Title

estat — Postestimation statistics for survey data

Syntax

Survey design characteristics

`estat svyset`

Design and misspecification effects for point estimates

`estat effects` [, *estat_effects_options*]

Design and misspecification effects for linear combinations of point estimates

`estat lceffects` *exp* [, *estat_lceffects_options*]

Subpopulation sizes

`estat size` [, *estat_size_options*]

Subpopulation standard-deviation estimates

`estat sd` [, *estat_sd_options*]

Singleton and certainty strata

`estat strata`

Coefficients of variation for survey data

`estat cv` [, *estat_cv_options*]

Goodness-of-fit test for binary response models using survey data

`estat gof` [, *estat_gof_options*]

Display covariance matrix estimates

`estat vce` [, *estat_vce_options*]

estat_effects_options	description
`deff`	report DEFF design effects
`deft`	report DEFT design effects
`srssubpop`	report design effects, assuming SRS within subpopulation
`meff`	report MEFF design effects
`meft`	report MEFT design effects
display_options	control spacing and display of omitted variables and base and empty cells

estat_lceffects_options	description
deff	report DEFF design effects
deft	report DEFT design effects
srssubpop	report design effects, assuming SRS within subpopulation
meff	report MEFF design effects
meft	report MEFT design effects

estat_size_options	description
obs	report number of observations (within subpopulation)
size	report subpopulation sizes

estat_sd_options	description
variance	report subpopulation variances instead of standard deviations
srssubpop	report standard deviation, assuming SRS within subpopulation

estat_cv_options	description
nolegend	suppress the table legend

estat_gof_options	description
group(#)	compute test statistic using # quantiles
total	compute test statistic using the total estimator instead of the mean estimator
all	execute test for all observations in the data

estat_vce_options	description
covariance	display as covariance matrix; the default
correlation	display as correlation matrix
equation(*spec*)	display only specified equations
block	display submatrices by equation
diag	display submatrices by equation; diagonal blocks only
format(%*fmt*)	display format for covariances and correlations
nolines	suppress lines between equations
display_options	control display of omitted variables and base and empty cells

Menu

Statistics > Survey data analysis > DEFF, MEFF, and other statistics

Description

estat svyset reports the survey design characteristics associated with the current estimation results.

estat effects displays a table of design and misspecification effects for each estimated parameter.

estat lceffects displays a table of design and misspecification effects for a user-specified linear combination of the parameter estimates.

estat size displays a table of sample and subpopulation sizes for each estimated subpopulation mean, proportion, ratio, or total. This command is available only after svy: mean, svy: proportion, svy: ratio, and svy: total.

estat sd reports subpopulation standard deviations based on the estimation results from mean and svy: mean. estat sd is not appropriate with estimation results that used direct standardization or poststratification.

estat strata displays a table of the number of singleton and certainty strata within each sampling stage. The variance scaling factors are also displayed for estimation results where singleunit(scaled) was svyset.

estat cv reports the coefficient of variation (CV) for each coefficient in the current estimation results. The CV for coefficient b is

$$\text{CV}(b) = 100\frac{\text{SE}(b)}{b}$$

estat gof reports a goodness-of-fit test for binary response models using survey data. This command is available only after svy: logistic, svy: logit, and svy: probit.

estat vce displays the covariance or correlation matrix of the parameter estimates of the previous model. See [R] **estat** for examples.

Options for estat effects

deff and deft request that the design-effect measures DEFF and DEFT be displayed. This is the default, unless direct standardization or poststratification was used.

The deff and deft options are not allowed with estimation results that used direct standardization or poststratification. These methods obscure the measure of design effect because they adjust the frequency distribution of the target population.

srssubpop requests that DEFF and DEFT be computed using an estimate of simple random sampling (SRS) variance for sampling within a subpopulation. By default, DEFF and DEFT are computed using an estimate of the SRS variance for sampling from the entire population. Typically, srssubpop is used when computing subpopulation estimates by strata or by groups of strata.

meff and meft request that the misspecification-effect measures MEFF and MEFT be displayed.

display_options: noomitted, vsquish, noemptycells, baselevels, allbaselevels; see [R] **estimation options**.

Options for estat lceffects

deff and deft request that the design-effect measures DEFF and DEFT be displayed. This is the default, unless direct standardization or poststratification was used.

The deff and deft options are not allowed with estimation results that used direct standardization or poststratification. These methods obscure the measure of design effect because they adjust the frequency distribution of the target population.

srssubpop requests that DEFF and DEFT be computed using an estimate of simple random sampling (SRS) variance for sampling within a subpopulation. By default, DEFF and DEFT are computed using an estimate of the SRS variance for sampling from the entire population. Typically, srssubpop is used when computing subpopulation estimates by strata or by groups of strata.

meff and meft request that the misspecification-effect measures MEFF and MEFT be displayed.

Options for estat size

obs requests that the number of observations used to compute the estimate be displayed for each row of estimates.

size requests that the estimate of the subpopulation size be displayed for each row of estimates. The subpopulation size estimate equals the sum of the weights for those observations in the estimation sample that are also in the specified subpopulation. The estimated population size is reported when a subpopulation is not specified.

Options for estat sd

variance requests that the subpopulation variance be displayed instead of the standard deviation.

srssubpop requests that the standard deviation be computed using an estimate of SRS variance for sampling within a subpopulation. By default, the standard deviation is computed using an estimate of the SRS variance for sampling from the entire population. Typically, srssubpop is given when computing subpopulation estimates by strata or by groups of strata.

Option for estat cv

nolegend prevents the table legend identifying the subpopulations from being displayed.

Options for estat gof

group(*#*) specifies the number of quantiles to be used to group the data for the goodness-of-fit test. The minimum allowed value is group(2). The maximum allowed value is group(*df*), where *df* is the design degrees of freedom (e(df_r)). The default is group(10).

total requests that the goodness-of-fit test statistic be computed using the total estimator instead of the mean estimator.

all requests that the goodness-of-fit test statistic be computed for all observations in the data, ignoring any if or in restrictions specified with the model fit.

Options for estat vce

covariance displays the matrix as a variance–covariance matrix; this is the default.

correlation displays the matrix as a correlation matrix rather than a variance–covariance matrix. rho is a synonym for correlation.

`equation(`*spec*`)` selects the part of the VCE to be displayed. If *spec* = *eqlist*, the VCE for the listed equations is displayed. If *spec* = *eqlist1* \ *eqlist2*, the part of the VCE associated with the equations in *eqlist1* (rowwise) and *eqlist2* (columnwise) is displayed. * is shorthand for all equations. `equation()` implies `block` if `diag` is not specified.

`block` displays the submatrices pertaining to distinct equations separately.

`diag` displays the diagonal submatrices pertaining to distinct equations separately.

`format(%`*fmt*`)` specifies the display format for displaying the elements of the matrix. The default is `format(%10.0g)` for covariances and `format(%8.4f)` for correlations. See [U] **12.5 Formats: Controlling how data are displayed** for more information.

`nolines` suppresses lines between equations.

display_options: `noomitted`, `noemptycells`, `baselevels`, `allbaselevels`; see [R] **estimation options**.

Remarks

▷ Example 1

Using data from the Second National Health and Nutrition Examination Survey (NHANES II) (McDowell et al. 1981), let's estimate the population means for total serum cholesterol (`tcresult`) and for serum triglycerides (`tgresult`).

```
. use http://www.stata-press.com/data/r11/nhanes2
. svy: mean tcresult tgresult
(running mean on estimation sample)
Survey: Mean estimation
Number of strata =      31        Number of obs    =      5050
Number of PSUs   =      62        Population size  =  56820832
                                  Design df        =        31

--------------------------------------------------------------
             |             Linearized
             |       Mean   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
    tcresult |   211.3975   1.252274      208.8435    213.9515
    tgresult |    138.576   2.071934      134.3503    142.8018
--------------------------------------------------------------
```

We can use `estat svyset` to remind us of the survey design characteristics that were used to produce these results.

```
. estat svyset
      pweight: finalwgt
          VCE: linearized
  Single unit: missing
     Strata 1: strata
         SU 1: psu
        FPC 1: <zero>
```

(*Continued on next page*)

`estat effects` reports a table of design and misspecification effects for each mean we estimated.

```
. estat effects, deff deft meff meft
```

	Mean	Linearized Std. Err.	DEFF	DEFT	MEFF	MEFT
tcresult	211.3975	1.252274	3.57141	1.88982	3.46105	1.86039
tgresult	138.576	2.071934	2.35697	1.53524	2.32821	1.52585

`estat size` reports a table that contains sample and population sizes.

```
. estat size
```

	Mean	Linearized Std. Err.	Obs	Size
tcresult	211.3975	1.252274	5050	56820832
tgresult	138.576	2.071934	5050	56820832

`estat size` can also report a table of subpopulation sizes.

```
. svy: mean tcresult, over(sex)
```
(*output omitted*)

```
. estat size

        Male: sex = Male
      Female: sex = Female
```

Over	Mean	Linearized Std. Err.	Obs	Size
tcresult				
Male	210.7937	1.312967	4915	56159480
Female	215.2188	1.193853	5436	60998033

`estat sd` reports a table of subpopulation standard deviations.

```
. estat sd

        Male: sex = Male
      Female: sex = Female
```

Over	Mean	Std. Dev.
tcresult		
Male	210.7937	45.79065
Female	215.2188	50.72563

(*Continued on next page*)

`estat cv` reports a table of coefficients of variations for each estimate.

```
. estat cv

        Male: sex = Male
      Female: sex = Female
```

Over	Mean	Linearized Std. Err.	CV
tcresult			
Male	210.7937	1.312967	.622868
Female	215.2188	1.193853	.554716

◁

▷ Example 2: Design effects with subpopulations

When there are subpopulations, `estat effects` can compute design effects with respect to one of two different hypothetical SRS designs. The default design is one in which SRS is conducted across the full population. The alternate design is one in which SRS is conducted entirely within the subpopulation of interest. This alternate design is used when the `srssubpop` option is specified.

Deciding which design is preferable depends on the nature of the subpopulations. If we can imagine identifying members of the subpopulations before sampling them, the alternate design is preferable. This case arises primarily when the subpopulations are strata or groups of strata. Otherwise, we may prefer to use the default.

Here is an example using the default with the NHANES II data.

```
. use http://www.stata-press.com/data/r11/nhanes2b
. svy: mean iron, over(sex)
  (output omitted)
. estat effects

        Male: sex = Male
      Female: sex = Female
```

Over	Mean	Linearized Std. Err.	DEFF	DEFT
iron				
Male	104.7969	.557267	1.36097	1.16661
Female	97.16247	.6743344	2.01403	1.41916

Thus the design-based variance estimate is about 36% larger than the estimate from the hypothetical SRS design including the full population. We can get DEFF and DEFT for the alternate SRS design by using the `srssubpop` option.

```
. estat effects, srssubpop

        Male: sex = Male
      Female: sex = Female
```

Over	Mean	Linearized Std. Err.	DEFF	DEFT
iron				
Male	104.7969	.557267	1.348	1.16104
Female	97.16247	.6743344	2.03132	1.42524

Because the NHANES II did not stratify on sex, we think it problematic to consider design effects with respect to SRS of the female (or male) subpopulation. Consequently, we would prefer to use the default here, although the values of DEFF differ little between the two in this case.

For other examples (generally involving heavy oversampling or undersampling of specified subpopulations), the differences in DEFF for the two schemes can be much more dramatic.

Consider the NMIHS data (Gonzalez Jr., Krauss, and Scott 1992), and compute the mean of `birthwgt` over `race`:

```
. use http://www.stata-press.com/data/r11/nmihs
. svy: mean birthwgt, over(race)
  (output omitted)
. estat effects

    nonblack: race = nonblack
       black: race = black
```

Over	Mean	Linearized Std. Err.	DEFF	DEFT
birthwgt				
nonblack	3402.32	7.609532	1.44376	1.20157
black	3127.834	6.529814	.172041	.414778

```
. estat effects, srssubpop

    nonblack: race = nonblack
       black: race = black
```

Over	Mean	Linearized Std. Err.	DEFF	DEFT
birthwgt				
nonblack	3402.32	7.609532	.826842	.909308
black	3127.834	6.529814	.528963	.727298

Because the NMIHS survey was stratified on race, marital status, age, and birthweight, we believe it reasonable to consider design effects computed with respect to SRS within an individual race group. Consequently, we would recommend here the alternative hypothetical design for computing design effects; that is, we would use the `srssubpop` option.

◁

▷ Example 3: Misspecification effects

Misspecification effects assess biases in variance estimators that are computed under the wrong assumptions. The survey literature (for example, Scott and Holt 1982, 850; Skinner 1989) defines misspecification effects with respect to a general set of "wrong" variance estimators. `estat effects` considers only one specific form: variance estimators computed under the incorrect assumption that our *observed* sample was selected through SRS.

The resulting "misspecification effect" measure is informative primarily when an unweighted point estimator is approximately unbiased for the parameter of interest. See Eltinge and Sribney (1996a) for a detailed discussion of extensions of misspecification effects that are appropriate for *biased* point estimators.

Note the difference between a misspecification effect and a design effect. For a design effect, we compare our complex-design–based variance estimate with an estimate of the true variance that we would have obtained under a hypothetical true simple random sample. For a misspecification effect, we compare our complex-design–based variance estimate with an estimate of the variance from fitting the same model without weighting, clustering, or stratification.

`estat effects` defines MEFF and MEFT as

$$\text{MEFF} = \widehat{V}/\widehat{V}_{\text{msp}}$$

$$\text{MEFT} = \sqrt{\text{MEFF}}$$

where $\widehat{V}$ is the appropriate design-based estimate of variance and $\widehat{V}_{\text{msp}}$ is the variance estimate computed with a misspecified design—ignoring the sampling weights, stratification, and clustering.

Here we request that the misspecification effects be displayed for the estimation of mean zinc levels from our NHANES II data.

```
. use http://www.stata-press.com/data/r11/nhanes2b
. svy: mean zinc, over(sex)
  (output omitted)
. estat effects, meff meft
            Male: sex = Male
          Female: sex = Female
```

Over	Mean	Linearized Std. Err.	MEFF	MEFT
zinc				
Male	90.74543	.5850741	6.28254	2.5065
Female	83.8635	.4689532	6.32648	2.51525

If we run `ci` without weights, we get the standard errors that are $(\widehat{V}_{\text{msp}})^{1/2}$.

```
. sort sex
. ci zinc if sex == "Male":sex
```

Variable	Obs	Mean	Std. Err.	[95% Conf.	Interval]
zinc	4375	89.53143	.2334228	89.0738	89.98906

```
. display [zinc]_se[Male]/r(se)
2.5064994
. display ([zinc]_se[Male]/r(se))^2
6.2825393
. ci zinc if sex == "Female":sex
```

Variable	Obs	Mean	Std. Err.	[95% Conf.	Interval]
zinc	4827	83.76652	.186444	83.40101	84.13204

```
. display [zinc]_se[Female]/r(se)
2.515249
. display ([zinc]_se[Female]/r(se))^2
6.3264774
```

◁

▷ Example 4: Design and misspecification effects for linear combinations

Let's compare the mean of total serum cholesterol (`tcresult`) between men and women in the NHANES II dataset.

```
. use http://www.stata-press.com/data/r11/nhanes2
. svy: mean tcresult, over(sex)
(running mean on estimation sample)
Survey: Mean estimation

Number of strata =      31       Number of obs     =      10351
Number of PSUs   =      62       Population size   =  117157513
                                 Design df         =         31

         Male: sex = Male
       Female: sex = Female
```

Over	Mean	Linearized Std. Err.	[95% Conf.	Interval]
tcresult				
Male	210.7937	1.312967	208.1159	213.4715
Female	215.2188	1.193853	212.784	217.6537

We can use `estat lceffects` to report the standard error, design effects, and misspecification effects of the difference between the above means.

```
. estat lceffects [tcresult]Male - [tcresult]Female, deff deft meff meft
 ( 1)  [tcresult]Male - [tcresult]Female = 0
```

Mean	Coef.	Std. Err.	DEFF	DEFT	MEFF	MEFT
(1)	-4.425109	1.086786	1.31241	1.1456	1.27473	1.12904

◁

▷ Example 5: Using survey data to determine Neyman allocation

Suppose that we have partitioned our population into L strata and stratum h contains N_h individuals. Also let σ_h represent the standard deviation of a quantity we wish to sample from the population. According to Cochran (1977, sec. 5.5), we can minimize the variance of the stratified mean estimator, for a fixed sample size n, if we choose the stratum sample sizes according to Neyman allocation:

$$n_h = n \frac{N_h \sigma_h}{\sum_{i=1}^{L} N_i \sigma_i} \tag{1}$$

We can use `estat sd` with our current survey data to produce a table of subpopulation standard-deviation estimates. Then we could plug these estimates into (1) to improve our survey design for the next time we sample from our population.

Here is an example using birthweight from the NMIHS data. First, we need estimation results from `svy: mean` over the strata.

```
. use http://www.stata-press.com/data/r11/nmihs
. svyset [pw=finwgt], strata(stratan)
      pweight: finwgt
          VCE: linearized
  Single unit: missing
     Strata 1: stratan
         SU 1: <observations>
        FPC 1: <zero>
. svy: mean birthwgt, over(stratan)
  (output omitted )
```

Next we will use `estat size` to report the table of stratum sizes. We will also generate matrix `p_obs` to contain the observed percent allocations for each stratum. In the matrix expression, `r(_N)` is a row vector of stratum sample sizes and `e(N)` contains the total sample size. `r(_N_subp)` is a row vector of the estimated population stratum sizes.

```
. estat size
            1: stratan = 1
            2: stratan = 2
            3: stratan = 3
            4: stratan = 4
            5: stratan = 5
            6: stratan = 6
```

Over	Mean	Linearized Std. Err.	Obs	Size
birthwgt				
1	1049.434	19.00149	841	18402.98161
2	2189.561	9.162736	803	67650.95932
3	3303.492	7.38429	3578	579104.6188
4	1036.626	12.32294	710	29814.93215
5	2211.217	9.864682	714	153379.07445
6	3485.42	8.057648	3300	3047209.10519

```
. matrix p_obs = 100 * r(_N)/e(N)
. matrix nsubp = r(_N_subp)
```

Now we call `estat sd` to report the stratum standard-deviation estimates and generate matrix `p_neyman` to contain the percent allocations according to (1). In the matrix expression, `r(sd)` is a vector of the stratum standard deviations.

(*Continued on next page*)

```
. estat sd
                1: stratan = 1
                2: stratan = 2
                3: stratan = 3
                4: stratan = 4
                5: stratan = 5
                6: stratan = 6

            Over         Mean    Std. Dev.

birthwgt
               1     1049.434    2305.931
               2     2189.561    555.7971
               3     3303.492    687.3575
               4     1036.626    999.0867
               5     2211.217    349.8068
               6      3485.42    300.6945

. matrix p_neyman = 100 * hadamard(nsubp,r(sd))/el(nsubp*r(sd)',1,1)
. matrix list p_obs, format(%4.1f)
p_obs[1,6]
     birthwgt:  birthwgt:  birthwgt:  birthwgt:  birthwgt:  birthwgt:
             1          2          3          4          5          6
r1         8.5        8.1       36.0        7.1        7.2       33.2
. matrix list p_neyman, format(%4.1f)
p_neyman[1,6]
     birthwgt:  birthwgt:  birthwgt:  birthwgt:  birthwgt:  birthwgt:
             1          2          3          4          5          6
r1         2.9        2.5       26.9        2.0        3.6       62.0
```

We can see that strata 3 and 6 each contain about one-third of the observed data, with the rest of the observations spread out roughly equally to the remaining strata. However, plugging our sample estimates into (1) indicates that stratum 6 should get 62% of the sampling units, stratum 3 should get about 27%, and the remaining strata should get a roughly equal distribution of sampling units.

◁

▷ Example 6: Summarizing singleton and certainty strata

Use `estat strata` with `svy` estimation results to produce a table that reports the number of singleton and certainty strata in each sampling stage. Here is an example using (fictional) data from a complex survey with five sampling stages (the dataset is already `svyset`). If singleton strata are present, `estat strata` will report their effect on the standard errors.

```
. use http://www.stata-press.com/data/r11/strata5
. svy: total y
  (output omitted)
```

(Continued on next page)

```
. estat strata
```

Stage	Singleton strata	Certainty strata	Total strata
1	0	1	4
2	1	0	10
3	0	3	29
4	2	0	110
5	204	311	865

```
Note: missing standard error because of
      stratum with single sampling unit.
```

`estat strata` also reports the scale factor used when the `singleunit(scaled)` option is `svyset`. Of the 865 strata in the last stage, 204 are singleton strata and 311 are certainty strata. Thus the scaling factor for the last stage is

$$\frac{865 - 311}{865 - 311 - 204} \approx 1.58$$

```
. svyset, singleunit(scaled) noclear
  (output omitted )
. svy: total y
  (output omitted )
. estat strata
```

Stage	Singleton strata	Certainty strata	Total strata	Scale factor
1	0	1	4	1
2	1	0	10	1.11
3	0	3	29	1
4	2	0	110	1.02
5	204	311	865	1.58

```
Note: variances scaled within each stage to handle
      strata with a single sampling unit.
```

The `singleunit(scaled)` option of `svyset` is one of three methods in which Stata's `svy` commands can automatically handle singleton strata when performing variance estimation; see [SVY] **variance estimation** for a brief discussion of these methods.

◁

▷ Example 7: Goodness-of-fit test for svy: logistic

From example 2 in [SVY] **svy estimation**, we modeled the incidence of high blood pressure as a function of height, weight, age, and sex (using the `female` indicator variable).

```
. use http://www.stata-press.com/data/r11/nhanes2d
. svyset
      pweight: finalwgt
          VCE: linearized
  Single unit: missing
     Strata 1: strata
         SU 1: psu
        FPC 1: <zero>
. svy: logistic highbp height weight age female
(running logistic on estimation sample)
Survey: Logistic regression
Number of strata   =        31              Number of obs      =      10351
Number of PSUs     =        62              Population size    =  117157513
                                            Design df          =         31
                                            F(   4,     28)    =     178.69
                                            Prob > F           =     0.0000

------------------------------------------------------------------------------
             |             Linearized
      highbp | Odds Ratio   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      height |   .9688567   .0056822    -5.39   0.000     .9573369    .9805151
      weight |   1.052489   .0032829    16.40   0.000     1.045814    1.059205
         age |   1.050473   .0024816    20.84   0.000     1.045424    1.055547
      female |   .7250087   .0641188    -3.64   0.001      .605353    .8683158
------------------------------------------------------------------------------
```

We can use `estat gof` to perform a goodness-of-fit test for this model.

```
. estat gof
Logistic model for highbp, goodness-of-fit test
                        F(9,23) =          3.49
                       Prob > F =        0.0074
```

The F statistic is significant at the 5% level, indicating that the model is not a good fit for these data.
◁

(*Continued on next page*)

Saved results

estat svyset saves the following in r():

Scalars

r(stages)	number of sampling stages

Macros

r(wtype)	weight type
r(wexp)	weight expression
r(wvar)	weight variable name
r(su#)	variable identifying sampling units for stage #
r(strata#)	variable identifying strata for stage #
r(fpc#)	FPC for stage #
r(brrweight)	brrweight() variable list
r(fay)	Fay's adjustment
r(bsrweight)	bsrweight() variable list
r(bsn)	bootstrap mean-weight adjustment
r(jkrweight)	jkrweight() variable list
r(sdrweight)	sdrweight() variable list
r(sdrfpc)	fpc() value from within sdrweight()
r(vce)	*vcetype* specified in vce()
r(dof)	dof() value
r(mse)	mse, if specified
r(poststrata)	poststrata() variable
r(postweight)	postweight() variable
r(settings)	svyset arguments to reproduce the current settings
r(singleunit)	singleunit() setting

estat strata saves the following in r():

Matrices

r(_N_strata_single)	number of strata with one sampling unit
r(_N_strata_certain)	number of certainty strata
r(_N_strata)	number of strata
r(scale)	variance scale factors used when singleunit(scaled) is svyset

estat effects saves the following in r():

Matrices

r(deff)	vector of DEFF estimates
r(deft)	vector of DEFT estimates
r(deffsub)	vector of DEFF estimates for srssubpop
r(deftsub)	vector of DEFT estimates for srssubpop
r(meff)	vector of MEFF estimates
r(meft)	vector of MEFT estimates

estat lceffects saves the following in r():

Scalars

r(estimate)	point estimate
r(se)	estimate of standard error
r(df)	degrees of freedom
r(deff)	DEFF estimate
r(deft)	DEFT estimate
r(deffsub)	DEFF estimate for srssubpop
r(deftsub)	DEFT estimate for srssubpop
r(meff)	MEFF estimate
r(meft)	MEFT estimate

`estat size` saves the following in `r()`:

Matrices	
`r(_N)`	vector of numbers of nonmissing observations
`r(_N_subp)`	vector of subpopulation size estimates

`estat sd` saves the following in `r()`:

Macros	
`r(srssubpop)`	`srssubpop`, if specified
Matrices	
`r(mean)`	vector of subpopulation mean estimates
`r(sd)`	vector of subpopulation standard-deviation estimates
`r(variance)`	vector of subpopulation variance estimates

`estat cv` saves the following in `r()`:

Matrices	
`r(b)`	estimates
`r(se)`	standard errors of the estimates
`r(cv)`	coefficients of variation of the estimates

`estat gof` saves the following in `r()`:

Scalars	
`r(p)`	p-value associated with the test statistic
`r(F)`	F statistic, if `e(df_r)` was saved by estimation command
`r(df1)`	numerator degrees of freedom for F statistic
`r(df2)`	denominator degrees of freedom for F statistic
`r(chi2)`	χ^2 statistic, if `e(df_r)` was not saved by estimation command
`r(df)`	degrees of freedom for χ^2 statistic

`estat vce` saves the following in `r()`:

Matrices	
`r(V)`	VCE or correlation matrix

Methods and formulas

`estat` is implemented as an ado-file.

Methods and formulas are presented under the following headings:

Design effects
Linear combinations
Misspecification effects
Population and subpopulation standard deviations
Coefficients of variation
Goodness of fit for binary response models

Design effects

`estat effects` produces two estimators of design effect, DEFF and DEFT.

DEFF is estimated as described in Kish (1965) as

$$\text{DEFF} = \frac{\widehat{V}(\widehat{\theta})}{\widehat{V}_{\text{srswor}}(\widetilde{\theta}_{\text{srs}})}$$

where $\widehat{V}(\widehat{\theta})$ is the design-based estimate of variance for a parameter, θ, and $\widehat{V}_{\text{srswor}}(\widetilde{\theta}_{\text{srs}})$ is an estimate of the variance for an estimator, $\widetilde{\theta}_{\text{srs}}$, that would be obtained from a similar hypothetical survey conducted using SRS without replacement (wor) and with the same number of sample elements, m, as in the actual survey. For example, if θ is a total Y, then

$$\widehat{V}_{\text{srswor}}(\widetilde{\theta}_{\text{srs}}) = (1-f)\frac{\widehat{M}}{m-1}\sum_{j=1}^{m} w_j\left(y_j - \widehat{\overline{Y}}\right)^2 \tag{1}$$

where $\widehat{\overline{Y}} = \widehat{Y}/\widehat{M}$. The factor $(1-f)$ is a finite population correction. If the user sets an FPC for the first stage, $f = m/\widehat{M}$ is used; otherwise, $f = 0$.

DEFT is estimated as described in Kish (1987, 41) as

$$\text{DEFT} = \sqrt{\frac{\widehat{V}(\widehat{\theta})}{\widehat{V}_{\text{srswr}}(\widetilde{\theta}_{\text{srs}})}}$$

where $\widehat{V}_{\text{srswr}}(\widetilde{\theta}_{\text{srs}})$ is an estimate of the variance for an estimator, $\widetilde{\theta}_{\text{srs}}$, obtained from a similar survey conducted using SRS with replacement (wr). $\widehat{V}_{\text{srswr}}(\widetilde{\theta}_{\text{srs}})$ is computed using (1) with $f = 0$.

When computing estimates for a subpopulation, $\mathcal{S}$, and the `srssubpop` option is *not* specified (that is, the default), (1) is used with $w_{\mathcal{S}j} = I_{\mathcal{S}}(j)\, w_j$ in place of w_j, where

$$I_{\mathcal{S}}(j) = \begin{cases} 1, \text{ if } j \in \mathcal{S} \\ 0, \text{ otherwise} \end{cases}$$

The sums in (1) are still calculated over all elements in the sample, regardless of whether they belong to the subpopulation: by default, the SRS is assumed to be done across the full population.

When the `srssubpop` option is specified, the SRS is carried out within subpopulation $\mathcal{S}$. Here (1) is used with the sums restricted to those elements belonging to the subpopulation; m is replaced with $m_{\mathcal{S}}$, the number of sample elements from the subpopulation; $\widehat{M}$ is replaced with $\widehat{M}_{\mathcal{S}}$, the sum of the weights from the subpopulation; and $\widehat{\overline{Y}}$ is replaced with $\widehat{\overline{Y}}_{\mathcal{S}} = \widehat{Y}_{\mathcal{S}}/\widehat{M}_{\mathcal{S}}$, the weighted mean across the subpopulation.

Linear combinations

`estat lceffects` estimates $\eta = C\theta$, where θ is a $q \times 1$ vector of parameters (for example, population means or population regression coefficients) and C is any $1 \times q$ vector of constants. The estimate of η is $\widehat{\eta} = C\widehat{\theta}$, and its variance estimate is

$$\widehat{V}(\widehat{\eta}) = C\widehat{V}(\widehat{\theta})C'$$

Similarly, the SRS without replacement (srswor) variance estimator used in the computation of DEFF is

$$\widehat{V}_{\rm srswor}(\widetilde{\eta}_{\rm srs}) = C\widehat{V}_{\rm srswor}(\widehat{\theta}_{\rm srs})C'$$

and the SRS with replacement (srswr) variance estimator used in the computation of DEFT is

$$\widehat{V}_{\rm srswr}(\widetilde{\eta}_{\rm srs}) = C\widehat{V}_{\rm srswr}(\widehat{\theta}_{\rm srs})C'$$

The variance estimator used in computing MEFF and MEFT is

$$\widehat{V}_{\rm msp}(\widetilde{\eta}_{\rm msp}) = C\widehat{V}_{\rm msp}(\widehat{\theta}_{\rm msp})C'$$

`estat lceffects` was originally developed under a different command name; see Eltinge and Sribney (1996b).

Misspecification effects

`estat effects` produces two estimators of misspecification effect, MEFF and MEFT.

$$\text{MEFF} = \frac{\widehat{V}(\widehat{\theta})}{\widehat{V}_{\rm msp}(\widehat{\theta}_{\rm msp})}$$

$$\text{MEFT} = \sqrt{\text{MEFF}}$$

where $\widehat{V}(\widehat{\theta})$ is the design-based estimate of variance for a parameter, θ, and $\widehat{V}_{\rm msp}(\widehat{\theta}_{\rm msp})$ is the variance estimate for $\widehat{\theta}_{\rm msp}$. These estimators, $\widehat{\theta}_{\rm msp}$ and $\widehat{V}_{\rm msp}(\widehat{\theta}_{\rm msp})$, are based on the incorrect assumption that the observations were obtained through SRS with replacement: they are the estimators obtained by simply ignoring weights, stratification, and clustering. When θ is a total Y, the estimator and its variance estimate are computed using the standard formulas for an unweighted total:

$$\widehat{Y}_{\rm msp} = \widehat{M}\,\overline{y} = \frac{\widehat{M}}{m}\sum_{j=1}^{m} y_j$$

$$\widehat{V}_{\rm msp}(\widehat{Y}_{\rm msp}) = \frac{\widehat{M}^2}{m(m-1)}\sum_{j=1}^{m}\left(y_j - \overline{y}\right)^2$$

When computing MEFF and MEFT for a subpopulation, sums are restricted to those elements belonging to the subpopulation, and $m_{\mathcal{S}}$ and $\widehat{M}_{\mathcal{S}}$ are used in place of m and $\widehat{M}$.

Population and subpopulation standard deviations

For srswr designs, the variance of the mean estimator is

$$V_{\rm srswr}(\overline{y}) = \sigma^2/n$$

where n is the sample size and σ is the population standard deviation. `estat sd` uses this formula and the results from `mean` and `svy: mean` to estimate the population standard deviation via

$$\widehat{\sigma} = \sqrt{n\ \widehat{V}_{\rm srswr}(\overline{y})}$$

Subpopulation standard deviations are computed similarly, using the corresponding variance estimate and sample size.

Coefficients of variation

The coefficient of variation (CV) for estimate $\widehat{\theta}$ is

$${\rm CV}(\widehat{\theta}) = 100 \frac{\sqrt{\widehat{V}(\widehat{\theta})}}{|\widehat{\theta}|}$$

A missing value is reported when $\widehat{\theta}$ is zero.

Goodness of fit for binary response models

Let y_j be the jth observed value of the dependent variable, $\widehat{p}_j$ be the predicted probability of a positive outcome, and $\widehat{r}_j = y_j - \widehat{p}_j$. Let g be the requested number of groups from the `group()` option; then the $\widehat{r}_j$ are placed in g quantile groups as described in *Methods and formulas* for the `xtile` command in [D] **pctile**. Let $\overline{\mathbf{r}} = (\overline{r}_1, \dots, \overline{r}_g)$, where $\overline{r}_i$ is the subpopulation mean of the $\widehat{r}_j$ for the ith quantile group. The standard Wald statistic for testing $\overline{\mathbf{r}} = \mathbf{0}$ is

$$\widehat{X}^2 = \overline{\mathbf{r}}\{\widehat{V}(\overline{\mathbf{r}})\}^{-1}\overline{\mathbf{r}}'$$

where $\widehat{V}(\overline{\mathbf{r}})$ is the design-based variance estimate for $\overline{\mathbf{r}}$. Here $\widehat{X}^2$ is approximately distributed as a χ^2 with $g-1$ degrees of freedom. This Wald statistic is one of the three goodness-of-fit statistics discussed in Graubard, Korn, and Midthune (1997). `estat gof` reports this statistic when the design degrees of freedom is missing, such as with `svy bootstrap` results.

According to Archer and Lemeshow (2006), the F-adjusted mean residual test is given by

$$\widehat{F} = \widehat{X}^2(d-g+2)/(dg)$$

where d is the design degrees of freedom. Here $\widehat{F}$ is approximately distributed as an F with $g-1$ numerator and $d-g+2$ denominator degrees of freedom.

With the `total` option, `estat gof` uses the subpopulation total estimator instead of the subpopulation mean estimator.

References

Archer, K. J., and S. Lemeshow. 2006. Goodness-of-fit test for a logistic regression model fitted using survey sample data. *Stata Journal* 6: 97–105.

Cochran, W. G. 1977. *Sampling Techniques*. 3rd ed. New York: Wiley.

Eltinge, J. L., and W. M. Sribney. 1996a. Accounting for point-estimation bias in assessment of misspecification effects, confidence-set coverage rates and test sizes. Unpublished manuscript, Department of Statistics, Texas A&M University.

——. 1996b. svy5: Estimates of linear combinations and hypothesis tests for survey data. *Stata Technical Bulletin* 31: 31–42. Reprinted in *Stata Technical Bulletin Reprints*, vol. 6, pp. 246–259. College Station, TX: Stata Press.

Gonzalez Jr., J. F., N. Krauss, and C. Scott. 1992. Estimation in the 1988 National Maternal and Infant Health Survey. *Proceedings of the Section on Statistics Education, American Statistical Association* 343–348.

Graubard, B. I., E. L. Korn, and D. Midthune. 1997. Testing goodness-of-fit for logistic regression with survey data. In *Proceedings of the Section on Survey Research Methods, Joint Statistical Meetings*, 170–174. Alexandria, VA: American Statistical Association.

Kish, L. 1965. *Survey Sampling*. New York: Wiley.

——. 1987. *Statistical Design for Research*. New York: Wiley.

McDowell, A., A. Engel, J. T. Massey, and K. Maurer. 1981. Plan and operation of the Second National Health and Nutrition Examination Survey, 1976–1980. *Vital and Health Statistics* 1(15): 1–144.

Scott, A. J., and D. Holt. 1982. The effect of two-stage sampling on ordinary least squares methods. *Journal of the American Statistical Association* 77: 848–854.

Skinner, C. J. 1989. Introduction to part A. In *Analysis of Complex Surveys*, ed. C. J. Skinner, D. Holt, and T. M. F. Smith, 23–58. New York: Wiley.

Also see

[SVY] **svy postestimation** — Postestimation tools for svy

[SVY] **svy estimation** — Estimation commands for survey data

[SVY] **subpopulation estimation** — Subpopulation estimation for survey data

[SVY] **variance estimation** — Variance estimation for survey data