \quad \text{s.t.} \quad \sum_{i=1}^{N} w_i = 1 \quad \sum_{i=1}^{N} w_i h(Y_i, X_i)$$

(150)

The weights are equal to

$$w_i = w(Y_i, X_i) = \frac{1}{N(1 + \lambda h(Y_i, X_i))}$$

(151)

\textsuperscript{23}To be precise, they only consider linear regression with additional moment restrictions, but their argument applies generally.
with \( \hat{\lambda} \) the Lagrange multiplier on the second restriction. It is the solution to

\[
\frac{1}{N} \sum_{i=1}^{N} \frac{h(Y_i, X_i)}{1 + \hat{\lambda} h(Y_i, X_i)} = 0 \quad (152)
\]

Now consider the case that a biased sample is combined with marginal information from the population. As an illustration we consider the example of a 0-1 dependent variable with conditional density

\[
f(y|x; \theta) = G(x'\theta)^y(1 - G(x'\theta))^{1-y}.
\]

The endogenously stratified sample has strata \( S_1 = 1 \times X' \) and \( S_2 = 0 \times X' \) with \( X' \) the support of \( X \). The multinomial sampling probabilities are \( H_1, H_2 \) and the population fractions of the two strata are \( Q_1, Q_2 \). Also \( h(y, x) = y - Q_1 \). In large samples \( \hat{\lambda} \) in (152) converges to the solution to the equation that is obtained by replacing the sample average in (152) by the corresponding expectation over the sample distribution

\[
\int \sum_{y=0}^{1} \frac{y - Q_1}{1 + \lambda(y - Q_1)} \left( \frac{H_1}{Q_1} G(x'\theta) \right)^y \left( \frac{H_2}{Q_2} (1 - G(x'\theta)) \right)^{1-y} g(x) dx = 0 \quad (153)
\]

The solution is

\[
\lambda = \frac{H_1 - Q_1}{Q_1 Q_2} \quad (154)
\]

so that the weights that depend on the value of \( y \) only are

\[
w(y, x) = \frac{1}{N} \frac{1}{1 + \frac{H_1 - Q_1}{Q_1 Q_2} (y - Q_1)} = \frac{1}{N} \left( \frac{Q_1}{H_1} \right)^y \left( \frac{Q_2}{H_2} \right)^{1-y} \quad (155)
\]

These weights are used in the score based on the full sample to obtain the weighted likelihood equation

\[
\sum_{i=1}^{N} w(Y_i, X_i) \left( Y_i \frac{\partial \ln G(X_i'\theta)}{\partial \theta} + (1 - Y_i) \frac{\partial \ln(1 - G(X_i'\theta))}{\partial \theta} \right) = 0 \quad (156)
\]

This corresponds to the Weighted Exogenous Sampling MLE of Manski and
Lerman (1977). This estimator is not fully efficient because it does not use the parametric model in the additional moment condition.

We conclude that if the additional population moments combined with the biased sample identify the population parameters, then the weighted estimator proposed by Imbens and Hellerstein (1999) is the efficient GMM that imposes the population moments. If the conditional density is specified, the estimator is not fully efficient. Hence their weighted estimator provides a constructive method to combine biased samples with population moments.

If the combination of the biased sample(s) and the population moment does not identify the population parameters, the weighted GMM estimator converges to the solution of

$$
\int \int m_1(y, x; \theta) \frac{f_s(y, x)}{1 + \lambda' h(y, x)} dy dx
$$

with \( \lambda \) the solution of (152) if we replace the (biased) sample average by the (biased) sample expected value. Hence the GMM estimator is consistent for the parameters in a distribution that satisfies the population moments and is also consistent with the biased sample. It is obtained from the distribution in the biased sample by weighting, which is the general approach (see section 3.2). The weights reproduce the population distribution if the parameters are identified. If not then they produce a GMM estimate that is consistent with the available information. However, in that case the weight (and hence the GMM estimator) are not unique. In the optimization problem (150) we can replace \( \ln w_i \) by \( K(p_i) \) with \( K \) any concave function. This reflects the fact that the parameters are not point identified.
4 Repeated cross sections

4.1 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 available 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 available longitudinal panels, 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 dynamic 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 throughout). 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, and thus rule out in- and out-migration, births, and mortality.

At issue is what distributions can be identified from the set of cross sections and what parameters in what types of models can be identified. 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 dynamic models, i.e., 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 and restrictions elucidated in Section 2.4 above, understanding 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 2.4.3 above:

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

and with associated GMM-IV moment condition, following on (72), of:

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

111
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\textsuperscript{24}. In most applications, $f(Y_t; \theta) = Y_t$ and $g_1$ and $g_2$ represent the conditional expectations of $Y_t$. The function $g_1$ contains only $X_t$ and $Z_0$ and hence appears to be estimable without matching across cross-sections but, as will be shown below, is problematic in fixed effects models because $X_t$ is endogenous 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 are matched by variables $Z_0$ and $Z_{1t}$, with the latter constituting the exclusion restriction. In most applications to date, $Z_{1t} = t$. Matching on $t$ is possible under the closed population assumption that the population at each $t$ is the same\textsuperscript{25}. The critical exclusion restriction in all RCS models is that $Z_{1t}$ and its interactions with $Z_0$ not appear arbitrarily in $g_1$ and $g_2$ and yet are correlated with those functions. For the $Z_{1t} = t$ case, this implies that individual age, year, unemployment duration, or firm lifetimes (depending on the application) cannot be entered nonparametrically in $g_1$ and $g_2$. Although restrictions on the functional forms in which $t$ appears in $h$, $g_1$, and $g_2$ are possible, we do not consider those here. Thus the essential restriction in RCS

\textsuperscript{24}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.

\textsuperscript{25}However, it is possible that some history information is available in each cross-section which allows matching on time-varying variables (e.g., employment or marital status histories in the case of household survey data; ages of children are another).
estimation is that some 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 estimation can be conducted by OLS of $Y_t$ on $E(X_t|h(Z_0, Z_{1t}))$, $E(Y_{t-1}|h(Z_0, Z_{1t}))$, and $Z_0$, where the two expectations are replaced by estimates obtained by first-stage projections. If there are no $Z_0$ in the data, $Z_{1t} = t$, and $h(t)$ 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 $Y_{t-1}$. Most interesting cases arise instead 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 $h$ is assumed to be a vector of indicators for all combinations of $Z_0$ values and discrete values of $t$, estimation using (159) is equivalent to a regression of the cell means of $Y_t$ on the cell means of $X_t$, $Y_{t-1}$, and dummy variables for $Z_0$. 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 aggregation approach is not possible in the first place.

Two leading examples fit into this framework. One is the linear first-order

\[^{26}\text{Projections onto } Z_0 \text{ and } Z_{1t} \text{ directly are an alternative}\]
autoregression

\[ Y_t = \alpha + \beta Y_{t-1} + \gamma X_t + \delta Z_0 + \epsilon_t \quad (160) \]

If \( Z_{1t} = t \) then the exclusion restriction requires that the parameters of (160) not be arbitrarily time-varying. The restriction that the instrument must be relevant implies that the mean of \( E(Y_{t-1}|Z_0,t) \) must vary with \( t \). Thus a purely stationary process is unlikely to be estimable. If \( Y_{t-1} \) is correlated with \( \epsilon_t \), an instrument \( Z_{1t} \) must be found which is orthogonal to \( \epsilon_t \).

A second example is the linear fixed effects model:

\[ Y_t = \gamma X_t + \delta Z_0 + f + \epsilon_t \quad (161) \]

where \( f \) is an individual fixed 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 (159) proceeds by using the elements of \( h \) as instruments for \( X_t \), which again requires some minimal time-invariance of the parameters of (161). 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 (161) by the aggregation method mentioned previously was proposed by Deaton (1985). 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 (161) 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).

### 4.2 Consistency and Related Issues

The conditions for consistency of moment estimators in the form (159) are well-known in general (Hansen, 1982). The special form they take in the two sample case were considered in Section 2.4 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 for the asymptotic orthogonality of the instruments

$$\frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} h(Z_{0i}, Z_{1it}) \varepsilon_{it} \xrightarrow{P} 0 \quad (162)$$

where $\varepsilon_{it} = f(Y_{it}; \theta) - g_1(X_{it}, Z_{0i}; \theta) - g_2(Y_{i,t-1}, Z_{0i}; \theta)$. Asymptotics are taken w.r.t. $N$ with fixed $T$, as conventional in panel data applications where $N$ is generally large and $T$ is not. For fixed effects and other models with time-invariant individual-specific components of $e_t$, (162) is fulfilled so long as those components converge to a fixed, time invariant value which is the same at all $t$ and which will necessarily be orthogonal to the (time-varying) $Z_{1it}^{27}$. Condition (162) is also required given the presence of $Y_{t-1}$ in the equation and represents an IV solution familiar to panel data models with dynamics and lagged endoge-

---

27The time-invariant vector $Z_0$ is rather different, and for (162) to hold for those variables requires that the distribution of the fixed effect be the same for all $Z_0$. The parameters governing the effect of the (time invariant) $Z_0$ are identified because individual fixed effects are not being controlled for.
nous variables. However, with a lagged dependent variable in the equation the errors in successive periods have a MA(1) covariance because the errors in not observing the same individuals in each cross section are correlated (McKenzie, forthcoming).

The case of $T \to \infty$ conditional on fixed $N$ is more problematic if there are time-invariant individual specific components in the error term like fixed effects. If $f(t)$ is the mean fixed effect in cross section $t$ then the error $[f(t) - f^*]$, where $f^*$ is asymptotic value as $N \to \infty$, does not disappear under a $T$ asymptotic. Consequently there is less confidence that (162) will hold. Solutions to this problem have not been addressed in the general GMM-IV context we are considering here, although they have been considered in a different context described below.

There is a third asymptotic that can be considered as well, which is an asymptotic in the number of cohorts (Deaton, 1985; Verbeek, 1996). Up to this 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 $N$ and $T$, the error in estimating fixed effects will not disappear and, indeed, asymptotically the number of fixed effects grows. Once again, therefore, condition (162) is unlikely to hold.

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

$$Y_{ct} = \gamma X_{ct} + \delta Z_0 + \epsilon_{ct}$$

(163)

where means are taking over observations within each cohort ($c$) and year ($t$) cell ($Z_0$ represents a set of cohort dummies and represents its own mean) yields biased estimates for finite $N$ because the $f_{ct}$ which are the means of the within-cohort deviations in fixed effects, do not disappear because of the fluctuations in each cross-section induced by random sampling. Put differently, the cohort dummies which are aimed to estimate mean fixed effects in each cohort are estimated only over means for the $N$ sampled individuals in each cross-section, which fluctuates with $t$. These $f_{ct}$ fluctuations will be is correlated with $X_{ct}$. Deaton instead considers the "population" equation

$$Y^*_{ct} = \gamma X^*_{ct} + \delta Z_0 + f^*_c + \epsilon^*_ct$$

(164)

where variables with asterisks represent population values. Estimation of 164 has no error in the estimation of mean cohort fixed effects but does have a measurement error problem because $X_{ct}$ and $Y_{ct}$ 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. His proposed estimator is

\[
(\hat{\gamma} \quad \hat{\delta})' = (M - \Sigma)^{-1}(m - \sigma)
\]

where \(M\) is the moment matrix of the regressors (including \(Z_0\), which has no error), \(m\) is the moment matrix between the regressors and \(Y\), and \(\Sigma\) and \(\sigma\) are the estimated within-cell covariance matrices of the regressors and of the regressors and \(Y\), respectively.

As \(N \to \infty\) the measurement errors go to zero and the least squares estimate of the aggregate model is consistent. Deaton noted that the estimator is 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 measurement error adjustment. 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" measurement error in (165). Devereaux (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. Devereaux also proposes a modification of the Deaton estimator which is approximately unbiased but has a smaller finite sample variance.

Deaton does not set up his model from the IV framework used here but it can be done so. To focus on the key issues, assume that only one cohort of \(N\) individuals is observed for \(T\) periods and that there is only one regressor. The
individual model is

\[ y_{it} = \beta x_{it} + f_{it} + \varepsilon_{it} \]  

(166)

where individual subscripts \( i = 1, \ldots, N \) are now added\(^{28}\). The IV equation when using dummies for \( t = l, \ldots, T \) as instruments is

\[ \bar{y}_t = \beta \bar{x}_t + \bar{f}_t + \bar{\varepsilon}_t \]  

(167)

Consequently,

\[ \text{Cov}(\bar{y}_t, \bar{x}_t) = \beta \text{Var}(\bar{x}_t) + \text{Cov}(\bar{f}_t, \bar{x}_t) \]  

(168)

Representing the correlation of the regressor and the fixed effect by \( x_{it} = \psi + \phi f_{it} + \omega_{it} \) and assuming, for illustration, that \( f_{it} \sim N(0, \sigma^2) \), the bias term in (168) is

\[ \text{Cov}(\bar{f}_t, \bar{x}_t) = \frac{\phi \sigma^2}{N} \]  

(169)

This bias term is small if \( N \) is large, if the fixed effects are unimportant (\( \sigma^2 \) small or zero), or the correlation between the regressor and the fixed 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, \( \text{Cov}(f_{it}, x_{it}) = \text{Cov}(y_{it}, x_{it}) - \beta \text{Var}(x_{it}) \) within cells of \( t \). Hence \( \text{Cov}(\bar{f}_t, \bar{x}_t) = \hat{\sigma}_{yx} - \beta \hat{\sigma}_x^2 \) where \( \hat{\sigma}_{yx} \) and \( \hat{\sigma}_x^2 \) are estimated from the individual within-cell observations. Inserting this into (168) and solving for \( \beta \), we obtain the Deaton estimator

\[ \hat{\beta}_D = \frac{\text{Cov}(\bar{y}_t, \bar{x}_t) - \hat{\sigma}_{yx}}{\text{Var}(\bar{x}_t) - \hat{\sigma}_x^2} \]  

(170)

\(^{28}\)We leave it to be understood that individual \( i \) at time \( t \) is not the same individual \( i \) at time \( t' \).
This estimator specializes to the IV estimator \( \hat{\beta}_{IV} = \frac{\text{Cov}(y_t, x_t)}{\text{Var}(x_t)} \) when it and the within-cell estimator \( \hat{\beta}_W = \frac{\hat{\sigma}_{yx}}{\hat{\sigma}_{x}^2} \) are the same. The two estimators are the same precisely when \( \phi^2 = \sigma^2 = 0 \) and hence the bias in (169) is zero. Consequently, the Deaton adjustment in (170) is an indirect adjustment for the size of the bias from the presence of the fixed effect.

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 Devereaux (2003) suggests that should be higher, possibly 2000 or more. The necessary \( N \) is sensitive to the specification of the model. Devereaux 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 within 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

\[ Y_t = \alpha + \beta Y_{t-1} + \delta Z_0 + f + \varepsilon_t \]  (171)

which is a combination of (160) and (161) (individual subscripts are again omitted). 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 Honor, 2001). Using simple functions oft 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 (forthcoming) 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. \( T \).

### 4.3 Binary Choice Models

In the binary choice model we return to (158) 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

\[
Y_t = F\left(g_1(X_t, Z_0; \theta) + g_2(Y_{t-1}, Z_0; \theta)\right) + \nu_t \tag{172}
\]

which does not fit into the framework of the moment condition in (160) because \( X_t \) and \( Y_{t-1} \) are not separable. Let us therefore initially assume \( g_2 = 0 \) and consider lagged indicators below. Now (160) applies directly assuming the availability of a suitable exclusion restriction, as before. The moment conditions are a simple extension of those shown in eqns (77)-(79). The method is applicable to the fixed effects binary choice model or to any binary choice model with endogenous \( X_t \). However, in parametric estimation where the \( F \) distribution is assumed to be known, a distributional assumption is needed for the fixed effect in order to derive \( F \), e.g., if \( f_t \) is the fixed effect component of \( \varepsilon_t \),

\[
f_t = v(Z_0; \phi) + \eta_t \tag{173}
\]

where \( v \) is assumed to be of known form and where \( \eta_t \) has a known parametric distribution from which the c.d.f. of the composite error \( \varepsilon_t \) can be derived.

If the instrument is a set of dummies for \( t \), possibly interacted with \( Z_0 \), the nonlinearity of the conditional expectation function makes this approach
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 (173), is

\[
X_t = w(Z_0, Z_{1t}; \psi) + \omega_t
\]  

(174)

where \( w \) is a function of known parametric form and \( \omega_t \) is an error term with a parametric distributional form. 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; Blundell and Smith, 1986; see Blundell and Smith, 1993 for a review). Options include replacing \( X_t \) in \( g_1 \) with its predicted value from (174); inserting an estimated residual from (174) into (172); and estimating (153) and (155) in reduced from by inserting (174) into (172). In this approach, the parameters of (172) are estimated by maximum likelihood, which implies that the weighting vector \( h \) in (159) is the binary choice weighting matrix

\[
\frac{F'}{(1-F)F'}
\]

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

\[
E(Y_t|Z_0, Y_{t-1}) = F(g_2(Y_{t-1}, Z_0; \theta))
\]  

(175)

where we have assumed that \( \varepsilon_t \) is distributed independently of \( Y_{t-1} \) (i.e., no serial correlation). Instrumental variable estimation of (175) 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

\[
E(Y_t|Z_0) = p_t(Z_0) =
\]

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

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

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

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

(176)

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 (176) is a familiar flow identity from renewal theory showing how the marginal probability at \( t - 1 \) is transformed by the two hazard rates 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 hazards 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 hazard rate functions is achieved by restricting their temporal dependence (indeed, in (176) they are assumed to be time invariant); identification is lost if the two hazards 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,

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

which can be jointly estimated with (176) imposing the commonality of the functions\(^{29}\). Alternatively, (176) can be expressed in fully solved back form and estimated as well.

Equation (176) has been used as the basis of RCS estimation at the aggregate level. Miller (1952) considered estimation of (176) with time-series data on the proportions of a variable, \( p_t \) which is special case of RCS data. Without data

\(^{29}\)Alternatively an initial conditions can be specified as a marginal \( p \) in the first period.
on individual regressors $Z_0$, he suggested simple least squares estimation of

$$p_t = \mu + \eta p_{t-1} = \nu_t$$  \hfill (178)

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 et al. (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 serial correlation of the errors, the two hazards require knowledge of the functional dependence of $\varepsilon_t$ on $Y_{t-1}$. The most straightforward approach would require replacing the simple hazards 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 likehood 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 (176) is problematic because

$$E(Y_t|X_t, Z_0) = \mu(X_t, Z_0; \theta) + \eta(X_t, Z_0; \theta)p_{t-1}(X_t, Z_0)$$ \hfill (179)

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 identi-
fied 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.

4.4 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 was 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 et al (1994), who estimated Euler equations under uncertainty for aggregate cohort profiles of consumption, applying 16 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, and can be estimated as well if training status is known.

There have been a few applications of the Markov model described above. Pelzer et al. (2002a, 2002b) 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

17 but 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 identifica-

tion.

Gell 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 unemploy-
ment have changed with calendar time and what that implies for the distribution
of unemployment between short term and long term. A simpler but similar ex-
cercise 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


[111] Vardi, Y. "Nonparametric estimation in the presence of length bias." *An-


Appendix

Theorem 1

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

Proof

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

$$m_N(\theta)'W_Nm_N(\theta) = (m_N(\theta) - m_N(\theta_0))'W_N(m_N(\theta) - m_N(\theta_0)) +$$

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

By the mean value theorem

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

with $\theta_*$ between $\theta$ and $\theta_0$. Substitution in (180) and taking the limit $N_1, N_2 \to \infty$ gives

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

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

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 $E\left[\frac{\partial m}{\partial \theta}(\theta)\right]$ 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

\[
\sqrt{N} (\hat{\theta}_N - \theta_0) \xrightarrow{d} N(0, V(\theta_0)) \tag{183}
\]

with

\[
V(\theta_0) = \left[ E \left( \frac{\partial m'}{\partial \theta} (\theta_0) \right) W(\theta_0) E \left( \frac{\partial m}{\partial \theta'} (\theta_0) \right) \right]^{-1}. \tag{184}
\]

\[
\cdot E \left( \frac{\partial m'}{\partial \theta} (\theta_0) \right) W(\theta_0) \left( \lambda \text{Var}(m_{1j}(\theta_0)) + \text{Var}(m_{2i}(\theta_0)) \right) W(\theta_0) E \left( \frac{\partial m}{\partial \theta'} (\theta_0) \right).
\]

\[
\cdot \left[ E \left( \frac{\partial m'}{\partial \theta} (\theta_0) \right) W(\theta_0) E \left( \frac{\partial m}{\partial \theta'} (\theta_0) \right) \right]^{-1}
\]

**Proof**

The first order conditions give

\[
0 = \frac{\partial m'}{\partial \theta} (\hat{\theta}_N) W_N \sqrt{N} m_N(\hat{\theta}_N) \tag{185}
\]

By the mean value theorem we have for some \( \bar{\theta}_N \) between \( \theta_0 \) and \( \hat{\theta}_N \)

\[
\sqrt{N} m_N(\hat{\theta}_N) = \sqrt{N} m_N(\theta_0) + \frac{m_N}{\partial \theta} (\bar{\theta}_N) \sqrt{N} \hat{\theta}_N - \theta_0 \tag{186}
\]

Substitution in (185) and solving for \( \sqrt{N} (\hat{\theta}_N - \theta_0) \) gives

\[
\sqrt{N} (\hat{\theta}_N - \theta_0) = - \left[ \frac{\partial m'}{\partial \theta} (\hat{\theta}_N) W_N \frac{m_N}{\partial \theta'} (\bar{\theta}_N) \right]^{-1} \frac{m_N}{\partial \theta} (\hat{\theta}_N) W_N \sqrt{N} m_N(\theta_0) \tag{187}
\]
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 $\sqrt{N}m_N(\theta_0)$.

**Theorem 3**

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

**Proof**

Substitution of (187) in (186) gives

$$\sqrt{N}m_N(\hat{\theta}_N) =$$

$$[I - \frac{\partial m_N}{\partial \theta'}(\hat{\theta}_N)\frac{\partial m_N}{\partial \theta'}(\hat{\theta}_N)^{-1} \frac{\partial m_N}{\partial \theta}(\hat{\theta}_N)W_N] \sqrt{N}m_N(\theta_0)$$

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

$$\sqrt{N}m_N(\hat{\theta}_N) = [I - A(\theta_0)'(A(\theta_0)WA(\theta_0)' - 1)A(\theta_0)W] \sqrt{N}m_N(\theta_0) + o_p(1)$$

Upon substitution of (189) in (97)

$$T_N = \sqrt{N}m_N(\theta_0)'[I - W'A(\theta_0)'(A(\theta_0)WA(\theta_0)' - 1)A(\theta_0)]W.$$

151
\[ \sqrt{N_{2m_N}(\theta_0)} = \sqrt{N_{2m_N}(\theta_0)} \left[ I - A(\theta_0)'(A(\theta_0)W \Lambda(\theta_0)'A(\theta_0)')^{-1}A(\theta_0)W \right] \sqrt{N_{2m_N}(\theta_0)} + o_p(1) = \]

\[ = \sqrt{N_{2m_N}(\theta_0)} \left[ W - W'A(\theta_0)'(A(\theta_0)W \Lambda(\theta_0)'A(\theta_0)')^{-1}A(\theta_0)W \right] \sqrt{N_{2m_N}(\theta_0)} + o_p(1) \]

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

\[ T_N = \sqrt{N_{2m_N}(\theta_0)}M(\theta_0)^{-\frac{1}{2}}. \] (191)

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

\[ . M(\theta_0)^{-\frac{1}{2}} \sqrt{N_{2m_N}(\theta_0)} + o_p(1) \]

Because \( \sqrt{N_{2m_N}(\theta_0)}M(\theta_0)^{-\frac{1}{2}} \overset{d}{\rightarrow} N(0, I) \) and the matrix between \([\cdot]\) is idempotent with rank equal to \( \dim(m_N) - \dim(\theta) \), the result follows.
Figure 2: Fréchet 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 $y < \min\{F^{-1}(p_k(z_l)), k = 1, 2, 3, l = 1, 2\}$. 