

9 Model Assumption Errors

David Draper & Russell Bowater⁴, University of Bath

9.1 Introduction

The original goal of design-based analysis methods in survey sampling was “the development of a sampling theory that is model-free” (Cochran 1977). Even within classical design-based methods, however, the incorporation of auxiliary information through such techniques as ratio and regression estimation is essentially (if perhaps somewhat covertly) model-based. Today overtly model-based methods are commonly employed in business statistics, in the calculation of index formulae, in the use of benchmarking and seasonal adjustment (where model-based outlier detection and correction are crucial), and in estimation when no data for a sub-population are available (for example, enterprises that fall below a size threshold, as in cut-off sampling, or small-area estimation from aggregate data). Models are thus ubiquitous in the analysis of business survey data (see, for example, Särndal *et al.* 1992), and the assumptions they make must be critically reviewed with an eye to quantifying model assumption errors.

We have already encountered the use of models in several previous chapters; in particular, in section 2.3.2 we examined the idea of treating the population from which the sample at hand was drawn as itself a sample from a *superpopulation* specified by a model. An example of this idea that is relevant to model assumption errors came up in the discussion of quota sampling in section 4.3: if the population values y_j in the cells of the quota-sampling grid are assumed to be random variables with $E_{\xi}(y_j) = \mu_h$, and $V_{\xi}(y_j) = \sigma_h^2$, where h indexes the cell in the grid in which y_j is observed, then model-unbiased estimates both of the population total t (\hat{t} , say) and the variance of \hat{t} are available and coincide with the usual design-unbiased estimates from stratified sampling. However, this is equivalent to the modelling assumption that the observed y_j values in the quota sample are stochastically indistinguishable from what one would obtain with simple random sampling (without replacement) from the cells in the grid, and there is no way to completely verify this assumption from the data. Errors in this model assumption could lead to a bias in the estimate of t whose magnitude and even direction are hard to quantify.

In the following sections we examine in turn the five leading areas in which model assumption errors appear crucial in business surveys: index formulae, benchmarking, seasonal adjustment, cut-off sampling, and coping with non-ignorable nonresponse. In the final section we offer some recommendations on best practice in the reporting of possible model assumption errors in business surveys.

⁴ We are grateful to Ray Chambers (University of Southampton), Eva Elvers (Statistics Sweden) and Paul Smith (UK Office for National Statistics) for comments and references, and to Paul Smith for some suggested text fragments. Membership on this list does not imply agreement with the ideas expressed here, nor are any of these people responsible for any errors or omissions that may be present.

9.2 Index numbers

As noted by Jazairi (1982), an *index number* is a measure of the magnitude of a variable at one point relative to its value at another point. The variable in question is often either the price or the (sales) quantity (or volume) of a commodity. The “points” in question may be different times, or locations, or groups of households; we will focus here on time, measured in months. In the simplest form of this idea there are only two points in time being compared; one, say t (often the earlier time-point), is selected as the *reference* or *base month*, and the other, say t' , is the *current month*.

Consider a set or *market basket*, C , of commodities c_1, \dots, c_m observed at n times, and let p_{it} and q_{it} be the price and volume, respectively, of commodity c_i at time t . The *money value* of c_i at time t is by definition simply the product $v_{it} \equiv p_{it}q_{it}$. The ratio $p_{it'}/p_{it}$ of the price of commodity c_i at time t' to its price at time t is the *price ratio*; the corresponding fraction $q_{it'}/q_{it}$ is the *volume ratio*. In attempting to measure how much the price of the market basket C has changed over time, an old (18th century) idea was simply to form the average $\frac{1}{m} \sum_{i=1}^m \frac{p_{it'}}{p_{it}}$ of the price ratios; in the 19th century the German economists Laspèyres and

Paasche introduced a refinement of this idea which is still used today. The *Laspèyres price* and *volume indices*, respectively, are ratios of weighted sums of the form

$$LP_{t'} = \frac{\sum_{i=1}^m p_{it'} q_{it}}{\sum_{i=1}^m p_{it} q_{it}}, \quad LV_{t'} = \frac{\sum_{i=1}^m q_{it'} p_{it}}{\sum_{i=1}^m q_{it} p_{it}}; \quad (9.1)$$

for example, the Laspèyres price index represents the ratio of the cost of the base month market basket at the current month prices to its cost at the prices of the base month. Similarly the *Paasche price* and *volume indices*, respectively, are

$$PP_{t'} = \frac{\sum_{i=1}^m p_{it'} q_{it'}}{\sum_{i=1}^m p_{it} q_{it'}}, \quad PV_{t'} = \frac{\sum_{i=1}^m q_{it'} p_{it'}}{\sum_{i=1}^m q_{it} p_{it'}}; \quad (9.2)$$

thus the Paasche indices are similar to those of Laspèyres except that in Laspèyres' weighted sums the weights are measured in the base month and Paasche's weights are those in the current month. With any given market basket, and base and current months, the Laspèyres and Paasche price indices will typically not agree (essentially for the same reason that the relative change of a quantity q_i from time t to t' , $((q_{i'} - q_i)/q_i)$, does not coincide with the relative change from t' to t , $((q_i - q_{i'})/q_{i'})$); the *Fisher ideal index*, the geometric mean of the Laspèyres and Paasche formulae, is frequently used as a compromise. There are many