This same situation may arise, as noted above, in the context of other individual-level series. Earnings may be more variable for self-employed workers, or those who depend on commissions or tips than salaried workers. In the context of firm data, we might expect that profits (or revenues, or capital investment) might be much more variable in some industries than others. Capital-goods makers face a much more cyclical demand for their product than do, for example, electric utilities.

Testing for heteroskedasticity between groups of observations

How might we test for groupwise heteroskedasticity? By the assumption that each group's regression equation satisfies the classical assumptions (including that of homoskedasticity), the s^2 computed by [R] **regress** is a consistent estimate of the group-specific variance of the disturbance process. For two groups, an F-test may be constructed, with the larger variance in the numerator; the degrees of freedom are the residual degrees of freedom of each group's regression. This can easily be accomplished if both groups' residuals are stored in a single variable, with a group variable indicating group membership (in this case 1 or 2). The third form of [R] **sdtest** may then be employed, using the `by(groupvar)` option, to conduct the F-test.

What if there are more than two groups across which we wish to test for equality of disturbance variance: for instance, a set of ten industries? We may then employ the `robvar` command ([R] **sdtest**), which like `sdtest` expects to find a single variable containing each group's residuals, with a group membership variable identifying them. The `by(groupvar)` option is employed here as well. The test conducted is that of Levene (1960), labeled as W_0, which is robust to non-normality of the error distribution. Two variants of the test proposed by Brown and Forsythe (1992) which employs more robust estimators of central tendency (e.g., median rather than mean), W_{50} and W_{10}, are also computed.

We illustrate groupwise heteroskedasticity with state-level data from the `NEdata` dataset. These data are comprised of one observation per year for each of the six U.S. states in the New England region for 1981–2000. Descriptive statistics are generated by [R] **summarize** for `dpipc`, state disposable personal income per capita.

```
. use NEdata, clear
. summarize dpipc
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
       dpipc |       120    18.15802    5.662848   8.153382   33.38758
```

We fit a linear trend model to `dpipc` by regressing that variable on `year`. The residuals are tested for equality of variances across states with `robvar`.

```
. regress dpipc year
      Source |       SS       df       MS              Number of obs =     120
-------------+------------------------------           F(  1,   118) =  440.17
       Model |  3009.33617     1  3009.33617           Prob > F      =  0.0000
```

```
    Residual |  806.737449   118  6.83675804           R-squared     =  0.7886
-------------+------------------------------           Adj R-squared =  0.7868
       Total |  3816.07362   119  32.0678456           Root MSE      =  2.6147

------------------------------------------------------------------------------
       dpipc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        year |   .8684582   .0413941    20.98   0.000     .7864865    .9504298
       _cons |  -1710.508   82.39534   -20.76   0.000    -1873.673   -1547.343
------------------------------------------------------------------------------

. predict double eps, residual

. robvar eps, by(state)

            |      Summary of Residuals
      state |        Mean   Std. Dev.       Freq.
------------+------------------------------------
         CT |    4.167853   1.3596266          20
         MA |    1.618796   .86550138          20
         ME |  -2.9841056   .93797625          20
         NH |   .51033312   .61139299          20
         RI |   -.8927223   .63408722          20
         VT |  -2.4201543   .71470977          20
------------+------------------------------------
      Total |  -6.063e-14   2.6037101         120

W0  = 4.3882072   df(5, 114)     Pr > F = .00108562

W50 = 3.2989849   df(5, 114)     Pr > F = .00806752

W10 = 4.2536245   df(5, 114)     Pr > F = .00139064
```

The hypothesis of equality of variances is soundly rejected by all three `robvar` test statistics, with the residuals for Connecticut, Massachusetts and Maine possessing a standard deviation considerably larger than those of the other three states.

Estimation with FGLS

If we discern that different groups of observations have different error variances, we may apply the generalized least squares estimator using analytical weights, as described above in Section 6.2.1. In the groupwise context, we define the analytical weight (`aw`) series as a constant value for each observation in a group. That value is calculated as the estimated variance of that group's OLS residuals. Using the residual series calculated above, we construct an estimate of its variance for each New England state with [R] **egen** and generate the analytical weight series:

```
. bysort state: egen sd_eps = sd(eps)

. generate double gw_wt = 1/sd_eps^2

. tabstat sd_eps gw_wt, by(state)

Summary statistics: mean
  by categories of: state

state |    sd_eps     gw_wt
------+--------------------
   CT |  1.359627  .5409545
   MA |  .8655014  1.334948
   ME |  .9379762  1.136623
```

```
      NH    .611393  2.675218
      RI   .6340872   2.48715
      VT   .7147098  1.957675

   Total   .8538824  1.688761
```

The [R] **tabstat** command reveals that the standard deviations of New Hampshire and Rhode Island's residuals are much more sizable than those of the other four states. We now reestimate the regression using FGLS employing the analytical weight series:

```
. regress dpipc year [aw=gw_wt]
(sum of wgt is   2.0265e+02)
      Source         SS       df       MS              Number of obs =     120
                                                       F(  1,   118) =  698.19
       Model   2845.55409     1  2845.55409           Prob > F      =  0.0000
    Residual   480.921278   118  4.07560405           R-squared     =  0.8554
                                                       Adj R-squared =  0.8542
       Total   3326.47537   119  27.9535745           Root MSE      =  2.0188

       dpipc       Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]

        year    .8444948   .0319602    26.42   0.000     .7812049    .9077847
       _cons    -1663.26   63.61705   -26.14   0.000    -1789.239   -1537.281
```

In comparison to the unweighted estimates' `Root MSE` of 2.6147, FGLS yields a considerably smaller value of 2.0188.

6.2.3 Heteroskedasticity in grouped data

We spoke in Section 6.2 above of a third case in which heteroskedasticity arises in cross-sectional data: where our observations are grouped or aggregated data, representing differing numbers of microdata records. This situation arises when the variables in our dataset are averages or standard deviations of groups' observations: for instance, a set of 50 U.S. state observations. Since we know the population of each state, we know precisely how much more accurate California's observation (based on 30+ million individuals) is than Vermont's (based on fewer than a million). This situation would also arise in the context of observations representing average attainment scores for individual schools or school districts, where we know that each school (or school district) has a different-sized student population. In these cases we know that heteroskedasticity will occur in the grouped or aggregated data, and are in the unique case of knowing Ω, since it depends only on the N_g underlying each observation.

You could consider this a problem of non-random sampling. In the first example above, when 30 million California records are replaced by one state record, an individual has very little weight in the average. In a smaller state, each individual would have a greater weight in her state's average values. If we want to conduct inference in terms of a national random sample, we must equalize those weights, leading to a heavier weight