

jackknife — Jackknife estimation

Description	Quick start	Menu	Syntax
Options	Remarks and examples	Stored results	Methods and formulas
References	Also see		

Description

`jackknife` performs jackknife estimation of the specified statistics (or expressions) for a Stata command or a user-written program. Statistics are jackknifed by estimating the command once for each observation or cluster in the dataset, leaving the associated observation or cluster out of the calculations. `jackknife` is designed for use with nonestimation commands, functions of coefficients, or user-written programs. To jackknife coefficients, we recommend using the `vce(jackknife)` option when allowed by the estimation command.

`jknife` is a synonym for `jackknife`.

Quick start

Jackknife estimate of the standard deviation of `v1` returned by `summarize` in `r(sd)`

```
jackknife sd=r(sd), rclass: summarize v1
```

Jackknife estimate of the statistic `r(mystat)` returned by `rclass` program `myprog1` that returns the sample size in `r(N)`

```
jackknife stat=r(mystat), rclass: myprog1 v1
```

As above, and save the results from each replication in `mydata.dta`

```
jackknife stat=r(mystat), rclass saving(mydata): myprog1 v1
```

Jackknife estimate of a difference in coefficients estimated by `regress`

```
jackknife diff=(_b[x2]-_b[x1]): regress y x1 x2 x3
```

Jackknife estimate of the statistic `e(mystat)` returned by `eclass` program `myprog2` that returns the sample size in `e(N)`

```
jackknife stat=e(mystat), eclass: myprog2 y x1 x2 x3
```

Jackknife estimates of coefficients stored in `e(b)` by `myprog2`

```
jackknife _b, eclass: myprog2 y x1 x2 x3
```

Add variables containing the pseudovalues of the coefficients to the dataset

```
jackknife _b, eclass keep: myprog2 y x1 x2 x3
```

Menu

Statistics > Resampling > Jackknife estimation

Syntax

```
jackknife exp_list [ , options eform_option ] : command
```

<i>options</i>	Description
Main	
<code>eclass</code>	number of observations used is stored in <code>e(N)</code>
<code>rclass</code>	number of observations used is stored in <code>r(N)</code>
<code>n(<i>exp</i>)</code>	specify <i>exp</i> that evaluates to the number of observations used
Options	
<code>cluster(<i>varlist</i>)</code>	variables identifying sample clusters
<code>idcluster(<i>newvar</i>)</code>	create new cluster ID variable
<code>saving(<i>filename</i>, ...)</code>	save results to <i>filename</i> ; save statistics in double precision; save results to <i>filename</i> every # replications
<code>keep</code>	keep pseudovalues
<code>mse</code>	use MSE formula for variance estimation
Reporting	
<code>level(#)</code>	set confidence level; default is <code>level(95)</code>
<code>notable</code>	suppress table of results
<code>noheader</code>	suppress table header
<code>nolegend</code>	suppress table legend
<code>verbose</code>	display the full table legend
<code>nodots</code>	suppress replication dots
<code>noisily</code>	display any output from <i>command</i>
<code>trace</code>	trace <i>command</i>
<code>title(<i>text</i>)</code>	use <i>text</i> as title for jackknife results
<code>display_options</code>	control columns and column formats, row spacing, line width, display of omitted variables and base and empty cells, and factor-variable labeling
<code>eform_option</code>	display coefficient table in exponentiated form
Advanced	
<code>nodrop</code>	do not drop observations
<code>reject(<i>exp</i>)</code>	identify invalid results
<code>coeflegend</code>	display legend instead of statistics

`svy` is allowed; see [SVY] [svy jackknife](#).

command is any command that follows standard Stata syntax. All weight types supported by *command* are allowed except `aweight`s; see [U] [11.1.6 weight](#).

`coeflegend` does not appear in the dialog box.

See [U] [20 Estimation and postestimation commands](#) for more capabilities of estimation commands.

```

exp_list contains    (name: elist)
                    elist
                    eexp
elist contains      newvar = (exp)
                    (exp)
eexp is             specname
                    [eqno]specname
specname is        _b
                    _b[]
                    _se
                    _se[]
eqno is            ##
                    name

```

exp is a standard Stata expression; see [U] 13 Functions and expressions.

Distinguish between [], which are to be typed, and [], which indicate optional arguments.

Options

Main

`eclass`, `rclass`, and `n(exp)` specify where *command* stores the number of observations on which it based the calculated results. We strongly advise you to specify one of these options.

`eclass` specifies that *command* store the number of observations in $e(N)$.

`rclass` specifies that *command* store the number of observations in $r(N)$.

`n(exp)` specifies an expression that evaluates to the number of observations used. Specifying `n(r(N))` is equivalent to specifying the `rclass` option. Specifying `n(e(N))` is equivalent to specifying the `eclass` option. If *command* stores the number of observations in $r(N1)$, specify `n(r(N1))`.

If you specify no options, `jackknife` will assume `eclass` or `rclass`, depending on which of $e(N)$ and $r(N)$ is not missing (in that order). If both $e(N)$ and $r(N)$ are missing, `jackknife` assumes that all observations in the dataset contribute to the calculated result. If that assumption is incorrect, the reported standard errors will be incorrect. For instance, say that you specify

```
. jackknife coef=_b[x2]: myreg y x1 x2 x3
```

where `myreg` uses $e(n)$ instead of $e(N)$ to identify the number of observations used in calculations. Further assume that observation 42 in the dataset has `x3` equal to missing. The 42nd observation plays no role in obtaining the estimates, but `jackknife` has no way of knowing that and will use the wrong N . If, on the other hand, you specify

```
. jackknife coef=_b[x2], n(e(n)): myreg y x1 x2 x3
```

`jackknife` will notice that observation 42 plays no role. The `n(e(n))` option is specified because `myreg` is an estimation command but it stores the number of observations used in $e(n)$ (instead of the standard $e(N)$). When `jackknife` runs the regression omitting the 42nd observation, `jackknife` will observe that $e(n)$ has the same value as when `jackknife` previously ran the regression using all the observations. Thus `jackknife` will know that `myreg` did not use the observation.

Options

`cluster(varlist)` specifies the variables identifying sample clusters. If `cluster()` is specified, one cluster is left out of each call to `command`, instead of 1 observation.

`idcluster(newvar)` creates a new variable containing a unique integer identifier for each resampled cluster, starting at 1 and leading up to the number of clusters. This option may be specified only when the `cluster()` option is specified. `idcluster()` helps identify the cluster to which a pseudo-value belongs.

`saving(filename [, suboptions])` creates a Stata data file (.`dta` file) consisting of (for each statistic in `exp_list`) a variable containing the replicates.

`double` specifies that the results for each replication be saved as `doubles`, meaning 8-byte reals. By default, they are saved as `floats`, meaning 4-byte reals. This option may be used without the `saving()` option to compute the variance estimates by using double precision.

`every(#)` specifies that results be written to disk every `#`th replication. `every()` should be specified only in conjunction with `saving()` when `command` takes a long time for each replication. This option will allow recovery of partial results should some other software crash your computer. See [P] [postfile](#).

`replace` specifies that `filename` be overwritten if it exists. This option does not appear in the dialog box.

`keep` specifies that new variables be added to the dataset containing the pseudo-values of the requested statistics. For instance, if you typed

```
. jackknife coef=_b[x2], eclass keep: regress y x1 x2 x3
```

new variable `coef` would be added to the dataset containing the pseudo-values for `_b[x2]`. Let b be the value of `_b[x2]` when all observations are used to fit the model, and let $b(j)$ be the value when the j th observation is omitted. The pseudo-values are defined as

$$\text{pseudo-value}_j = N \{b - b(j)\} + b(j)$$

where N is the number of observations used to produce b .

When the `cluster()` option is specified, each cluster is given at most one nonmissing pseudo-value. The `keep` option implies the `nodrop` option.

`mse` specifies that `jackknife` compute the variance by using deviations of the replicates from the observed value of the statistics based on the entire dataset. By default, `jackknife` computes the variance by using deviations of the pseudo-values from their mean.

Reporting

`level(#)`; see [R] [estimation options](#).

`notable` suppresses the display of the table of results.

`noheader` suppresses the display of the table header. This option implies `nolegend`.

`nolegend` suppresses the display of the table legend. The table legend identifies the rows of the table with the expressions they represent.

`verbose` specifies that the full table legend be displayed. By default, coefficients and standard errors are not displayed.

`nodots` suppresses display of the replication dots. By default, one dot character is displayed for each successful replication. A red 'x' is displayed if `command` returns an error or if one of the values in `exp_list` is missing.

`noisily` specifies that any output from *command* be displayed. This option implies the `nodots` option.

`trace` causes a trace of the execution of *command* to be displayed. This option implies the `noisily` option.

`title(text)` specifies a title to be displayed above the table of jackknife results; the default title is `Jackknife results` or what is produced in `e(title)` by an estimation command.

display_options: `nocl`, `nopvalues`, `noomitted`, `vsquish`, `noemptycells`, `baselevels`, `allbaselevels`, `nofvlabel`, `fvwrap(#)`, `fvwrapon(style)`, `cformat(%fmt)`, `pformat(%fmt)`, `sformat(%fmt)`, and `nolstretch`; see [R] [estimation options](#).

eform_option causes the coefficient table to be displayed in exponentiated form; see [R] [eform_option](#). *command* determines which *eform_option* is allowed (`eform(string)` and `eform` are always allowed).

command determines which of the following are allowed (`eform(string)` and `eform` are always allowed):

<i>eform_option</i>	Description
<code>eform(<i>string</i>)</code>	use <i>string</i> for the column title
<code>eform</code>	exponentiated coefficient, <i>string</i> is <code>exp(b)</code>
<code>hr</code>	hazard ratio, <i>string</i> is <code>Haz. Ratio</code>
<code>shr</code>	subhazard ratio, <i>string</i> is <code>SHR</code>
<code>irr</code>	incidence-rate ratio, <i>string</i> is <code>IRR</code>
<code>or</code>	odds ratio, <i>string</i> is <code>Odds Ratio</code>
<code>rrr</code>	relative-risk ratio, <i>string</i> is <code>RRR</code>

Advanced

`nodrop` prevents observations outside `e(sample)` and the `if` and `in` qualifiers from being dropped before the data are resampled.

`reject(exp)` identifies an expression that indicates when results should be rejected. When *exp* is true, the resulting values are reset to missing values.

The following option is available with `jackknife` but is not shown in the dialog box:

`coeflegend`; see [R] [estimation options](#).

Remarks and examples

[stata.com](http://www.stata.com)

Remarks are presented under the following headings:

- [Introduction](#)
- [Using jackknife](#)
- [Jackknifed standard deviation](#)
- [Collecting multiple statistics](#)
- [Collecting coefficients](#)

Introduction

Although the jackknife—developed in the late 1940s and early 1950s—is of largely historical interest today, it is still useful in searching for overly influential observations. This feature is often forgotten. In any case, the jackknife is

- an alternative, first-order unbiased estimator for a statistic;
- a data-dependent way to calculate the standard error of the statistic and to obtain significance levels and confidence intervals; and
- a way of producing measures called pseudovalues for each observation, reflecting the observation’s influence on the overall statistic.

The idea behind the simplest form of the jackknife—the one implemented here—is to repeatedly calculate the statistic in question, each time omitting just one of the dataset’s observations. Assume that our statistic of interest is the sample mean. Let y_j be the j th observation of our data on some measurement y , where $j = 1, \dots, N$ and N is the sample size. If \bar{y} is the sample mean of y using the entire dataset and $\bar{y}_{(j)}$ is the mean when the j th observation is omitted, then

$$\bar{y} = \frac{(N - 1)\bar{y}_{(j)} + y_j}{N}$$

Solving for y_j , we obtain

$$y_j = N\bar{y} - (N - 1)\bar{y}_{(j)}$$

These are the pseudovalues that `jackknife` calculates. To move this discussion beyond the sample mean, let $\hat{\theta}$ be the value of our statistic (not necessarily the sample mean) using the entire dataset, and let $\hat{\theta}_{(j)}$ be the computed value of our statistic with the j th observation omitted. The pseudovalue for the j th observation is

$$\hat{\theta}_j^* = N\hat{\theta} - (N - 1)\hat{\theta}_{(j)}$$

The mean of the pseudovalues is the alternative, first-order unbiased estimator mentioned above, and the standard error of the mean of the pseudovalues is an estimator for the standard error of $\hat{\theta}$ (Tukey 1958).

The jackknife estimate of variance has been largely replaced by the bootstrap estimate (see [R] [bootstrap](#)), which is widely viewed as more efficient and robust. The use of jackknife pseudovalues to detect outliers is too often forgotten and is something the bootstrap does not provide. See Mosteller and Tukey (1977, 133–163) and Mooney and Duval (1993, 22–27) for more information.

Using jackknife

Typing

```
. jackknife exp_list: command
```

executes *command* once for each observation in the dataset, leaving the associated observation out of the calculations that make up *exp_list*.

command defines the statistical command to be executed. Most Stata commands and user-written programs can be used with `jackknife`, as long as they follow standard Stata syntax and allow the `if` qualifier; see [U] [11 Language syntax](#). The `by` prefix may not be part of *command*.

exp_list specifies the statistics to be collected from the execution of *command*. If *command* changes the contents in `e(b)`, *exp_list* is optional and defaults to `_b`.


```

. clear
. input x
      x
1. 0.1
2. 0.1
3. 0.1
4. 0.4
5. 0.5
6. 1.0
7. 1.1
8. 1.3
9. 1.9
10. 1.9
11. 4.7
12. end

. jackknife sd=r(sd), rclass keep: summarize x
(running summarize on estimation sample)

Jackknife replications (11)
-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
.....

Jackknife results
Number of obs      =      11
Replications       =      11

command: summarize x
sd: r(sd)
n(): r(N)

```

	Coef.	Jackknife Std. Err.	t	P> t	[95% Conf. Interval]	
sd	1.343469	.624405	2.15	0.057	-.047792	2.73473

Interpreting the output, the standard deviation reported by `summarize mpg` is 1.34. The jackknife standard error is 0.62. The 95% confidence interval for the standard deviation is -0.048 to 2.73 .

By specifying `keep`, `jackknife` creates in our dataset a new variable, `sd`, for the pseudovalues.

```
. list, sep(4)
```

	x	sd
1.	.1	1.139977
2.	.1	1.139977
3.	.1	1.139977
4.	.4	.8893147
5.	.5	.824267
6.	1	.632489
7.	1.1	.6203189
8.	1.3	.6218889
9.	1.9	.835419
10.	1.9	.835419
11.	4.7	7.703949

The jackknife estimate is the average of the `sd` variable, so `sd` contains the individual values of our statistic. We can see that the last observation is substantially larger than the others. The last observation is certainly an outlier, but whether that reflects the considerable information it contains or indicates that it should be excluded from analysis depends on the context of the problem. Here Mosteller

Collecting multiple statistics

▷ Example 4

jackknife is not limited to collecting just one statistic. For instance, we can use `summarize`, `detail` and then obtain the jackknife estimate of the standard deviation and skewness. `summarize`, `detail` stores the standard deviation in `r(sd)` and the skewness in `r(skewness)`, so we might type

```
. use http://www.stata-press.com/data/r14/auto, clear
(1978 Automobile Data)
. jackknife sd=r(sd) skew=r(skewness), rclass: summarize mpg, detail
(running summarize on estimation sample)

Jackknife replications (74)
-----|----- 1 -----|----- 2 -----|----- 3 -----|----- 4 -----|----- 5
..... 50
.....

Jackknife results                                Number of obs   =       74
                                                Replications   =       74

      command: summarize mpg, detail
             sd: r(sd)
             skew: r(skewness)
             n(): r(N)
```

	Coef.	Jackknife Std. Err.	t	P> t	[95% Conf. Interval]	
sd	5.785503	.6072509	9.53	0.000	4.575254	6.995753
skew	.9487176	.3367242	2.82	0.006	.2776272	1.619808

◀

Collecting coefficients

▷ Example 5

jackknife can also collect coefficients from estimation commands. For instance, using `auto.dta`, we might wish to obtain the jackknife standard errors of the coefficients from a regression in which we model the mileage of a car by its weight and trunk space. To do this, we could refer to the coefficients as `_b[weight]`, `_b[trunk]`, `_se[weight]`, and `_se[trunk]` in the `exp_list`, or we could simply use the extended expressions `_b`. In fact, `jackknife` assumes `_b` by default when used with estimation commands.

```
. use http://www.stata-press.com/data/r14/auto
(1978 Automobile Data)
```

```
. jackknife: regress mpg weight trunk
(running regress on estimation sample)
```

```
Jackknife replications (74)
```

```
-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
1       2       3       4       5
..... 50
.....
```

```
Linear regression      Number of obs      =      74
                      Replications      =      74
                      F( 2,      73)     =      78.10
                      Prob > F          =      0.0000
                      R-squared         =      0.6543
                      Adj R-squared     =      0.6446
                      Root MSE        =      3.4492
```

mpg	Coef.	Jackknife Std. Err.	t	P> t	[95% Conf. Interval]	
weight	-.0056527	.0010216	-5.53	0.000	-.0076887	-.0036167
trunk	-.096229	.1486236	-0.65	0.519	-.3924354	.1999773
_cons	39.68913	1.873324	21.19	0.000	35.9556	43.42266

If you are going to use `jackknife` to estimate standard errors of model coefficients, we recommend using the `vce(jackknife)` option when it is allowed with the estimation command; see [R] [vce_option](#).

```
. regress mpg weight trunk, vce(jackknife, nodots)
```

```
Linear regression      Number of obs      =      74
                      Replications      =      74
                      F( 2,      73)     =      78.10
                      Prob > F          =      0.0000
                      R-squared         =      0.6543
                      Adj R-squared     =      0.6446
                      Root MSE        =      3.4492
```

mpg	Coef.	Jackknife Std. Err.	t	P> t	[95% Conf. Interval]	
weight	-.0056527	.0010216	-5.53	0.000	-.0076887	-.0036167
trunk	-.096229	.1486236	-0.65	0.519	-.3924354	.1999773
_cons	39.68913	1.873324	21.19	0.000	35.9556	43.42266

◀

John Wilder Tukey (1915–2000) was born in Massachusetts. He studied chemistry at Brown and mathematics at Princeton and afterward worked at both Princeton and Bell Labs, as well as being involved in a great many government projects, consultancies, and committees. He made outstanding contributions to several areas of statistics, including time series, multiple comparisons, robust statistics, and exploratory data analysis. Tukey was extraordinarily energetic and inventive, not least in his use of terminology: he is credited with inventing the terms *bit* and *software*, in addition to *ANOVA*, *boxplot*, *data analysis*, *hat matrix*, *jackknife*, *stem-and-leaf plot*, *trimming*, and *winsorizing*, among many others. Tukey's direct and indirect impacts mark him as one of the greatest statisticians of all time.

Stored results

jackknife stores the following in `e()`:

Scalars

<code>e(N)</code>	sample size
<code>e(N_reps)</code>	number of complete replications
<code>e(N_misreps)</code>	number of incomplete replications
<code>e(N_clust)</code>	number of clusters
<code>e(k_eq)</code>	number of equations in <code>e(b)</code>
<code>e(k_extra)</code>	number of extra equations
<code>e(k_exp)</code>	number of expressions
<code>e(k_eeexp)</code>	number of extended expressions (<code>_b</code> or <code>_se</code>)
<code>e(df_r)</code>	degrees of freedom

Macros

<code>e(cmdname)</code>	command name from <i>command</i>
<code>e(cmd)</code>	same as <code>e(cmdname)</code> or <code>jackknife</code>
<code>e(command)</code>	<i>command</i>
<code>e(cmdline)</code>	command as typed
<code>e(prefix)</code>	<code>jackknife</code>
<code>e(wtype)</code>	weight type
<code>e(wexp)</code>	weight expression
<code>e(title)</code>	title in estimation output
<code>e(cluster)</code>	cluster variables
<code>e(pseudo)</code>	new variables containing pseudovalues
<code>e(nfunction)</code>	<code>e(N)</code> , <code>r(N)</code> , <code>n()</code> option, or empty
<code>e(exp#)</code>	expression for the <i>#</i> th statistic
<code>e(mse)</code>	from <code>mse</code> option
<code>e(vce)</code>	<code>jackknife</code>
<code>e(vcetype)</code>	title used to label Std. Err.
<code>e(properties)</code>	<code>b V</code>

Matrices

<code>e(b)</code>	observed statistics
<code>e(b_jk)</code>	jackknife estimates
<code>e(V)</code>	jackknife variance–covariance matrix
<code>e(V_modelbased)</code>	model-based variance

When `exp_list` is `_b`, `jackknife` will also carry forward most of the results already in `e()` from *command*.

Methods and formulas

Let $\hat{\theta}$ be the observed value of the statistic, that is, the value of the statistic calculated using the original dataset. Let $\hat{\theta}_{(j)}$ be the value of the statistic computed by leaving out the *j*th observation (or cluster); thus $j = 1, 2, \dots, N$ identifies an individual observation (or cluster), and *N* is the total number of observations (or clusters). The *j*th pseudovalue is given by

$$\hat{\theta}_j^* = \hat{\theta}_{(j)} + N\{\hat{\theta} - \hat{\theta}_{(j)}\}$$

When the `mse` option is specified, the standard error is estimated as

$$\hat{s}\hat{e} = \left\{ \frac{N-1}{N} \sum_{j=1}^N (\hat{\theta}_{(j)} - \hat{\theta})^2 \right\}^{1/2}$$

and the jackknife estimate is

$$\bar{\theta}_{(\cdot)} = \frac{1}{N} \sum_{j=1}^N \hat{\theta}_{(j)}$$

Otherwise, the standard error is estimated as

$$\widehat{\text{se}} = \left\{ \frac{1}{N(N-1)} \sum_{j=1}^N (\hat{\theta}_j^* - \bar{\theta}^*)^2 \right\}^{1/2} \quad \bar{\theta}^* = \frac{1}{N} \sum_{j=1}^N \hat{\theta}_j^*$$

where $\bar{\theta}^*$ is the jackknife estimate. The variance–covariance matrix is similarly computed.

References

- Brillinger, D. R. 2002. John W. Tukey: His life and professional contributions. *Annals of Statistics* 30: 1535–1575.
- Gould, W. W. 1995. [sg34: Jackknife estimation](#). *Stata Technical Bulletin* 24: 25–29. Reprinted in *Stata Technical Bulletin Reprints*, vol. 4, pp. 165–170. College Station, TX: Stata Press.
- Mooney, C. Z., and R. D. Duval. 1993. *Bootstrapping: A Nonparametric Approach to Statistical Inference*. Newbury Park, CA: Sage.
- Mosteller, C. F., and J. W. Tukey. 1977. *Data Analysis and Regression: A Second Course in Statistics*. Reading, MA: Addison–Wesley.
- Tukey, J. W. 1958. Bias and confidence in not-quite large samples. Abstract in *Annals of Mathematical Statistics* 29: 614.

Also see

- [R] [jackknife postestimation](#) — Postestimation tools for jackknife
- [R] [bootstrap](#) — Bootstrap sampling and estimation
- [R] [permute](#) — Monte Carlo permutation tests
- [R] [simulate](#) — Monte Carlo simulations
- [SVY] [svy jackknife](#) — Jackknife estimation for survey data
- [U] [13.5 Accessing coefficients and standard errors](#)
- [U] [13.6 Accessing results from Stata commands](#)
- [U] [20 Estimation and postestimation commands](#)