Intro 10 — Fitting models with survey data

Description

Sometimes the data are not a simple random sample from the underlying population but instead are based on a complex survey design that can include stages of clustered sampling and stratification. Estimates produced by \texttt{sem} can be adjusted for these issues.

Adjustments for survey data are provided by \texttt{sem} and \texttt{gsem}.

Remarks and examples

Data obtained from surveys, properly treated, produce different point estimates because some observations represent a greater proportion of the underlying population than others. They also produce different standard errors because the observation-to-observation (sample-to-sample) variation is a function of the survey’s design.

To obtain survey-corrected results, you first describe the characteristics of the survey with \texttt{svyset}:

\begin{verbatim}
.svysset county [pw=samplewgt], fpc(n_counties) strata(states) || ///
    school, fpc(n_schools) || ///
    student, fpc(n_students)
\end{verbatim}

In the above, we are telling Stata that our data are from a three-stage sampling design. The first stage samples without replacement counties within state; the second, schools within each sampled county; and the third, students within schools.

Once we have done that, we can tell Stata to make the survey adjustment by prefixing statistical commands with the \texttt{svy:} prefix:

\begin{verbatim}
.svy: regress test_result teachers_per_student sex ...
\end{verbatim}

Point estimates and standard errors will be adjusted.

You can use the \texttt{svy:} prefix with \texttt{sem} and \texttt{gsem}:

\begin{verbatim}
.svy: sem (test_result<-...) ... (teachers_per_student->...) ...
\end{verbatim}

See the \textit{Stata Survey Data Reference Manual} for more information on this. From a survey perspective, \texttt{sem} is not different from any other statistical command of Stata. When \texttt{gsem} is used to fit a multilevel model, stage-level sampling weights specified in the \texttt{svyset} command are applied to the corresponding hierarchical group level in the model.

Once results are estimated, you do not include the \texttt{svy:} prefix in front of the postestimation commands. You type, for instance,

\begin{verbatim}
.estat eqtest ...
\end{verbatim}

You do not type \texttt{svy: estat eqtest} ....

Some postestimation procedures you might ordinarily perform can be inappropriate with survey estimation results. This is because you no longer have a sample likelihood value. The postestimation command \texttt{lrtest} is an example. If you attempt to use an inappropriate postestimation command, you will be warned.
. lrtest ...
\texttt{lrtest} is not appropriate with survey estimation results
r(322);

\section*{Also see}

[SEM] \textbf{Intro 9} — Standard errors, the full story

[SEM] \textbf{Intro 11} — Fitting models with summary statistics data (sem only)

[SVY] \textit{Stata Survey Data Reference Manual}