NEPS
National Educational Panel Study

NEPS Working Papers

Götz Rohwer

Using Sampling Weights for Model Estimation?

NEPS Working Paper No. 4

Bamberg, December 2011

SPONSORED BY THE
Federal Ministry
of Education
and Research

Working Papers of the German National Educational Panel Study (NEPS)
at the University of Bamberg

The NEPS Working Papers publish articles, expertises, and findings related to the German National Educational Panel Study (NEPS).
The NEPS Working Papers are edited by a board of researchers representing the wide range of disciplines covered by NEPS. The series started in 2011.

Papers appear in this series as work in progress and may also appear elsewhere. They often represent preliminary studies and are circulated to encourage discussion. Citation of such a paper should account for its provisional character.

Any opinions expressed in this series are those of the author(s) and not those of the NEPS consortium.

The NEPS Working Papers are available at
http://www.uni-bamberg.de/neps/publikationen/neps-working-papers/

Contact: German National Educational Panel Study (NEPS) − University of Bamberg − 96045 Bamberg − Germany − contact.neps@uni-bamberg.de

Using Sampling Weigths for Model Estimation?

Götz Rohwer, Ruhr-Universität Bochum

Dezember 2011

E-Mail-Adresse des Autors:
goetz.rohwer@rub.de

Bibliographische Angaben:

Rohwer, G. (2011). Using Sampling Weights for Model Estimation? (NEPS Working Paper No. 4). Bamberg: Otto-Friedrich-Universität, Nationales Bildungspanel.

Using Sampling Weights for Model Estimation?

Abstract
This paper discusses the question whether one should use (design-based) sampling weights when estimating statistical models. It is argued that the answer depends, in particular, on the kind of model to be estimated. The paper distinguishes three kinds. (1) Descriptive models that intend to provide simplified descriptions of the distribution of variables defined for a target population. It is argued that, except for some special situations, sampling weights should be taken into account when estimating such models. (2) Probabilistic data models which start from the idea that the data in a given sample can be viewed as realizations of random variables. It is argued that thinking about the usage of sampling weights in the estimation of such models depends on the understanding of the relationship between the model and the random variables serving to represent the given data. (3) Probabilistic functional models which intend to formulate rules for a generic unit defined without reference to any particular target population. It is argued that using sampling weights in the estimation of such models is required only if the selection probabilities used in the sampling procedure depend on endogenous variables of the model.

Keywords
sampling weights, stratified sampling, descriptive estimation, model estimation

1. Introduction

This paper discusses the question whether one should use sampling weights when estimating statistical models. It is argued that the answer depends, in particular, on the kind of model to be estimated. I distinguish three kinds:

- *Descriptive models* that intend to provide simplified descriptions of the distribution of variables defined for a target population. I argue that, except for some special situations, sampling weights should be taken into account when estimating such models.
- *Probabilistic data models* which start from the idea that the data in a given sample can be viewed as realizations of random variables. I argue that thinking about the usage of sampling weights in the estimation of such models depends on the understanding of the relationship between the model and the random variables serving to represent the given data.
- *Probabilistic functional models* which intend to formulate rules for a generic unit defined without reference to any particular target population. I argue that using sampling weights in the estimation of such models is required only if the selection probabilities used in the sampling procedure depend on endogenous variables of the model. I further suggest to rethink, and possibly reformulate, the model if weighted and unweighted estimates differ significantly.

I consider only sampling weights that can be derived from a stratified sampling design. In particular, I do not discuss the usage of weights intended to compensate for unequal response rates. A further limitation is that I only discuss models for cross-sectional data.

2. Descriptive Models

I consider two kinds of descriptive models: Models intended to represent distributions of statistical variables in a target population, and descriptive regression models intended to describe dependency relations between statistical variables in a target population. I begin with briefly explaining my understanding of 'descriptive estimation'.

2.1 Descriptive Estimation

The conceptual framework is given by a *statistical variable*

$$X : \Omega \longrightarrow \mathcal{X}$$

which is defined for a target population Ω consisting of a finite number of units. To each unit ω, the variable X assigns an element $X(\omega)$ of the variable's property space $\mathcal{X}$. X could consist of several components: $X = (X_1, \ldots, X_q)$ with a property space $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_q$. In any case, one can refer to the variable's distribution by quantities $\mathrm{P}(X = x)$ defined as the proportion of units in Ω whose value of X is x (a specified element of $\mathcal{X}$).[1] *Descriptive estimation* can then be understood as intending to estimate the distribution $\mathrm{P}(X = x)$, or quantities derived from this distribution (e.g. the mean of X), based on knowing the values of X for the members of a random sample $\mathcal{S} \subset \Omega$.

Knowing the values of X for the members of a sample $\mathcal{S}$ entails that one can refer to a statistical variable

$$X^s : \mathcal{S} \longrightarrow \mathcal{X}$$

[1] Analogously, one can use expressions like $\mathrm{P}(X \leq x)$ and $\mathrm{P}(X \in M)$ where M is a subset of $\mathcal{X}$.

with $X^s(\omega) = X(\omega)$ for all $\omega \in \mathcal{S}$; and one can calculate the quantities $\mathrm{P}(X^s = x)$ describing the distribution of X^s in the sample. How these quantities can be used to estimate $\mathrm{P}(X = x)$ depends on the sampling procedure. As a starting point, I take it that $\mathrm{P}(X^s = x)$ provides a *plausible estimate* of $\mathrm{P}(X = x)$ if $\mathcal{S}$ is a simple random sample defined by equal selection probabilities for all members of the population. (I discuss below in which sense one can also speak of an 'unbiased estimate'.)

In practice, sampling procedures often entail unequal selection probabilities. As a general framework I consider stratified sampling based on information about a stratification variable

$$H : \Omega \longrightarrow \{1, \ldots, m\}$$

Knowing this variable, one can define m subpopulations (strata), $\Omega_j = \{\omega \,|\, H(\omega) = j\}$, and one also knows their sizes, N_j, such that $\Sigma_j N_j = N$, the size of the target population Ω.

Given this framework, one can define several different kinds of stratified sampling procedures. In the following I consider just one form: From each subpopulation Ω_j one takes a simple random sample, $\mathcal{S}_j$, having a predefined size n_j. The overall sample is then defined as the union $\mathcal{S} = \mathcal{S}_1 \cup \cdots \cup \mathcal{S}_m$ having the size $n = n_1 + \cdots + n_m$.

If all sampling fractions n_j/N_j equal n/N, the overall sample $\mathcal{S}$ is a simple random sample from the target population and (by definition) provides plausible estimates. If, however, the sampling fractions differ across the strata, one should use sampling weights. This can be seen as follows. One intends to estimate

$$\mathrm{P}(X = x) = \sum\nolimits_j \mathrm{P}(X = x \,|\, H = j)\, \frac{N_j}{N} \tag{1}$$

Then, given that[2]

$$\mathrm{P}(X^s = x \,|\, H = j) = \frac{1}{n_j} \sum\nolimits_{\omega \in \mathcal{S}_j} I[X = x](\omega)$$

is a plausible estimate of $\mathrm{P}(X = x \,|\, H = j)$, and defining weights

$$w_\omega = \frac{N_j}{n_j\, N} \quad \text{for } \omega \in \mathcal{S}_j \tag{2}$$

(entailing the normalization $\Sigma_{\omega \in \mathcal{S}}\, w_\omega = 1$),

$$\sum\nolimits_{\omega \in \mathcal{S}} w_\omega\, I[X = x](\omega) \tag{3}$$

is a plausible estimate of $\mathrm{P}(X = x)$.

Example

A simple example will be used for illustration. The target population consists of $N = 100000$ school-children having completed a specified grade. There are three variables: X records the school type (0 or 1), Y records whether the grade was completed successfully (1) or not (0), and Z records the parents' educational level (0 = low or 1 = high). Table 1 shows the distribution of these variables in the population and in a stratified sample.

The construction of the sample uses X (school type) as a stratification variable. The two subpopulations are Ω_1 consisting of $N_1 = 70000$ school-children with $X = 0$, and Ω_2 consisting of

[2] I use $I[X = x]$ as an indicator variable defined for Ω: $I[X = x](\omega) = 1$ if $X(\omega) = x$, and $I[X = x](\omega) = 0$ otherwise.

Table 1 Distribution of (X, Z, Y) in the population and in a stratified sample.

X	Z	Y	Population	Sample
0	0	0	20000	100
0	0	1	30000	150
0	1	0	4000	20
0	1	1	16000	80
1	0	0	2400	24
1	0	1	5600	56
1	1	0	2200	22
1	1	1	19800	198

$N_2 = 30000$ school-children with $X = 1$. The sampling fractions are, respectively, 0.5 % and 1 %, so that $n_1 = 350$ and $n_2 = 300$, and the overall sample size is $n = 650$. For ease of presentation, and since we are not concerned with sampling errors, it is assumed that the variables' distribution in the subsamples equals their distribution in the corresponding subpopulations.

Now assume that we want to estimate $\mathrm{P}(Y = 1) = 0.714$. Using the sample without weights would result in a distorted estimate: $484/650 = 0.745$. On the other hand, using the weights $w_\omega = N_1/(n_1 N) = 0.002$ for $\omega \in \mathcal{S}_1$, and $w_\omega = N_2/(n_2 N) = 0.001$ for $\omega \in \mathcal{S}_2$, would give the plausible estimate $\sum_{\omega\in\mathcal{S}} w_\omega\, I[Y = 1](\omega) = 0.714$.

Referring to a sampling design

So far the discussion was in terms of 'plausible estimates' based on a given simple or stratified random sample $\mathcal{S}$ from a target population Ω. In order to introduce estimators, the reference to a sampling design is required. A *sampling design* describes a method that can be used to create random samples from a target population and specifies a probability distribution for the set of possible samples. I assume that all possible samples have the same size, n, that is fixed in advance.

Reference to a sampling design allows one to define, for each unit $\omega \in \Omega$, an *inclusion variable*, that is, a random variable indicating whether the unit is included in a randomly generated sample:

$$\dot{I}_\omega(\mathcal{S}) = \begin{cases} 1 \text{ if } \omega \in \mathcal{S} \\ 0 \text{ otherwise} \end{cases} \tag{4}$$

To remind that this is a random variable (defined by the sampling design), and not a statistical variable, it is marked by a dot. In addition one can define *inclusion probabilities*

$$\pi(\omega) = \sum\nolimits_{\mathcal{S}} \dot{I}_\omega(\mathcal{S}) \Pr(\mathcal{S}) \tag{5}$$

to be interpreted as the probability of generating a sample that includes the unit ω.

These notions can now be used to define estimators. For example, an estimator for $\mathrm{P}(X = x)$ can be defined as

$$\dot{\mathrm{P}}[X = x](\mathcal{S}) = \sum\nolimits_{\omega\in\Omega} w_\omega\, I[X = x](\omega)\, \dot{I}_\omega(\mathcal{S}) \tag{6}$$

with weights defined by $w_\omega = 1/(\pi(\omega)N)$. Given the values of X for the units in a sample $\mathcal{S}$, one can use this estimator to calculate a specific estimate of $\mathrm{P}(X = x)$. The weights are chosen as to make the estimator *unbiased*, meaning that its expectation (defined w.r.t. the sampling

design) equals the quantity to be estimated:

$$\begin{aligned}\mathrm{E}\big(\dot{\mathrm{P}}[X=x]\big) &= \sum\nolimits_{\mathcal{S}}\sum\nolimits_{\omega\in\Omega} w_\omega\, I[X=x](\omega)\,\dot{I}_\omega(\mathcal{S})\,\Pr(\mathcal{S}) \\ &= \sum\nolimits_{\omega\in\Omega} w_\omega\, I[X=x](\omega)\,\pi(\omega) = \mathrm{P}(X=x)\end{aligned}$$

The weights are actually the same as defined in (2) since, in the stratified sampling design, the inclusion probability of units belonging to subpopulation Ω_j is $\pi(\omega) = n_j/N_j$. The argument shows that the plausible estimate defined in (3) can be considered as produced by an unbiased estimator. This allows one to speak of an 'unbiased estimate' in the sense that it is an estimate generated with an unbiased estimator.

Joint and conditional distributions

Formula (3) can easily be extended to estimate joint distributions. For example, to estimate the joint distribution of X and Z, one could use

$$\sum\nolimits_{\omega\in\mathcal{S}} w_\omega\, I[X=x, Z=z](\omega) \tag{7}$$

Using the data from the sample in Table 1, one would get the following estimates:

x	z	Estimate of $\mathrm{P}(X=x, Z=z)$
0	0	$0.002 \cdot 250 = 0.50$
0	1	$0.002 \cdot 100 = 0.20$
1	0	$0.001 \cdot \ \ 80 = 0.08$
1	1	$0.001 \cdot 220 = 0.22$

Referring to the argument of the preceding subsection, these values can be considered as unbiased estimates of the corresponding population quantities.

Estimates of conditional distributions can be derived from estimates of joint distributions. If the selection probabilities used for sampling only depend on variables serving as conditions, weights do not vary in the subsample specified by the conditioning, and it is not necessary to employ weights. In the example, $\mathrm{P}(Y^s = y \,|\, X^s = x, Z^s = z)$ already is a plausible estimate of $\mathrm{P}(Y = y \,|\, X = x, Z = z)$. Since the correspondingly defined estimator is not unbiased, this is an example of a biased estimate that is, nevertheless, plausible.

2.2 Models for Distributions

The idea is to represent the distribution of a variable X in the target population, Ω, by a function depending on parameters to be estimated. I use $g(x;\theta)$ as a generic formulation, to be interpreted as a frequency or a density function, depending on whether the variable's property space, $\mathcal{X}$, is conceived of as discrete or continuous.

How to estimate such models? In order to set up a well-defined estimation problem, one needs to define the model to be estimated (that is, $g(x;\theta)$ *and* the particular value of the parameter θ that should be estimated). The central idea of descriptive estimation, in my understanding, is that one intends to estimate the model that could be calculated if complete data for all units in the target population would be available. Of course, the definition is incomplete until one also has specified a particular method of fitting the model to the data. The general approach is to define a distance function that allows one to quantify the size of the difference between the distribution of X and the model, and then to use the parameter value that minimizes this

difference. Several possibilities exist. Here I refer to the maximum likelihood method which is based on the distance function

$$\mathrm{d_{ML}}(\theta) = \sum\nolimits_{x \in X(\Omega)} \mathrm{P}(X = x) \left(\log(\mathrm{P}(X = x)) - \log(g(x;\theta)) \right) \tag{8}$$

($X(\Omega)$ is the set of different values of X in the target population).[3] Minimizing this distance function is equivalent with maximizing the following log-likelihood function:

$$\ell(\theta) = \sum\nolimits_{x \in X(\Omega)} \mathrm{P}(X = x)\ \log(g(x;\theta)) \tag{9}$$

The model to be estimated is $g(x;\theta^*)$ where θ^* is the value of θ that maximizes $\ell(\theta)$.

The log-likelihood function immediately shows how to estimate the model with data from a sample: $\mathrm{P}(X = x)$ should be substituted by a plausible (unbiased) estimate that can be derived from the sample. Consequently, if it is a simple random sample, one can use the log-likelihood function

$$\ell^s(\theta) = \sum\nolimits_{x \in X(\mathcal{S})} \mathrm{P}(X^s = x)\ \log(g(x;\theta)) \tag{10}$$

Representing the sample as $\mathcal{S} = \{\omega_1, \ldots, \omega_n\}$, and defining $x_i = X(\omega_i)$, this can also be written as

$$\ell^s(\theta) = \frac{1}{n} \sum\nolimits_{i=1,n} \log(g(x_i;\theta)) \tag{11}$$

If $\hat{\theta}$ maximizes this function, $g(x;\hat{\theta})$ can be considered as a plausible estimate of the model $g(x;\theta^*)$.

Stratified sampling

If the data result from stratified sampling, one has to use sampling weights. The log-likelihood function (derived by substituting $\mathrm{P}(X = x)$ in (9) by the plausible (unbiased) estimate (3)) is

$$\ell^s(\theta) = \sum\nolimits_{i=1,n} w_i \log(g(x_i;\theta)) \tag{12}$$

where w_i denotes the sampling weights as defined in (2): $w_i = \frac{N_j}{n_j\, N}$ if x_i belongs to subpopulation Ω_j.

To illustrate, I use the example introduced in Section 2.1. The goal is to estimate a model for the distribution of the variable Y. Since it is a binary variable, a single parameter suffices for a complete representation of the distribution; one can simply use a frequency function

$$g(y;\theta) = \begin{cases} \theta & \text{if } y = 1 \\ 1 - \theta & \text{if } y = 0 \end{cases}$$

Inserting this into (12), one finds the maximand $\hat{\theta} = \sum_{i=1,n} w_i\, y_i$. With the sample data from Table 1 one gets the value

$$\hat{\theta} = 0.002 \cdot 150 + 0.002 \cdot 80 + 0.001 \cdot 56 + 0.001 \cdot 198 = 0.714$$

which equals the value found in Section 2.1.

[3] For further discussion of this distance function see Rohwer and Pötter (2001: 148ff.).

Continuous distribution models

Now assume that Y records school-children's scores in a competence test. One might then use a model based on a continuous distribution, for example a normal density function, say $\phi(y; \mu, \sigma)$. The log-likelihood function to be maximized would be

$$\ell^s(\mu, \sigma) = \sum\nolimits_{i=1,n} w_i \log(\phi(y_i; \mu, \sigma)) \tag{13}$$

The estimated model would be $\phi(y; \hat{\mu}, \hat{\sigma})$ where $\hat{\mu}$ and $\hat{\sigma}$ maximize $\ell^s(\mu, \sigma)$.

2.3 Descriptive Regression Models

Given two variables, X with property space $\mathcal{X}$ and Y with property space $\mathcal{Y}$, defined for a target population Ω, I define a *descriptive regression function* as a function that assigns to each value $x \in \mathcal{X}$ the conditional distribution of Y given $X = x$; symbolically depicted:

$$x \longrightarrow \mathrm{P}[Y \,|\, X = x] \tag{14}$$

Descriptive regression models are correspondingly defined functions which substitute $\mathrm{P}[Y \,|\, X = x]$, which is itself a function and not a single number,[4] by a simplified description. Notice that the approach of descriptive estimation allows one to make a clear distinction between the model and the thing to be represented by the model (here a descriptive regression function).

If Y is a discrete variable, a straightforward approach employs conditional frequencies. There is then, for each value $y \in \mathcal{Y}$, a specific model for the regression function

$$x \longrightarrow \mathrm{P}(Y = y \,|\, X = x) \tag{15}$$

Examples of such models will be considered in Section 2.4. If Y is a quantitative variable, regression models often concern conditional mean values, that is, are models of the regression function

$$x \longrightarrow \mathrm{M}(Y \,|\, X = x) \tag{16}$$

Some examples will be discussed in Section 2.5.

A further possibility is to start from a parametric model for Y, say $g(y; \theta)$, and to make θ a function of x. For example, if Y records the school-children's scores in a competence test, one might use a normal density function, say $\phi(y; \mu, \sigma)$, and a linear link function $\mu = \alpha + x\beta$. Obviously, the possibilities to specify regression models with this approach are nearly unlimited.

An important goal to be served by regression models becomes visible if X consists of several components, say $X = (X_1, \ldots, X_q)$. Then the regression function to be described by a model is

$$(x_1, \ldots, x_q) \longrightarrow \mathrm{P}[Y \,|\, X_1 = x_1, \ldots, X_q = x_q] \tag{17}$$

and one might be interested in finding a simpler description of the dependence on the independent variables. The simplest possibility would be to use a linear link function, say $\theta = \beta_0 + x_1\beta_1 + \cdots + x_q\beta_q$. Of course, this ignores all interaction effects that might be important.

I speak of *descriptive* regression models in order to stress that these models are intended to describe (represent) regression functions which are defined for statistical variables in a target population. This entails that it will most often be necessary to employ sampling weights when estimating such models with data resulting from stratified sampling. I briefly consider this for two kinds of regression models.

[4] $\mathrm{P}[Y \,|\, X = x]$ is a short-cut for the function $y \longrightarrow \mathrm{P}(Y = y \,|\, X = x)$.

2.4 Regression Models for Proportions

These are models for the regression function (15). I use $g_y(x;\theta)$ as a generic formulation for representing $\mathrm{P}(Y = y \,|\, X = x)$. If the model to be estimated is defined with the maximum likelihood method, its definition is based on minimizing the distance function

$$\sum\nolimits_{x,y} \mathrm{P}(Y = y, X = x) \left(\log(\mathrm{P}(Y = y \,|\, X = x)) - \log(g_y(x;\theta)) \right)$$

This is equivalent with maximizing the log-likelihood function

$$\ell(\theta) = \sum\nolimits_{x,y} \mathrm{P}(Y = y, X = x) \, \log(g_y(x;\theta)) \tag{18}$$

The parameter value θ^* that maximizes this function *defines* the model for the target population. If this model is to be estimated with sample data, one should use a log-likelihood function where $\mathrm{P}(Y = y, X = x)$ is substituted by a plausible (unbiased) estimate.

Illustration with a binary logit model

To illustrate, I assume a binary dependent variable ($Y = 0$ or 1), and use a logit model defined by

$$g_1(x;\alpha,\beta) = \frac{\exp(\alpha + x\,\beta)}{1 + \exp(\alpha + x\,\beta)}$$

and $g_0(x;\alpha,\beta) = 1 - g_1(x;\alpha,\beta)$. With data from a stratified random sample one should use the log-likelihood function

$$\ell^s(\theta) = \sum\nolimits_{i=1,n} w_i \left(y_i \log(g_1(x_i;\theta)) + (1 - y_i) \log(g_0(x_i;\theta)) \right) \tag{19}$$

with sampling weights w_i defined by (2). Notice that, except when estimating a saturated model, the weights are required even if the selection probabilities employed in the sampling design depend only on variables that are used as independent variables in the model. To illustrate, I use a logit model for the regression function

$$(x, z) \longrightarrow \mathrm{P}(Y = y \,|\, X = x, Z = z) \tag{20}$$

based on the example introduced in Section 2.1. Sampling weights are not required when estimating a saturated model

$$g_1(x;\alpha,\beta_x,\beta_z,\beta_{xz}) = \frac{\exp(\alpha + x\,\beta_x + z\,\beta_z + x\,z\,\beta_{xz})}{1 + \exp(\alpha + x\,\beta_x + z\,\beta_z + x\,z\,\beta_{xz})}$$

They should be used, however, as soon as the model is simplified by omitting an interaction term. In this example, a simplified model would be

$$g_1(x;\alpha,\beta_x,\beta_z) = \frac{\exp(\alpha + x\,\beta_x + z\,\beta_z)}{1 + \exp(\alpha + x\,\beta_x + z\,\beta_z)}$$

Estimating this model with the data in Table 1, one gets the following results:

	population	sample with weights	sample without weights
$\hat{\alpha}$	0.3850	0.3850	0.3734
$\hat{\beta_x}$	0.6136	0.6136	0.5917
$\hat{\beta_z}$	1.0801	1.0801	1.1383

Obviously, sampling weights are required for plausible estimates of the model parameters defined for the target population.

Sampling weights are required, in particular, when the selection probabilities depend on the dependent variable of the model. In our example, this would be a model in which the probability of attending one or the other school type (X) is made dependent on the parents' educational level (Z). The log-likelihood function (18) shows that model estimation would require a plausible estimate of the joint distribution $\mathrm{P}(X = x, Z = z)$.

2.5 Least Squares Estimation

Another estimation method is least squares (LS) estimation. It is often used for the definition and estimation of regression models for conditional mean values as depicted in (16). In order to refer to a model for mean values $\mathrm{M}(Y \mid X = x)$, I use $m(x;\theta)$ as a generic formulation. The model for the target population is then defined by a parameter value θ^* that minimizes the LS distance function

$$\mathrm{LS}(\theta) = \sum\nolimits_{x \in X(\Omega)} \mathrm{P}(X = x)\big(\mathrm{M}(Y \mid X = x) - m(x;\theta)\big)^2 \tag{21}$$

This is equivalent (see Rohwer and Pötter 2001: 135f.) with minimizing the function

$$\sum\nolimits_{\omega \in \Omega} \big(Y(\omega) - m(X(\omega);\theta)\big)^2$$

This immediately leads to the usual formulation of LS estimation with data from a simple random sample, namely,

$$\sum\nolimits_{i=1,n} \big(y_i - m(x_i;\theta)\big)^2 \tag{22}$$

If the data result from stratified sampling, it is helpful to start from (21) because this formulation shows what should be done in order to find plausible parameter estimates: $\mathrm{P}(X = x)$ and $\mathrm{M}(Y|X = x)$ should be substituted by plausible (unbiased) estimates. In order to derive a formula for weighted LS regression, one can start from rewriting (21) as

$$\mathrm{LS}(\theta) = \sum\nolimits_{x} \mathrm{P}(X = x)\Big(\sum\nolimits_{y} \frac{y\,\mathrm{P}(Y = y, X = x)}{\mathrm{P}(X = x)} - m(x;\theta)\Big)^2$$

Minimizing this function is equivalent with minimizing

$$\sum\nolimits_{x} \mathrm{P}(X = x)\, m(x;\theta)^2 - 2\, m(x;\theta) \sum\nolimits_{y} y\,\mathrm{P}(Y = y, X = x)$$

Substituting $\mathrm{P}(X = x)$ and $\mathrm{P}(Y = y, X = x)$ by plausible (unbiased) estimates can now be done by using weights in the following way:

$$\sum\nolimits_{i=1,n} w_i\, m(x_i;\theta)^2 - 2\, w_i\, m(x_i;\theta)\, y_i$$

where the weights w_i are defined by (2). Finally, minimizing this function is equivalent with minimizing

$$\sum\nolimits_{i=1,n} w_i \big(y_i - m(x_i;\theta)\big)^2 \tag{23}$$

which is the standard formulation for LS estimation with sampling weights.

It is noteworthy that this formulation also covers the case where the selection probabilities depend on the dependent variable of the model.

3. Probabilistic Data Models

I now consider *probabilistic data models*. The following quotation from D. R. Cox and N. Wermuth (1996: 12) explains the basic ideas.

> The basic assumptions of *probabilistic analyses* are as follows: 1. The data are observed values of random variables, i.e. of variables having a probability distribution. 2. Reasonable working assumptions can be made about the nature of these distributions, usually that they are of a particular mathematical form involving, however, unknown constants, called parameters. We call this representation a model, or more fully a probability model, for the data. 3. Given the form of the model, we regard the objective of the analysis to be the summarization of evidence about either the unknown parameters in the model or, occasionally, about the values of further random variables connected with the model, and, very importantly, the interpretation of that evidence.

The most important assumption underlying this modeling approach is that the data in a given sample can be considered as values of random variables. To make this explicit, I refer to a population, Ω, for which a statistical variable $X : \Omega \longrightarrow \mathcal{X}$ is defined, and to a sample $\mathcal{S} = \{\omega_1, \ldots, \omega_n\}$ from this population. The observed data are $x_i = X(\omega_i)$, for $i = 1, \ldots, n$. The basic assumption then is that one can think of these values as realizations of random variables

$$\dot{X}_1, \ldots, \dot{X}_n \tag{24}$$

Such variables will be called *data representing random variables*. To indicate that these are random variables, and therefore conceptually different from statistical variables, they are marked by a dot.

Probabilistic modeling consists in making assumptions about the probability distributions of these random variables. Unfortunately, it is unclear how to understand these random variables, and different interpretations exist (two interpretations will be discussed briefly in Section 3.4).[5] A further obscurity concerns the goal of the modeling. It is often said that the goal is to model 'data-generating processes'. However, usage of this term easily obscures an important distinction between two kinds of processes:

a) Processes that generate real-world facts. Referring to the example introduced in Section 2.1, one can think of the processes that generate for each school-child specific values of the variables X (school type) and Y (outcome). Such processes will be called *fact-generating processes*.[6]

b) Processes that generate data, that is, information about already existing facts. Such processes can properly be called *data-generating processes*. They include, in particular, the selection of units to be included in a (random) sample. Such processes obviously presuppose that fact-generating processes have taken place.

Distributions of data representing random variables result from both kinds of processes. I suppose that the theoretical interest concerns the fact-generating processes. One therefore has to decide whether one needs to distinguish the data representing random variables from the theoretical interesting variables intended to represent the fact-generating processes.

[5] I will not discuss so-called superpopulation models which start from random variables defined for the population Ω (and not just for the given sample).

[6] In Rohwer (2010), they have been called 'substantial processes'. Since the term 'substantial' is ambiguous, I now prefer to speak of 'fact-generating processes'.

3.1 Simple Stochastic Estimators

For ease of notation, I conceive of the random variables $\dot{X}_i$ as having a discrete property space so that one can refer to probability functions $f_i(x) = \Pr(\dot{X}_i = x)$. It is often supposed (not only if the data result from simple random sampling) that the variables have identical distributions, $f_i(x) = f(x)$ (being the distribution of a random variable $\dot{X}$), and are stochastically independent (briefly: i.i.d.). Now assume that one wants to estimate $f(x)$. Descriptive estimation in the understanding of Section 2.1 is no longer appropriate. Instead, one can use the data representing random variables to define an estimator that is itself a random variable. One can start from random variables

$$\dot{I}[\dot{X}_i = x] = \begin{cases} 1 & \text{if } \dot{X}_i = x \\ 0 & \text{otherwise} \end{cases} \tag{25}$$

and use these variables to define the estimator

$$\dot{U}_x = \frac{1}{n} \sum\nolimits_{i=1,n} \dot{I}[\dot{X}_i = x] \tag{26}$$

The sampled values $x_1, \ldots, x_n$ can then be viewed as providing a specific value of this estimator, to be interpreted as an estimate of $f(x)$. That the estimator is a random variable opens the opportunity to define 'unbiased' in a way that is not available with descriptive estimation:

$$\mathrm{E}(\dot{U}_x) = f(x)$$

Notice that the expectation, E(.), is here not defined w.r.t. the probability distribution which is associated with the sampling procedure used to generate the actual sample. Instead, it is defined w.r.t. the distribution of the data representing random variables.

Moreover, one can think that the estimator has a variance that can be estimated and used to assess the 'precision' of the estimate; in the current example,

$$\mathrm{V}(\dot{U}_x) = \frac{f(x)\,(1 - f(x))}{n}$$

Stratified Samples

Now consider data from stratified sampling based on a stratification variable H that distinguishes m subpopulations (I use the notation introduced in Section 2.1). One then needs to distinguish the data representing variables $\dot{X}_i$ from a theoretically interesting random variable, say $\dot{X}^*$. Of course, this variable must be defined before its distribution, say $f^*(x)$, can be estimated. One possibility is as follows. One starts from the assumption that in each subsample, $\mathcal{S}_j$, the variables $\dot{X}_i$ are i.i.d. with $f_{(j)}(x)$. This allows one to define

$$f^*(x) = \sum\nolimits_{j=1,m} \frac{N_j}{N}\, f_{(j)}(x)$$

Using then

$$\dot{U}_{x,j} = \frac{1}{n_j} \sum\nolimits_{i \in \mathcal{S}_j} \dot{I}[\dot{X}_i = x]$$

as an estimator for $f_{(j)}(x)$, an appropriate estimator for $f^*(x)$ is

$$\dot{U}^*_x = \sum\nolimits_{i=1,n} w_i\, \dot{I}[\dot{X}_i = x] \tag{27}$$

where w_i are the sampling weights defined in (2). This is formally analogous to the function (3); the main difference is that (27) allows one to think in terms of a stochastic estimator that can be defined without reference to a sampling design for the given sample.

3.2 Models for Distributions

I now briefly discuss models for distributions of data representing random variables. I begin again with the supposition that the variables $\dot{X}_i$ are i.i.d. with $f(x)$. The theoretically assumed model is given by $g(x;\theta)$. Estimation with the maximum likelihood method proceeds by minimizing the distance function

$$\sum\nolimits_x f(x)\,\big(\log(f(x)) - \log(g(x;\theta))\big)$$

This is equivalent with maximizing the log-likelihood function

$$\sum\nolimits_x f(x)\log(g(x;\theta)) \tag{28}$$

The model to be estimated is $g(x;\theta^*)$ where θ^* is the value of θ that maximizes this function.

This approach can also be viewed as providing a stochastic estimator. Substituting $f(x)$ by the estimator (26), one gets

$$\frac{1}{n}\sum\nolimits_{i=1,n}\sum\nolimits_x \dot{I}[\dot{X}_i = x]\log(g(x;\theta)) \tag{29}$$

This also shows how to estimate a model for a distribution $f^*(x)$ supposed to exist if the data result from stratified sampling (see Section 3.1). One substitutes $f^*(x)$ by the estimator (27), and then gets

$$\sum\nolimits_{i=1,n}\sum\nolimits_x w_i\,\dot{I}[\dot{X}_i = x]\log(g(x;\theta)) \tag{30}$$

covering (29) as a special case. The corresponding estimator $\dot{U}(\dot{X}_1,\ldots,\dot{X}_n)$ is defined as follows: If $\dot{X}_1 = x_1,\ldots,\dot{X}_n = x_n$ and $\hat{\theta}$ maximizes

$$\ell(\theta) = \sum\nolimits_{i=1,n} w_i\log(g(x_i;\theta)) \tag{31}$$

then $\dot{U}(\dot{X}_1,\ldots,\dot{X}_n) = \hat{\theta}$. In this sense one can consider $\hat{\theta}$ as an estimate of θ^* that results from the sample $x_1,\ldots,x_n$.

3.3 Probabilistic Regression Models

I now consider probabilistic regression models that are based on data representing random variables

$$(\dot{X}_1,\dot{Y}_1),\ldots,(\dot{X}_n,\dot{Y}_n) \tag{32}$$

I first assume that these variables are i.i.d. with a probability function $f(x,y) = \mathrm{P}(\dot{X} = x, \dot{Y} = y)$. As an example, I consider a model for the regression function

$$x \longrightarrow \mathrm{E}(\dot{Y} \mid \dot{X} = x) \tag{33}$$

Notice that, given x, $\mathrm{E}(\dot{Y} \mid \dot{X} = x)$ is a fixed value, defined by assuming a distribution $f(x,y)$. The regression model is intended to model these conditional expectations. I use $m(x;\theta)$ as a generic formulation. In addition, one must define the parameter, say θ^*, that one intends to estimate. As was done in Section 2.5, I use the least squares method, that is, θ^* is defined as the parameter value that minimizes

$$\sum\nolimits_x f(x)\big(\mathrm{E}(\dot{Y} \mid \dot{X} = x) - m(x;\theta)\big)^2 \tag{34}$$

Starting from this definition, one gets an estimator of θ^* by substituting $f(x)$ and $\mathrm{E}(\dot{Y}|\dot{X} = x) = \sum_y yf(x,y)/f(x)$ by suitable estimators.

The same approach can be followed when the data representing variables (32) relate to a stratified sample. It is assumed, then, that the model concerns a regression function

$$x \longrightarrow \mathrm{E}(\dot{Y}^* \,|\, \dot{X}^* = x) \tag{35}$$

where the theoretically interesting variable, $(\dot{X}^*, \dot{Y}^*)$, is defined by

$$f^*(x,y) = \sum\nolimits_{j=1,m} \frac{N_j}{N}\, f_{(j)}(x,y) \tag{36}$$

(based on assuming that, in each subsample $\mathcal{S}_j$, the variables $(\dot{X}_i, \dot{Y}_i)$ are i.i.d. with $f_{(j)}(x,y)$). The model to be estimated is now defined by the parameter θ^* that minimizes

$$\sum\nolimits_x f^*(x)\big(\mathrm{E}(\dot{Y}^* \,|\, \dot{X}^* = x) - m(x;\theta)\big)^2 \tag{37}$$

In order to find a suitable estimator, one can substitute $f^*(x)$ by the estimator (27), and $\mathrm{E}(\dot{Y}^* \,|\, \dot{X}^* = x)$ by the estimator

$$\sum\nolimits_y \frac{y \sum_i w_i \dot{I}[\dot{X}_i = x, \dot{Y}_i = y]}{\sum_i w_i \dot{I}[\dot{X}_i = x]} \tag{38}$$

The resulting formula looks complicated, namely,

$$\sum\nolimits_x \sum\nolimits_i w_i\, \dot{I}[\dot{X}_i = x] \Big(\sum\nolimits_y \frac{y \sum_i w_i \dot{I}[\dot{X}_i = x, \dot{Y}_i = y]}{\sum_i w_i \dot{I}[\dot{X}_i = x]} - m(x;\theta) \Big)^2 \tag{39}$$

However, the corresponding estimator for θ^*, say $\dot{U}(\dot{X}_1, \dot{Y}_1, \ldots, \dot{X}_n, \dot{Y}_n)$, is quite simple: If the sample is $(x_1, y_1), \ldots, (x_n, y_n)$, the estimator provides the value $\dot{U}(\dot{X}_1, \dot{Y}_1, \ldots, \dot{X}_n, \dot{Y}_n) = \hat{\theta}$ that minimizes the function

$$\sum\nolimits_{i=1,n} w_i \big(y_i - m(x_i;\theta)\big)^2 \tag{40}$$

(see the formally analogous derivation in Section 2.5).

When one should use sampling weights?

Notice that (40) requires weights even if the selection probabilities used in the stratified sampling depend only on variables included as independent variables in the regression model (the only exception occurs when there is a separate model for each subsample (stratum)). This is a consequence of understanding the conditional expectations, $\mathrm{E}(\dot{Y}|\dot{X} = x)$, as quantities that are fixed independently of the modeling exercise. This entails that the regression model, $m(x;\theta^*)$, must be understood as approximately representing these quantities.

There is, however, another understanding of probabilistic regression models that starts from the idea that one can use a model to make assumptions about the probability distribution of the data representing variables (see above the quotation from Cox and Wermuth). An often made assumption is that there is a parameter value θ^* such that

$$\mathrm{E}(\dot{Y} \,|\, \dot{X} = x) = m(x;\theta^*) \tag{41}$$

This assumption allows one to argue that sampling weights are not required if they depend only on variables included as independent variables in the model. This can be seen in the following

way. If sampling weights depend only on values of $\dot{X}$, the weights w_i in the estimator (38) cancel because they do no vary. If furthermore (41) holds, also the probability function $f^*(x)$ in (37) can be dropped. Consequently, the function (39) which is used to define the estimator can be simplified into

$$\sum\nolimits_x \Big(\sum\nolimits_y \frac{y \sum_i \dot{I}[\dot{X}_i = x, \dot{Y}_i = y]}{\sum_i \dot{I}[\dot{X}_i = x]} - m(x;\theta) \Big)^2$$

Finding then the minimum for a sample $(x_1, y_1), \ldots, (x_n, y_n)$ is equivalent with minimizing the unweighted least squares function

$$\sum\nolimits_{i=1,n} \big(y_i - m(x_i;\theta) \big)^2 \tag{42}$$

If (41) holds, $m(x;\theta^*)$ could be called a 'true regression model'.[7] This notion also motivates a specific understanding of 'omitted variables': variables that should be added to the model in order to make it a true regression model.

3.4 How to Understand the Approach?

The modeling approach based on data representing random variables is quite flexible. As soon as one has introduced these variables one can make arbitrary assumptions about their distributions and then use the methods of formal probability theory to derive implications. Unfortunately, it is unclear how to understand these random variables.

The main obscurity is due to the fact that these variables are defined by reference to a given sample, say $\mathcal{S} = \{\omega_1, \ldots, \omega_n\}$. Otherwise it would not be possible to distinguish data representing random variables by indices, i, that refer to individual units. However, assume that the index i refers to a particular unit, ω_i, how then to make sense of realizations of the random variable $\dot{X}_i$? Except for measurement errors, this variable can only have a single value, the one that was recorded in the given sample.

In order to avoid these obscurities, one can try to define data representing random variables by a sampling procedure. Given the notion of a statistical variable $X : \Omega \longrightarrow \mathcal{X}$, one can define a random variable $\dot{X}$ in the following way: randomly draw with replacement a unit ω, and then take $X(\omega)$ as a realization of $\dot{X}$. This is a conceptually valid definition, and it entails a definite probability distribution for $\dot{X}$: $\Pr(\dot{X} = x) = \mathrm{P}(X = x)$.

Of course, this is not normally the method that is used to create $\mathcal{S}$. If it is a simple random sample, one might, nevertheless, use $\dot{X}$ as an approximately valid representation of the data generating process that has produced the sampled values.[8] However, these values must then be conceived of as realizations of one single random variable, $\dot{X}$, and indices referring to particular units cannot be used.

In a similar way one could define data representing random variables by referring to a stratified sampling procedure. For each subpopulation one can define a separate random variable, say $\dot{X}_{(j)}$. Assuming then known fractions N_j/N, their distributions could be mixed to define a random variable $\dot{X}$ representing the target population. Again, there would be no possibility to introduce random variables indexed by references to particular units.

As a consequence of following this approach to introduce data representing random variables, models must be understood as descriptions of probability distributions that are fixed before the

[7] The belief that one has specified a true regression model is often thought to be a prerequisite even for linear OLS regression; see e.g. Winship and Radbill (1994: 232).

[8] This assumption is often made, see e.g. DuMouchel and Duncan (1983: 536).

modeling takes place (and actually are derived from *statistical* distributions in a target population). This entails that assumption (41) is most often not reasonable. A different understanding of the data representing random variables is required in order to allow one to make arbitrary assumptions about their distributions and, in particular, to think in terms of a 'true regression model'. The random variables must then be understood as theoretical fictions invented to make a special form of probabilistic modeling possible (see, e.g., Berk 2004: 53ff.).

4. Probabilistic Functional Models

I now consider functional models that serve to formulate rules for generic units (or situations). Such models can be conceptualized either as deterministic or as probabilistic models (see Rohwer 2010, 2011). Here I only consider probabilistic functional models (subsequently I drop the adjective and simply speak of functional models).

4.1 Modeling Probabilistic Rules

As an example of a probabilistic rule consider the following: The probability that a child successfully completes a grade is higher in schools of type 1 than in schools of type 0. This is not a statement about any particular school-child, or any particular population of school-children. It is not a descriptive statement at all. Instead, it is a rule which refers to a generic school-child. How to formulate such rules? As a first step, one can think that the rule formulates a dependency relation between two variables; graphically depicted:

$$\ddot{X} \longrightarrow\!\!\!\!\!\rightarrow \dot{Y} \tag{43}$$

The variable $\ddot{X}$ serves to make an assumption about the school type (0 or 1), and $\dot{Y}$ serves to refer to possible outcomes (1 if success, 0 if no success). $\ddot{X}$ is an exogenous, $\dot{Y}$ is an endogenous variable of the model. Since values of $\ddot{X}$ can be arbitrarily fixed, $\ddot{X}$ can be conceived of neither as a statistical nor as a random variable. To remind of its special status as an exogenous variable without an associated distribution it is marked by two dots. Since $\ddot{X}$ has no distribution, there also is no distribution for $\dot{Y}$ (and it is therefore not a random variable in the usual sense of the word). However, in order to make quantitative statements possible, one can think of *conditional* distributions of $\dot{Y}$ if particular values of $\ddot{X}$ are fixed. To make this idea explicit, one uses a stochastic function

$$x \longrightarrow \Pr[\dot{Y} \mid \ddot{X} = x] \tag{44}$$

that assigns to each value x of $\ddot{X}$ a conditional probability distribution of the variable $\dot{Y}$.

In my understanding, these are epistemic probability distributions quantifying the uncertainty of using the rule for a prediction. The probabilities are not fixed by real-world facts but reflect the beliefs and the knowledge of people who are interested in the predictions. Possibilities to find values of these probabilities therefore depend on the application context. If one can refer to an artificial random generator, or an analogously conceivable process frame (a 'random experiment'), classical inference methods which presuppose objective probabilities can be used. In most social science applications process frames can be conceived of as random generators in only a very loose sense, and quantification of epistemic probabilities must be based on samples from historically changing populations. Of course, there is no other possibility than to rely on observed conditional frequencies. However, in contrast to distributions of statistical variables defined for specified target populations, the epistemic probabilities to be estimated cannot be assumed to be objectively fixed quantities, but must be understood as being defined by suitable estimation methods. (Notice that I here and subsequently speak of 'estimation' without presupposing defined quantities that could be estimated in a proper sense.)

4.2 Functional Models without Parameters

All models discussed in the present section refer to the example introduced in Section 2.1 (Table 1). The focus is on how to estimate the models with data which result from stratified sampling and, in particular, whether one should use sampling weights.

Selection depends only on exogenous variables

A first situation occurs when the selection probabilities which are used in the sampling procedure depend only on exogenous variables of the model. As an example, one can think of a dependency relation $\ddot{X} \longrightarrow\!\!\!\!\!\rightarrow \dot{Y}$ where $\ddot{X}$ (corresponding to X) specifies the school type, and $\dot{Y}$ (corresponding to Y) represents the outcome. Since both variables are binary, it suffices to consider stochastic function

$$x \longrightarrow \Pr(\dot{Y} = 1 \,|\, \ddot{X} = x) \tag{45}$$

as a quantitative functional model. The conditional probabilities can be estimated by corresponding frequencies:

$$\Pr(\dot{Y} = 1 \,|\, \ddot{X} = 0) \;\text{ estimated by }\; \mathrm{P}(Y^s = 1 \,|\, X^s = 0) = 0.657$$
$$\Pr(\dot{Y} = 1 \,|\, \ddot{X} = 1) \;\text{ estimated by }\; \mathrm{P}(Y^s = 1 \,|\, X^s = 1) = 0.847$$

Since the selection probabilities depend only on X, there is no need to employ sampling weights.

Selection depends on endogenous variables

A different situation occurs when the selection probabilities depend on an endogenous variable of a functional model. To illustrate with the example, one might be interested in the stochastic function

$$z \longrightarrow \Pr(\dot{X} = 1 \,|\, \ddot{Z} = z) \tag{46}$$

which assumes that the probability of attending a school of a particular type depends on the educational level of the child's parents. Since the selection probabilities used for the stratified sampling depend on the variable X, $\mathrm{P}(X^s = 1 \,|\, Z^s = z)$ is certainly not a good estimate of $\Pr(\dot{X} = 1 \,|\, \ddot{Z} = z)$, and one should use instead a plausible estimate of $\mathrm{P}(X = 1 \,|\, Z = z)$. For example, as an estimate of $\Pr(\dot{X} = 1 \,|\, \ddot{Z} = 0)$ one can use the estimate

$$\frac{0.001\,\mathrm{P}(X^s = 1, Z^s = 0)}{0.002\,\mathrm{P}(X^s = 0, Z^s = 0) + 0.001\,\mathrm{P}(X^s = 1, Z^s = 0)} = 0.138$$

This is the estimate that one would use in descriptive estimation. In the present context, the argument for using this estimate is, however, different. In the descriptive approach one is interested in estimating the quantity $\mathrm{P}(X = 1 \,|\, Z = 0)$ that is defined for a particular target population. Instead, when estimating a functional model, one uses observed conditional frequencies for the quantification of epistemic probabilities. The argument for using sampling weights is then based on the intention to avoid distortions produced by a data generating process.

Selection depends on not included variables

Still another situation occurs when the selection probabilities depend on variables that are not included in the model. As an example I consider the stochastic function

$$z \longrightarrow \Pr(\dot{Y} = 1 \,|\, \ddot{Z} = z) \tag{47}$$

which assumes that the probability of a child's success in completing a grade depends on the parents' educational level. Should one use sampling weights when estimating these conditional probabilities?

Since the goal is not to make descriptive statements about a target population, there is no immediate answer. The first question should therefore be why there might be relevant differences between weighted and unweighted estimates. Differences occur if the conditional probability distribution of $\dot{Y}$, in addition to being dependent on the model's exogenous variables, also depends on the variable used for the stratified sampling. In the example, this is the variable X. The easiest solution therefore is: If weighted and unweighted estimates differ (significantly), include the variable used for stratification as an additional exogenous variable in the model. In our example, the enlarged model would be

$$(x, z) \longrightarrow \Pr(\dot{Y} = 1 \,|\, \ddot{X} = x, \ddot{Z} = z) \tag{48}$$

Estimation of this model does not require to use sampling weights.

Reference to this enlarged model also provides a hint why the question of whether to use sampling weights when estimating (47) has no clear answer. The model (47) cannot be derived from the enlarged model. To do this would require a distribution for the variable $\ddot{X}$ which does not exist. In order to perform the derivation one would need to substitute $\ddot{X}$ by a statistical variable X (or a random variable $\dot{X}$). Using a statistical variable X, one could write:

$$\Pr(\dot{Y} = 1 \,|\, \ddot{Z} = z) = \sum\nolimits_x \Pr(\dot{Y} = 1 \,|\, \ddot{Z} = z, X = x)\, \mathrm{P}(X = x \,|\, \ddot{Z} = z)$$

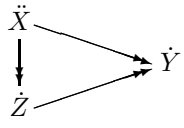
showing how $\Pr(\dot{Y} = 1 \,|\, \ddot{Z} = z)$ depends on a statistical distribution. The question which distribution should be used has no clear answer, however, because the model (47) does not refer to any particular population.

Of course, one can think of functional models intended to make predictions for units in a particular population. This could provide an argument for using conditional frequencies which are plausible (unbiased) estimates for the particular population. However, in social science applications where populations continuously change, one is seldom interested in functional models restricted to a particular point in time.

Adding structural relationships

The 'weighting problem' has no unique solution as long as both, $\ddot{X}$ (on which selection probabilities depend) and $\ddot{Z}$, are viewed as exogenous variables of the model. One should therefore think about possible relationships between the exogenous variables. There are three possibilities.

(a) $\ddot{Z}$ depends on $\ddot{X}$. This entails a new functional model in which the formerly exogenous variable $\ddot{Z}$ has become an endogenous variable, $\dot{Z}$. Graphically depicted:

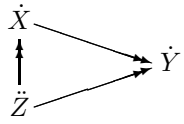


(49)

When estimating this model one would not need sampling weights.

(b) A second possibility is that $\ddot{X}$ depends on $\ddot{Z}$. Again, this entails a new functional model in

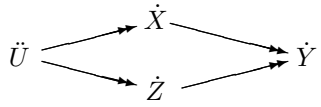
which the formerly exogenous variable $\ddot{X}$ has become an endogenous variable, $\dot{X}$. Graphically depicted:



(50)

Estimating this model would require to use sampling weights because the selection probabilities of the sampled data now depend on an endogenous variable of the model.

(c) A third possibility is to substitute both, $\ddot{X}$ and $\ddot{Z}$, by endogenous variables, $\dot{X}$ and $\dot{Z}$, which are assumed to depend on an exogenous variable, say $\ddot{U}$. Graphically depicted:



(51)

If $\ddot{U}$ is a meaningful variable and observations are included in the sample, one should use weights to estimate the model (since selection probabilities depend on an endogenous variable of the model). No solution of the weighting problem will be gained, however, if $\ddot{U}$ is an unobserved variable. Only arbitrary assumptions about the common distribution of $\dot{X}$ and $\dot{Z}$ would then be possible.

4.3 Parametric Functional Models

The basic functional model concerns a dependency relation $\ddot{X} \longrightarrow\!\!\!\!\!\rightarrow \dot{Y}$ and uses a stochastic function $x \longrightarrow \Pr[\dot{Y}|\ddot{X} = x]$ as a framework for quantitative statements. If no specific parametric model is assumed, one directly refers to the conditional probabilities $\Pr(\dot{Y} = y \,|\, \ddot{X} = x)$. Instead, one can set up a parametric model, say

$$x \longrightarrow g(x;\theta) \tag{52}$$

In my understanding, this model uses $g(x;\theta)$ for giving $\Pr[\dot{Y}|\ddot{X} = x]$, the epistemic probability distributions, a specific mathematical form; in a sense, $g(x;\theta)$ then *defines* how to conceive of $\Pr[\dot{Y}|\ddot{X} = x]$. This understanding entails that $g(x;\theta)$ is not intended to describe a conditional probability distribution that can be assumed to exist independently of the model. (Consequently, there is no question whether the model might be 'true' or not.)

This view is in accord with the understanding that functional models serve to formulate rules and do not intend to describe distributions in a target population. It follows that the principles of descriptive estimation are not applicable to the estimation of parametric functional models. In particular, there is no place for the argument that one should use sampling weights in order to get unbiased estimates of parameters that are defined by reference to a target population.

There is therefore a difference to the estimation of probabilistic data models. As has been argued in Section 3.3, when estimating these models one should use sampling weights even if the weights only depend on independent variables of the regression model. This is not required when estimating parametric functional models. Apart from this, all considerations of sampling weights discussed in the previous subsection can also be applied to the estimation of parametric functional models.

References

Berk, R. A. (2004). *Regression Analysis. A Constructive Critique*. Thousand Oakes: Sage.

Cox, D. R. and Wermuth, N. (1996) *Multivariate Dependencies – Models, Analysis and Interpretation*, London: Chapman & Hall.

DuMouchel, W. H., Duncan, G. J. (1983). Using Sample Survey Weights in Multiple Regression Analyses of Stratified Samples. *Journal of the American Statistical Association* 78, 535–543.

Rohwer, G. (2010). *Models in Statistical Social Research*. London: Routledge.

Rohwer, G. (2011). Qualitative Comparative Analysis. A Discussion of Interpretations. *European Sociological Review* 27, 728–740.

Rohwer, G., Pötter, U. (2001). *Grundzüge der sozialwissenschaftlichen Statistik*. Weinheim: Juventa.

Winship, C., Radbill, L. (1994). Sampling Weights and Regression Analysis. *Sociological Methods & Research* 23, 230–257.