10

Multivariate Analysis of Health Survey Data

The most basic description of health sector inequality is given by the bivariate relationship between a health variable and some indicator of socioeconomic status (SES) captured, for example, by the concentration curve and index. For a finer description, the analyst might want to standardize for demographic factors, such as age and gender (see chapters 5 and 15). Or, the analyst might want to explain the inequality through decomposition into its constituent parts (see chapters 12 and 13). More ambitiously, the analyst might want to test for the existence of a causal relationship between a health variable and SES and to examine the nature of any causality. All of these tasks require moving from bivariate to multivariate analysis. In this chapter, we discuss some issues that generally deserve consideration when undertaking multivariate analysis of survey data for the purpose of learning about health sector inequality or inequity. First, we distinguish between descriptive and causal analysis and identify the statistical issues that are relevant in each case. Second, because health data invariably derive from complex sample surveys, we consider the consequences of sample design for estimation and inference. To illustrate, we use a variety of methods to conduct multivariate analysis of child nutritional status in Vietnam. In the following chapter, we present some of the estimators most commonly used in analysis of health data.

Descriptive versus causal analysis

Descriptive analysis

As always, the appropriate statistical approach depends on the question to be answered. If the analyst is interested in simply describing SES-related inequality in health or health care, then statistical modeling issues are irrelevant. The analyst simply wants to describe how health varies with SES, conditional on other factors such as age, gender, and so on. Ordinary least squares (OLS) can be used to describe how the mean of health varies with SES, conditional on whatever factors the analyst wants to control for. The more variables the analyst controls for, the finer is the description of the relation between health and SES. Issues of omitted variable bias and endogeneity are not relevant. Of course, such simplicity comes at a price. The analyst cannot place any causal interpretation on the estimates. A significant OLS coefficient does not mean that SES has an effect on health, even if the analyst controls for a multitude of observable covariates. It simply means that health is observed to vary as SES varies. There is inequality.

The standardization and the decomposition methods covered in chapters 5 and 15 and in chapters 12 and 13, respectively, are examples of exploratory, or even explanatory, but still descriptive analysis. They are used to describe the distribution, primarily the mean, of health or health care conditional on SES, age, gender, and so forth.

Causal models

If the analyst wants to draw causal inferences, then the approach has to move from a descriptive one to a modeling approach. Causal relationships can arise through a number of pathways. Models and estimators vary in sophistication with the degree of detail of the causal relationship the analyst is aiming to uncover. For example, maternal education can affect child health either directly, through knowledge of healthy behavior, or indirectly, through preferences for child health. If the analyst is interested simply in whether educating women is an effective means of raising child health, irrespective of the mechanism through which it works, then the statistical model, and estimator, can be quite simple. A reduced form approach (see below) is adequate. However, if the analyst wants to establish whether educated mothers are better able to raise healthy children, abstracting from preference effects, then the model, as well as the methods, has to be more sophisticated. A structural model (see below) must be developed and estimated.

The household production model (Becker 1964, 1965) provides a useful framework for causal analysis of health variations (Grossman 1972a, 1972b; Rosenzweig and Schultz 1982, 1983; Schultz 1984; Wagstaff 1986). According to this approach, health, which is of intrinsic value, is "produced" by the household through the input of time and goods, such as food and medical care. The household selects such inputs given its members' physiological predispositions to good/bad health. These health endowments are observable to the household but not to the analyst. As a consequence, regressing outcomes, such as health, on inputs, such as medical care, will not render unbiased estimates of the causal impact of the latter because both the inputs and the outcomes reflect the values of the health endowments.

The most popular empirical strategy is to estimate reduced form demand relations. That is, to regress health outcomes on (exogenous) determinants of health inputs, for example, medical care prices. The resulting coefficients reflect both "technological" relationships between inputs and outcomes, and household preferences for health relative to other "goods." From such a reduced form health function it is not possible to conclude anything about the technological impact of a variable on health. For example, the relationship between female wage rates and child health reflects both the incentive effects of the wage on household time allocation and the effect of time use on child health. Nevertheless, for certain policy questions, reduced form estimation is appropriate. For example, say the analyst wants to know how population health will respond to an increased availability of medical care facilities, taking account both of the technological impact of medical care on health and the behavioral response with respect to utilization. Then, estimation of the reduced form correlation of area variations in medical facilities with individual levels of health is adequate.

If estimates of the health production technology are desired, then the problems of omitted variable bias and unobservable heterogeneity must be confronted. For example, regressing health on health care use, while omitting education, will give a biased estimate of the impact of health care in the likely instance that it is correlated with education. Resolution of the problem demands a sufficiently rich data set. The problem of heterogeneity bias arises from the unobservable health endowment, which induces correlation between the observable and unobservable components of a model of health determination. With cross-section data, correction of the resulting bias requires the availability of instruments, that is, variables that affect the health inputs but, conditional on these, not health itself. Appropriate instruments vary with the specific inputs being considered. At a general level, instruments used in the estimation of health production functions commonly come from geographic variation in market prices, from family endowments, for example, land rights, and from characteristics of public health programs at the regional level (Rosenzweig and Schultz 1983).

Instrumental variable (IV) estimation is fraught with danger. It is easy to claim that an endogenous regressor has been instrumented. It is somewhat more difficult to find a valid instrument. IV estimation should therefore be subjected to stringent testing (Bound, Jaeger, and Baker 1995; Staiger and Stock 1997). The variables proposed as instruments should be significant in a reduced form for the health input. Further, overidentification tests should be used to check whether exclusion of the proposed instruments from the health equation is justified.

Panel data have two important advantages with respect to estimation of health production functions. First, with data on the same individuals at different points in time, it is easier to account for the effect of unobservable health endowments, which generate much of the endogeneity problem. For example, the fixed effects estimator (see below) eliminates the time invariant unobservable effects and is consistent. The second important advantage of panel data is that they allow the time dynamics of health relationships to be investigated. The determination of health is essentially a dynamic process; health today reflects experiences of the past. For causal analysis of the determination of health, panel data are top priority.

Estimation and inference with complex survey data

Most surveys used for analysis of health sector inequalities in developing countries have complex sample designs. Typically, there is random sampling at some level or levels but there might be separate sampling from population subgroups (strata), groups of observations (clusters) may not be sampled independently, and there might be oversampling of certain groups. These three basic features of complex sample design—stratification, cluster sampling, and unequal selection probabilities—were introduced in chapter 2, in which we briefly discussed how the sample design should be taken into account in conducting inference with respect to population means. We now consider whether and how sample design should be taken into account in conducting multivariate analyses. A related issue, which we consider, is that of area effects—controlling for all observable determinants, area of residence exerts an independent effect on health. Such effects are characteristics of the population itself, but their sample importance depends on the sample design.

Stratified sampling

Samples can be stratified in a variety of ways. The design most typically employed in household surveys undertaken in developing countries, for example, the Living Standards Measurement Surveys, is standard stratified sampling. The population is divided into a relatively small number of strata—for example, urban/rural or large geographic regions. A random sample, of predetermined size, is selected independently from each of these strata. The sample proportions accounted for by each strata may or may not correspond to population proportions. In the case that they do not, the overall sample is not representative of the population and the issue of sample weights arises. This is a separate issue from stratification and is considered as such below.

If population means differ across strata, then predetermination of strata sample sizes reduces the sampling variance of estimators of these means. As a result, standard errors on estimates of population means, and other descriptive statistics, should be adjusted downward. In chapter 2, we demonstrated how to do this using the special routines for survey data available in Stata. It turns out that adjustment is not necessary in (nondescriptive) regression analysis and a wide variety of other multivariate modeling approaches, provided stratification is based on variables that are exogenous within the model (Wooldridge 2001, 2002). For example, say a sample stratified by urban/rural is used to estimate the determinants of child nutritional status, measured by height-for-age z-score. Provided, conditional on the regressors, unobservable determinants of height-for-age and of city dwelling are uncorrelated, the OLS estimator, for example, is consistent and efficient, and the usual standard errors are valid. In the likely presence of heteroscedasticity, the analyst would want to make the standard errors robust, but that is another issue. If stratification is based on an endogenous variable, however, then standard errors should be adjusted (Wooldridge 2002).

So, the need to adjust standard errors for stratification is situation specific. In practice, relative to simple standard errors, adjusting for stratification may inflate the standard errors. But with survey data, standard errors robust to heteroscedasticity, and possibly clustering (see below), will be required. Relative to those adjustments, the magnitude of that for stratification is usually modest and normally downward (see box 10.1). So, a conservative strategy is not to make any adjustment. If stratification is exogenous, there is no need for adjustment and, if endogenous, the adjustment will normally increase statistical significance.

It is often sensible to allow for intercept, and possibly slope, differences with respect to factors on which the sample is stratified. But this is in response to differences that exist in the population itself, not to the stratified design of the sample. For example, in many cases it is sensible to include an urban/rural dummy and to interact this with other regressors, to allow for differences in both the mean and responses between urban and rural locations, irrespective of whether the sample is stratified by urban/rural.

Computation Stata is the best package available for handling survey design issues. For the example presented in Box 10.1, OLS estimates with stratification adjusted SEs were obtained from the following:

```
svyset , strata(region)
svy, subpop(child): regr depvar varlist
```

where `strata(region)` instructs that the sample be stratified on the variable `region`, `depvar` denotes the dependent variable (height-for-age z-score [*–100] in

Box 10.1 *Standard Error Adjustment for Stratification Regression Analysis of Child Nutritional Status in Vietnam*

In the table below we present OLS coefficients from a regression of height-for-age z-scores (see chapter 4) using a sample of Vietnamese children under 10 years of age. The data are from the 1998 Vietnam Living Standards Survey (VLSS), which was stratified by 10 regions. The specification of the regression is based on that used by Wagstaff, van Doorslaer, and Watanabe (2003) (see also chapter 13). The dependent variable is actually the negative of the z-score (multiplied by 100), such that a positive coefficient indicates a negative correlation with height.

In addition to the OLS point estimates, we present standard errors (SEs) calculated with various degrees of adjustment. Relative to simple OLS SEs (column 2), adjustment for stratification alone (column 3) tends to inflate the SEs appreciably, but not dramatically. In some cases, the adjustment is slightly downward. In no case does the adjustment change the level of significance of the coefficient. In this example, making the SEs robust to heteroscedasticity of general form (column 4) has a very similar effect to that of adjusting for stratification. Besides stratification, the VLSS has a cluster sample design. Adjusting SEs for cluster sampling but not stratification (column 5) has a greater impact than stratification adjustment. In all cases, as expected, the adjustment is upward and in two cases it actually changes the level of significance. Finally, we adjust for both stratification and clustering (column 6). Comparing columns 5 and 6, it is apparent, for this example, that given adjustment for clustering, the marginal impact of stratification adjustment is small. In most cases, but not all, this marginal adjustment is downward. In no case does adjustment for stratification change the level of significance relative to that obtained by adjusting for clustering alone.

For this example, adjusting standard errors for clustering appears to be more important than adjustment for stratification. Although care must be taken not to draw general conclusions from an example, this is consistent with what is generally found in empirical work.

OLS analysis of height-for-age z-scores (–100), Vietnam 1998 (children <10 years)*

		Standard errors				
	Co-efficient	*Un-adjusted*	*Strati-fication adjusted*	*Hetero. robust*	*Cluster adjusted*	*Strat. & cluster adj.*
Child's age (months)	3.70***	0.1986	0.2466	0.2470	0.2885	0.2872
Child's age squared (/100)	–2.38***	0.1554	0.1755	0.1758	0.1966	0.1957
Child is male	12.31***	3.2927	3.2708	3.2792	3.3649	3.2844
(log) hhold. consumption per capita	–37.85***	3.9843	4.1046	4.1116	5.4035	5.4582
Safe drinking water	–7.43	4.9533	4.8300	4.8441	9.1538	9.2098
Satisfactory sanitation	–15.53***	5.1009	4.8199	4.8326	**6.1202**	**6.0937**
Years of schooling of household head	–0.87*	0.4804	0.4770	0.4786	**0.7302**	**0.7188**
Mother has primary school diploma	–2.33	4.0598	4.1309	4.1397	6.1913	6.2438
Sample size	5218					

Note: Dependent variable is **negative** of z-score, multiplied by 100. Bold indicates a change in significance level relative to that using unadjusted standard errors. Regression also contains region dummies at the level of stratification. ***, ** and * indicate 1%, 5% and 10% significance according to unadjusted standard errors.

Source: Authors.

the example), *varlist* is a list of regressors, and `subpop(child)` requests that the model be estimated for children (`child=1`) only. Restricting the sample to children and then estimating the model would not give the correct (stratification-adjusted) SEs. Computation for cluster sample adjustment is given below.

Cluster samples

Cluster samples arise from a two-stage (or more) sampling process. In the first stage, groups (clusters) of households are randomly sampled from either the population or the strata. Typically, these clusters are villages or neighborhoods of towns and cities. In the second stage, households are randomly sampled from each of the selected clusters. An important distinction from stratification is that strata are selected deterministically, whereas clusters are selected randomly. A further difference is that typically strata are few in number and contain many observations, whereas clusters are large in number and contain relatively few observations.

As a result of this design, observations are not independent within clusters, although most probably they are across clusters. There is likely to be more homogeneity within clusters than there is across the population as a whole. Within clusters, correlation of both observable and unobservable factors across households can be expected. Although these correlations exist in the population, the sample design increases their sample presence relative to that of a simple random sample. Consequences and remedies depend on the nature of the within-cluster correlation. Much of the analysis is analogous to that of unobservable individual effects in a panel data setting.

Case 1: Exogenous cluster effects Consider the following model:

$$y_{ic} = \mathbf{X}_{ic}\boldsymbol{\beta} + \lambda_c + \varepsilon_{ic},\ E\left[\varepsilon_{ic} \mid \mathbf{X}_{ic}, \lambda_c\right] = E\left[\varepsilon_{ic}\right] = 0, \tag{10.1}$$

where *i* and *c* are household (individual) and cluster indicators, respectively; $\mathbf{X}_{ic}$ is a vector of regressors; λ_c are cluster effects; and ε_{ic} idiosyncratic disturbances. If we assume that the cluster effects are independent of the regressors $\left(E\left[\lambda_c \mid \mathbf{X}_{ic}\right] = E\left[\lambda_c\right]\right)$, then so is the composite error $\left(u_{ic} = \lambda_c + \varepsilon_{ic}\right)$. This is the random effects model.

Conventional point estimators, for example, OLS, probit, and so on, depending on the nature of the dependent variable, are consistent, but inefficiency arises from the cluster-induced correlation in the composite errors which, in addition, requires adjustment of the standard errors. One option is to accept inefficiency and simply adjust the standard errors. In Stata, this is easily implemented through the option `cluster(varname)`, where `varname` defines the clusters. This option, which is available for most estimators, will adjust both for within-cluster correlation and for heteroscedasticity of unknown form.

An alternative strategy is to pursue efficiency by estimating the within-cluster correlation and taking account of this in estimation of the model parameters. In the linear case, for example, the analyst would use generalized least squares (GLS). A Lagrange multiplier test can be used to test the null that the cluster effects are insignificant and OLS is efficient (Wooldridge 2002). In the case of a binary discrete dependent variable, the analyst can estimate the random effects probit.

Case 2: Endogenous cluster effects The model is equation 10.1, but we relax the assumption of independence between the cluster effects and the regressors. That is, we allow $\left(E\left[\lambda_c \mid \mathbf{X}_{ic}\right] \neq E\left[\lambda_c\right]\right)$. This is the fixed effects model.

For example, in a regression of individual health on health service utilization, we would expect the latter to be correlated with the unobservable cluster-specific quality of those services. In instances in which there is such dependence, regressors are correlated with the composite error and standard estimators are inconsistent. The analyst must purge the cluster effects from the composite error. In a linear context, either include cluster dummies or, equivalently, transform the data by taking differences from within cluster means (i.e., the within-groups estimator). In a binary discrete choice context, the analyst can use the fixed effects logit estimator (Wooldridge 2002) (see chapter 11). Once the cluster effects have been purged from the composite error, there is no need to adjust standard errors for clustering (providing the linear specification of the cluster effects is correct). Adjustment for heteroscedasticity is likely to be a good idea.

The analyst can choose between the random and fixed effects models by reference to a Hausman test of the null of independence between the cluster effects and the regressors (Wooldridge 2002).

Box 10.2 *Taking Cluster Sampling into Account in Regression Analysis of Child Nutritional Status in Vietnam*

We continue with an examination of height-for-age z-scores of Vietnamese children using the 1998 VLSS, which has a cluster sample design. Cluster samples were actually drawn at two levels in this survey. At the first level, within each stratum a random sample of communes was drawn with probability of selection proportional to commune population size. Communes therefore represent the primary sampling units. Within each of the 194 selected communes, two villages/blocks were randomly selected with selection probabilities again proportional to population size. Finally, within each village/block a random sample of 20 households was selected. With this sample design, clusters could be defined at the level of the commune, village/block, or both. For simplicity, we will define clusters at the commune level.

We take three approaches to the cluster sample issue: OLS with standard errors adjusted for within-cluster correlation, random effects, and fixed effects. In each case, standard errors are made robust to heteroscedasticity of general form. The results of the respective z-score regressions are given in the table below.

Comparing the point estimates, it is apparent that the choice of estimator makes little difference for regressors that are clearly individual specific, but there is greater sensitivity in estimates for regressors that can be expected to display stronger within-commune correlation. So, for example, the point estimates for age and gender are near constant across the estimators. The estimate for household consumption is more sensitive, the effect weakening as we move from OLS, which takes no account of commune effects in the point estimates, to fixed effects, which purge the commune effects. This pattern is even more pronounced for indicators of safety of drinking water and sanitation, which can be expected to display fairly limited within-commune variation.

In general, standard errors are smaller for random and fixed effects. This is expected because these methods take into account the cluster effects in the (point) estimation and do not have to inflate the standard errors to allow for these correlated effects. In this example, however, the choice of estimator makes very little difference to levels of significance, reflecting the strength of the effect of some of the regressors.

The Lagrange multiplier test on the random effects model confirms that commune effects are, indeed, important. The Hausman test rejects the assumption of zero correlation between the commune effects and the regressors, indicating the superiority of the fixed effects estimator in this case.

Box 10.2 continued *Regression Analysis of Height-for-Age z-Scores (*-100), Vietnam 1998 (children <10 years)*

	OLS		*Random effects*		*Fixed effects*	
	Coeff.	*Cluster adjusted SE*	*Coeff.*	*Robust SE*	*Coeff.*	*Robust SE*
Child's age (months)	3.72***	0.2917	3.74***	0.2451	3.78***	0.2430
Child's age squared (1100)	–2.40***	0.1987	–2.40***	0.1742	–2.44***	0.1732
Child is male	12.26***	3.4527	12.19***	3.2394	12.97***	3.2443
(log) hhold. consumption p.c.	–50.93***	5.1149	–43.17***	4.0778	–30.37***	4.6090
Safe drinking water	–12.55	8.6438	–7.93	4.8984	–2.75	5.4247
Satisfactory sanitation	–22.90***	5.6974	–19.39***	4.8446	–9.77**	4.9364
Years of schooling of HoH	–0.39	0.6628	–0.33	0.4828	–0.55	0.5081
Mother has primary school diploma	2.67	5.3187	1.71	4.1140	1.74	4.3186
Intercept	445.00***	44.5600	377.01***	32.1941	276.19***	35.0991
Sample size	5,218	R^2	0.1527	B-P LM	485.84	(0.000)
				Hausman	50.54	(0.000)

Note: Dependent variable is **negative** of z-score, multiplied by 100.
SE = standard error, Robust SE-robust to general heteroskedasticity.
B-P LM = Breusch-Pagan Lagrange Multiplier test of significance of commune effects (*p*-value).
Hausman = Hausman test of random versus fixed effects (*p*-value).
***, ** & * indicate significance at 1%, 5% & 10%, respectively.
Source: Authors.

Computation Results such as those in box 10.2 can be generated in Stata as follows. For OLS with cluster corrected SEs,

```
svyset commune
svy, subpop(child): regr depvar varlist
```

where the `svyset` command instructs that clusters are defined by the variable `commune`. If the analysis were not restricted to part of the sample, the appropriate cluster corrected standard errors could be obtained from the following:

```
regr depvar varlist, cluster(commune)
```

To adjust SEs for both clustering and stratification, simply set the survey parameters appropriately,

```
svyset commune, strata(region)
```

and then run the `svy: regr` command as above. This was used to generate the SEs in the final column of the table in box 10.1. We do not adjust OLS SEs in the table in box 10.2 for stratification because the random and fixed effects estimators do not allow for that. Random effects estimates are obtained most easily from Stata's linear panel data estimator,

```
xtreg depvar varlist, re i(commune)
```

where `i(commune)` instructs to allow for common effects within each category of the variable `commune`. The Breusch-Pagan test statistic is obtained by following the command above with `xttest0`. To obtain (heteroskedasticity) robust SEs, as in the example, the analyst can implement the random effects (GLS) estimator through OLS on transformed data and request robust SEs. First, run the random effects estimator as above, and save the estimates of the variances of the error components,

```
scalar define sigma_e=e(sigma_e)^2
scalar define sigma_u=e(sigma_u)^2
```

Next calculate the variable that will be used to transform the data,

```
sort commune
by commune: gen T=_N
gen theta=1-sqrt(sigma_e/(sigma_e +(T*sigma_u)))
```

where the first two command lines generate a variable indicating the number of observations within each commune, and the third line gives the transformation variable. Now generate the quasi mean deviations (i.e., deviations from the transformed mean) for the dependent variable and each regressor,

```
local vbls "depvar varlist"
foreach var of local vbls {
  by commune: egen m_`var'=mean(`var')
  gen t_`var'=`var'-theta*m_`var'
}
```

Generate the variable from which the intercept will be estimated and run OLS,

```
gen intercept=1-theta
local vars "t_depvar t_var1 t_var2...  ."
regr `vars' intercept, noconstant robust
```

where the local `vars` contains the names of the transformed dependent variable and the regressors, `noconstant` requests that the regression be estimated without a constant, and `robust` requests heteroscedasticity robust SEs.

Fixed effects estimates can be obtained from the panel data command:

```
xtreg depvar varlist, fe i(commune)
```

Or to obtain the same point estimates but robust SEs, use the following:

```
areg depvar varlist, absorb(commune) robust
```

which requests OLS on deviations from `commune` specific means, that is, the within-groups or fixed-effects estimator.

The Hausman test statistic can be computed by the following:

```
xtreg depvar varlist, fe i(commune)
est store fixed
xtreg depvar varlist, re i(commune)
hausman fixed
```

Explaining community effects

The strategies outlined above for dealing with cluster samples are appropriate when the analyst is interested exclusively in the determinants of health/health care at the individual level. In this case, the cluster sample design is a problem to be

overcome. But cluster, or community, effects can be more than nuisance parameters. With respect to health inequality, for example, area variations in health, and their determinants, are of genuine interest. Not least because implementation of public health policies at the community, rather than the individual, level is often more feasible. In this case, a cluster sample design is an advantage rather than a problem. It facilitates examination of cross-community differences in health and their determinants, particularly if the household survey is accompanied by a community-level survey providing information on characteristics of the community.

Options for the analysis of community effects from individual-level data depend on whether the effects are exogenous or endogenous.

Case 1: Exogenous cluster (community) effects In this case, the analyst can explore the determinants of area variation in health outcomes or utilization by including community-level variables, if available, in the model. Define $\lambda_c = \mathbf{Z}_c\gamma + \lambda_c^*$, where $\mathbf{Z}_c$ are observable community-level factors, for example, health care facilities and personnel, quality of water provision and sewage, prices, and so forth and substitute this definition into equation 10.1. The (rewritten) model is as follows:

$$y_{ic} = \mathbf{X}_{ic}\beta + \mathbf{Z}_c\gamma + \lambda_c^* + \varepsilon_{ic},\ E\left[\varepsilon_{ic} \mid \mathbf{X}_{ic}, \mathbf{Z}_c, \lambda_c^*\right] = E\left[\varepsilon_{ic}\right] = 0. \tag{10.2}$$

To maintain the assumption of exogenous community effects, and therefore consistency (but not efficiency) of standard estimators, we now need the unobservable community effects (λ_c^*) to be independent of both the individual- and community-level regressors (i.e., $E\left[\lambda_c^* \mid X_{ic}, Z_c\right] = E\left[\lambda_c^*\right]$) (Wooldridge 2002). This is likely to be a stronger assumption than that placed on model 10.1 above. Assuming that the observable community factors capture all of the community effect, ($\lambda_c^* = 0, \forall c$) is even stronger. Excepting the latter restrictive case, standard errors still have to be adjusted (upward) for correlation induced by the (unobservable) community effects. However, the efficiency loss from employing OLS, for example, in this setting may not be large (Deaton 1997).

This random effects model is known as the hierarchical model in some fields (see, e.g., Rice and Jones [1997]). Although the models are equivalent, the hierarchical approach places more emphasis on decomposition of the overall variance into that arising at the individual and the community level. This approach is particularly useful in cases in which the analyst wants to focus on such a distinction between individual- and community-level effects.

Case 2: Endogenous cluster (community) effects In cases in which the (unobservable) community effects are correlated with individual-level regressors, it is not possible to include community-level variables in a model to be estimated from a single cross section. With a dummy variable approach, the community variables would be perfectly correlated with the community dummies. With a fixed-effects approach, community variables would be wiped out of the model along with the unobservable community effects. If one has panel data, then these problems are avoided provided there is sufficient across-time variation in the community-level variables. With a single cross section, a feasible two-stage approach in a linear context is to estimate a fixed-effects model, obtain estimates of the community effects, and then regress these on community-level variables. In the first stage, the bias arising from the community effects is removed from the individual-level analysis of, say, health determination. In the second stage, sources of community variation in health are examined.

In box 10.3 we continue with the example of child nutritional status in Vietnam, examining the sources of community-level variation, assuming in turn exogenous and endogenous community effects.

Box 10.3 *Explaining Community-Level Variation in Child Nutritional Status in Vietnam*

In box 10.2 we saw that commune effects are an important source of variation in height-for-age z-scores of Vietnamese children. The VLSS offers the opportunity to uncover factors underlying these commune effects through the examination of data from commune-level surveys that accompanied, and can be linked to, the household survey data. For demonstration purposes, we limit attention to the characteristics of commune health centers (CHCs). The analysis is necessarily restricted to children living in rural areas and small towns because the commune surveys were conducted in those areas only.

Again we compare OLS, random effects, and fixed effects. In the case of OLS and random effects, the estimates are obtained from entering the CHC characteristics directly into the individual-level regressions. We present, in the table below, the estimates for the CHC regressors only. Estimates for the individual-level regressors are similar to those given in the table in box 10.2. For the fixed-effects model, we take the two-stage approach outlined above. In the table, we present results from the second stage regression of the estimated commune effects on the CHC characteristics. The first stage estimates are similar to those in the table in box 10.2.

The results indicate a lower prevalence of stunting in communes in which the CHC has electricity, a sanitary toilet and, at marginal significance, a child growth chart. The number of inpatient beds available in a CHC and, at lower significance, the employment of a doctor is positively correlated with the prevalence of stunting. These latter results may reflect the targeting on resources in the communes of greatest need.

Analysis of Commune-Level Variation in Height-for-Age z-Scores (–100), Rural Vietnam 1998 (children <10 years)*

	OLS		*Random effects*		*2nd-stage fixed effects*	
Commune health center vbls.	*Coeff.*	*Cluster adj. SE*	*Coeff.*	*Robust SE*	*Coeff.*	*SE*
Vitamin A available ≥ 1/2 time	–10.11	6.6530	–6.86143	6.5927	–8.27114	6.7506
Has electricity	–38.79***	11.4558	–50.56***	12.1861	–45.34***	10.7991
Has clean water source	9.57	7.6534	7.2341	8.4061	7.0070	8.7610
Has sanitary toilet	–27.53***	7.0928	–24.50***	7.6694	–24.30***	7.8715
Has child growth chart	–13.85*	7.2046	–10.2623	7.5879	–11.732	7.6292
Number of inpatient beds	1.52*	0.8298	2.12**	0.9242	2.09**	0.9744
Has a doctor	11.39	6.9765	9.6255	7.1834	10.1856	7.5207
Intercept	371.89***	48.8784	344.71***	41.5639	279.13***	41.6264
Sample size 4,099	R^2	0.1313	B-P LM	248.42 (0.0000)		

Note: Dependent variable is **negative** of z-score, multiplied by 100. OLS & random effects = Coefficients on commune-level regressors only are presented. 2nd stage fixed effects = Estimated commune effects from fixed effects regressed on commune vbls.

SE = standard error, robust SE = robust to general heteroskedasticity.

B-P LM = Breusch-Pagan Lagrange Multiplier test of significance of community effects (*p*-value).

***, ** and * indicate significance at 1%, 5% and 10%, respectively.

Source: Authors.

Computation OLS and random effects estimates and standard errors can be generated exactly as above, with the inclusion of community-level regressors. The two-stage fixed-effects approach can be implemented in Stata by first running the linear (fixed effects) panel estimator and saving the predicted commune effects:

```
xtreg depvar varlist, fe i(commune)
predict ce, u
```

where `ce` is the variable name given to the predicted commune effects, `u`. OLS regression of these commune effects on commune-level variables (*varlist2*) is most easily implemented by using the between-groups panel estimator:

```
xtreg ce varlist2, be i(commune)
```

Sample weights

The probability of observing an individual in a survey may differ from the probability that the individual is randomly selected from the population. There are a number of reasons for this. The survey may be stratified, with strata sample proportions differing from respective population proportions. For example, there may be oversampling of the urban population. Besides sample design, differential nonresponse will lead to a sample that is not representative of the population. For those reasons, survey data typically come with a set of sample weights that, for each observation, indicate the (inverse of the) probability of being a sample member. In a standard stratified sample with differential sampling by strata, weights or expansion factors are given by the ratio of the population size of each stratum to its sample size.

Sample weights must be applied to obtain unbiased estimates of population means, concentration indices, and so forth and correct standard errors for these estimates. Application of the weights allows for the fact that observations with lower sample probabilities represent a greater number of (similar) individuals in the population. With respect to multivariate analysis, the case for applying sample weights is less clear-cut. In part, the appropriateness of weighting depends on the objective of the analysis. As we stressed at the beginning of this chapter, appropriate methods depend on the purpose of the analysis. If regression is being used simply as a descriptive device, and not for estimation of behavioral parameters, then weights should be applied (Deaton 1997). The regression function describes the means of one variable conditional on others. Application of sample weights will ensure that the conditional means estimated are those that would have been estimated from a simple random sample of the population. In this case, weights are applied for the same reason they are used in univariate analysis. For example, in standardization exercises (see chapters 5 and 15), regression is used simply to obtain conditional means, and it would be appropriate to apply sample weights.

If the purpose of the analysis is more ambitious—to uncover causal relationships—then the crucial factor determining whether weights need to be applied in estimating the model parameters is the source of differences between sample and population proportions. If proportions differ because of selection on factors that are exogenous within the model under consideration, then there is no need to apply weights. Unweighted estimators are consistent and more efficient than weighted counterparts (Wooldridge 2002). Usual or, in the presence of heteroscedasticity,

robust standard errors are valid. However, if selection is on endogenous factors, then a weighted estimator is required for consistency (Wooldridge 2002). In the case of the linear model, for example, weighted least squares could be used with the data weighted by the inverse of sample probabilities. If the sample weights derive from stratification with differential sampling by strata, then standard errors need to be calculated taking account of both the weights and the stratification. Alternatively, if there are sample weights but not stratification, then (robust) standard errors are calculated by applying the usual formula to the weighted data.

So, as with sample stratification, the need to take account of sample weights in estimation is situation specific. Consider a model of health determination to be estimated from a survey that oversamples the urban population. If, conditional on all regressors, unobservable determinants of health are uncorrelated with city dwelling, then there is no need to apply weights. Conditioning on an urban dummy is sufficient. In this example, the exogeneity assumption might be considered reasonably weak, although its validity would be challenged if migration were strongly influenced by health status. If, however, there were differential sampling by health itself, say the sick were oversampled, then sample weights would need to be applied.

The discussion above assumes parameter homogeneity across the differentially sampled groups. There might be different (conditional) group means, but that is easily dealt with through the inclusion of dummy variables. A more serious problem is differences in slope parameters across groups. Consider the following model:

$$y_{is} = \mathbf{X}_{is}\beta_s + \varepsilon_{is} \tag{10.3}$$

where *i* and *s*, respectively, indicate individual and group, for example, urban/rural, gender, ethnicity, and so on, and the parameter vector $\boldsymbol{\beta}_s$ is indexed on *s*, indicating parameter heterogeneity across groups. If differences in parameters across groups are of inherent interest, then the analyst can estimate either a separate model for each group or a single model with dummies for each group and their interactions with other regressors. The former is more general. In both cases, parameter homogeneity can be tested by standard methods.

For various reasons, the analyst might want an estimate of the average effect across the population. Such an average might be defined as follows: $\boldsymbol{\beta} = \frac{1}{N}\sum_{s=1}^{S} N_s \boldsymbol{\beta}_s$, that is, the weighted average of the group-specific parameters with weights provided by the population group proportions $\left(\frac{N_s}{N}\right)$ (Deaton 1997). If degrees of freedom are not a problem, this parameter can be consistently estimated by applying OLS to each sector to obtain estimates of the sector-specific parameters, $\hat{\boldsymbol{\beta}}_s$, and taking the population-weighted average of these. For degrees of freedom reasons or otherwise, it is often preferred to estimate the average parameter directly from one regression. In the case in which sample group proportions do not correspond to population proportions, it might be anticipated that unweighted OLS on the whole sample will not be consistent for the average parameter defined. That is correct. It is reasonable to ask whether sample weights can solve the problem. The answer is "no." Weighted regression will give an estimate that corresponds to that which would be obtained from a simple random sample, but that is not consistent for the population average parameter, apart from the extreme case in which regressor values are identical across all groups (Deaton 1997).

Box 10.4 *Applying Sample Weights in Regression Analysis of Child Nutritional Status in Vietnam*

We reproduce the analysis of box 10.2, but with sample weights applied to all estimators. One other difference is that the OLS standard errors are adjusted for stratification because, within a modeling approach, the logic for applying sample weights and for adjusting for stratification is the same. That is, selection on an endogenous variable.

By comparing the estimates presented in the table below with those given in the table in box 10.2, it is apparent that the application of sample weights makes very little difference to the results. A possible explanation is that the application of sample weights is not necessary in this particular example. That is, differential sampling is exogenous, and so the unweighted estimators are consistent.

Weighted Regression Analyses of Height-for-Age z-Scores Vietnam 1998 (children <10 years)

		OLS		*Random effects*		*Fixed effects*	
		Coeff.	*Adjusted SE*	*Coeff.*	*Robust SE*	*Coeff.*	*Robust SE*
Child's age (months)		3.90***	0.3218	3.90***	0.2652	3.91***	0.2642
Child's age squared (/100)		–2.51***	0.2206	–2.50***	0.1875	–2.51***	0.1875
Child is male		14.86***	3.5718	14.56***	3.3595	14.89***	3.3731
(log) hhold. consumption p.c.		–50.14***	5.5131	–40.67***	4.3511	–26.05***	5.0196
Safe drinking water		–12.16	10.2770	–6.92	5.1624	–2.07	5.6079
Satifactory sanitation		–22.01***	5.9503	–19.81***	5.3653	–10.48*	5.4439
Years of schooling of HoH		–0.21	0.7355	–0.15	0.5122	–0.42	0.5363
Mother has primary school diploma		3.62	5.6510	3.04	4.2925	2.19	4.4958
Intercept		428.15***	48.9827	347.47***	34.9686	236.12***	38.5646
Sample size	5,218	R^2	0.1496	R^2	0.4320	R^2	0.2457

Note: Dependent variable is **negative** of z-score, multiplied by 100.

Adjusted SE = standard error adjusted for clustering and stratification and robust to heteroskedasticity.

Robust SE = standard error robust to general heteroskedasticity.

***, ** and * indicate significance at 1%, 5% and 10%, respectively.

Source: Authors.

The issue here is parameter heterogeneity, which exists in the population, and is not simply a feature of sample design. Sample weights cannot be used to address an issue that arises from the population itself.

Sensitivity of estimates and their standard errors to the application of weights is examined in box 10.4.

Further reading

Deaton (1997) is a wonderfully useful guide to the analysis of survey data. Wooldridge (2002) and Cameron and Trivedi (2005) are both excellent, comprehensive textbooks covering the relevant econometric theory.

References

Becker, G. 1964. *Human Capital: A Theoretical and Empirical Analysis With Special Reference to Education.* New York: National Bureau of Economic Research.

———. 1965. "A Theory of the Allocation of Time." *Economic Journal* 75: 492–517.

Bound, J., D. Jaeger, and R. Baker. 1995. "Problems with Instrumental Variables Estimation When the Correlation between the Instruments and the Endogenous Explanatory Variables Is Weak." *Journal of American Statistical Association* 90: 443–450.

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

Deaton, A. 1997. *The Analysis of Household Surveys: A Microeconometric Approach to Development Policy.* Baltimore, MD: Johns Hopkins University Press.

Grossman, M. 1972a. *The Demand for Health: A Theoretical and Empirical Investigation.* New York: National Bureau of Economic Research.

Grossman, M. 1972b. "On the Concept of Health Capital and the Demand for Health." *Journal of Political Economy* 80: 223–55.

Rice, N., and A. M. Jones. 1997. "Multilevel Models and Health Economics." *Health Economics* 6: 561–75.

Rosenzweig, M. R., and T. P. Schultz. 1982. "Market Opportunities, Genetic Endowments and Infrafamily Resource Allocation: Child Survival in Rural India." *American Economic Review* 72:803–15.

———. 1983. "Estimating a Household Production Function: Heterogeneity, the Demand for Health Inputs, and Their Impact on Birth Weight." *Journal of Political Economy* 91(5): 723–46.

Schultz, T. P. 1984. "Studying the Impact of Household Economic and Community Variables on Child Mortality." *Population and Development Review* 10: 215–35.

Staiger, D., and J. H. Stock. 1997. "Instrumental Variables Regression with Weak Instruments." *Econometrica* 65(3): 557–86.

Wagstaff, A. 1986. "The Demand for Health: Some New Empirical Evidence." *Journal of Health Economics* 5(3): 195–233.

Wagstaff, A., E. van Doorslaer, and N. Watanabe. 2003. "On Decomposing the Causes of Health Sector Inequalities, with an Application to Malnutrition Inequalities in Vietnam." *Journal of Econometrics* 112(1): 219–27.

Wooldridge, J. M. 2001. "Asymptotic Properties of Weighted M-Estimators for Standard Stratified Samples." *Econometric Theory* 17: 451–70.

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