Breusch and Pagan (1979), in which the squared OLS residuals are regressed on variables that are thought to be likely candidates for causing the heteros-kedasticity, usually including the levels, squares, and interactions of the original explanatory variables. Indeed, as is easily checked, this is the correct specification if /3 in (20) is taken to be distributed randomly in the population. Under the assumption that the original regression errors are normally distributed, the null of homoskedasticity implies that the explained sum of squares of this supplementary regression will be distributed as x^2 with degrees of freedom equal to the number of regressors in the supplementary regression. This test is closely related to the information matrix test proposed by White (1980).

It is clearly good practice to calculate and report standard errors and other test statistics that are robust to departures from homoskedasticity. Furthermore, my own experience suggests that it is difficult to pass the Breusch—Pagan test in practical applications, and that heteroskedasticity is usually revealed not just by this test, but by others, such as the quantile regression techniques discussed below. That said, the heteroskedasticity-consistent standard errors and tests are rarely very different from those given by the standard formulas. An upward correction of about 30 percent to standard errors appears to be common, and this correction would not normally lead to startling differences in inference.

2.1.2. Clustering and linear regression

In Section 1 above, I showed that when observations within survey clusters are correlated, survey cluster sampling requires a revision of the formula for the standard error of an estimated mean. In particular, the usual variance, which is the population variance divided by the sample size, has to be multiplied by the Kish design effect (16), which depends on the average number of observations per cluster and the size of the intracluster correlation coefficient. Similar considerations apply to the estimation of linear regressions when there are grounds for believing that the errors are correlated within clusters. The fact that the sample is clustered does not in itself imply that there must be a non-zero intracluster correlation once other explanatory variables have been taken into account. However, survey clusters in rural areas in LDCs are typically geographically dispersed villages, so that there are likely to he unobserved communalities that are shared between households in the same village, and that differentiate them from those in other villages. Note too that there may be intrahousehold correlations between households beyond the cluster levels, for example across provinces or regions, correlations that could come from ethnic factors, from the way in which markets operate, or from the way that the government allocates services across administrative areas.

To illustrate the issues, I shall suppose that the survey is clustered, that there