Title

loneway — Large one-way ANOVA, random effects, and reliability

Syntax

`loneway` *response_var group_var* [*if*] [*in*] [*weight*] [, *options*]

options	Description
Main	
`mean`	expected value of F distribution; default is 1
`median`	median of F distribution; default is 1
`exact`	exact confidence intervals (groups must be equal with no weights)
`level(#)`	set confidence level; default is `level(95)`

`by` is allowed; see [D] **by**.

`aweight`s are allowed; see [U] **11.1.6 weight**.

Menu

Statistics > Linear models and related > ANOVA/MANOVA > Large one-way ANOVA

Description

`loneway` fits one-way analysis-of-variance (ANOVA) models on datasets with many levels of *group_var* and presents different ancillary statistics from `oneway` (see [R] **oneway**):

Feature	`oneway`	`loneway`
Fit one-way model	x	x
on fewer than 376 levels	x	x
on more than 376 levels		x
Bartlett's test for equal variance	x	
Multiple-comparison tests	x	
Intragroup correlation and SE		x
Intragroup correlation confidence interval		x
Est. reliability of group-averaged score		x
Est. SD of group effect		x
Est. SD within group		x

Options

Main

`mean` specifies that the expected value of the $F_{k-1,N-k}$ distribution be used as the reference point F_m in the estimation of ρ instead of the default value of 1.

`median` specifies that the median of the $F_{k-1,N-k}$ distribution be used as the reference point F_m in the estimation of ρ instead of the default value of 1.

`exact` requests that exact confidence intervals be computed, as opposed to the default asymptotic confidence intervals. This option is allowed only if the groups are equal in size and weights are not used.

`level(#)` specifies the confidence level, as a percentage, for confidence intervals of the coefficients. The default is `level(95)` or as set by `set level`; see **[U] 20.7 Specifying the width of confidence intervals**.

Remarks and examples

stata.com

Remarks are presented under the following headings:

The one-way ANOVA model

▷ Example 1

`loneway`'s output looks like that of `oneway`, except that `loneway` presents more information at the end. Using our automobile dataset, we have created a (numeric) variable called `manufacturer_grp` identifying the manufacturer of each car, and within each manufacturer we have retained a maximum of four models, selecting those with the lowest mpg. We can compute the intraclass correlation of mpg for all manufacturers with at least four models as follows:

```
. use http://www.stata-press.com/data/r13/auto7
(1978 Automobile Data)

. loneway mpg manufacturer_grp if nummake == 4
              One-way Analysis of Variance for mpg: Mileage (mpg)

                                              Number of obs =        36
                                                  R-squared =    0.5228

    Source                SS         df      MS            F     Prob > F
-------------------------------------------------------------------------

Between manufactur~p   621.88889      8    77.736111      3.70     0.0049
Within manufactur~p       567.75     27    21.027778
-------------------------------------------------------------------------

Total                  1189.6389     35    33.989683

         Intraclass       Asy.
         correlation      S.E.       [95% Conf. Interval]
         ------------------------------------------------

            0.40270     0.18770      0.03481      0.77060

         Estimated SD of manufactur~p effect      3.765247
         Estimated SD within manufactur~p         4.585605
         Est. reliability of a manufactur~p mean   0.72950
              (evaluated at n=4.00)
```

◁

In addition to the standard one-way ANOVA output, `loneway` produces the R-squared, the estimated standard deviation of the group effect, the estimated standard deviation within group, the intragroup correlation, the estimated reliability of the group-averaged mean, and, for unweighted data, the asymptotic standard error and confidence interval for the intragroup correlation.

R-squared

The R-squared is, of course, simply the underlying R^2 for a regression of *response_var* on the levels of *group_var*, or `mpg` on the various manufacturers here.

The random-effects ANOVA model

`loneway` assumes that we observe a variable, y_{ij}, measured for n_i elements within k groups or classes such that

$$y_{ij} = \mu + \alpha_i + \epsilon_{ij}, \quad i = 1, 2, \ldots, k, \quad j = 1, 2, \ldots, n_i$$

and α_i and ϵ_{ij} are independent zero-mean random variables with variance σ^2_α and σ^2_ϵ, respectively. This is the random-effects ANOVA model, also known as the components-of-variance model, in which it is typically assumed that the y_{ij} are normally distributed.

The interpretation with respect to our example is that the observed value of our response variable, `mpg`, is created in two steps. First, the ith manufacturer is chosen, and a value, α_i, is determined—the typical `mpg` for that manufacturer less the overall `mpg` μ. Then a deviation, ϵ_{ij}, is chosen for the jth model within this manufacturer. This is how much that particular automobile differs from the typical `mpg` value for models from this manufacturer.

For our sample of 36 car models, the estimated standard deviations are $\sigma_\alpha = 3.8$ and $\sigma_\epsilon = 4.6$. Thus a little more than half of the variation in `mpg` between cars is attributable to the car model, with the rest attributable to differences between manufacturers. These standard deviations differ from those that would be produced by a (standard) fixed-effects regression in that the regression would require the sum within each manufacturer of the ϵ_{ij}, $\epsilon_{i.}$ for the ith manufacturer, to be zero, whereas these estimates merely impose the constraint that the sum is *expected* to be zero.

Intraclass correlation

There are various estimators of the intraclass correlation, such as the pairwise estimator, which is defined as the Pearson product-moment correlation computed over all possible pairs of observations that can be constructed within groups. For a discussion of various estimators, see Donner (1986). `loneway` computes what is termed the analysis of variance, or ANOVA, estimator. This intraclass correlation is the theoretical upper bound on the variation in *response_var* that is explainable by *group_var*, of which R-squared is an overestimate because of the serendipity of fitting. This correlation is comparable to an R-squared—you do not have to square it.

In our example, the intra-`manu` correlation, the correlation of `mpg` within manufacturer, is 0.40. Because `aweights` were not used and the default correlation was computed (that is, the `mean` and `median` options were not specified), `loneway` also provided the asymptotic confidence interval and standard error of the intraclass correlation estimate.

Estimated reliability of the group-averaged score

The estimated reliability of the group-averaged score or mean has an interpretation similar to that of the intragroup correlation; it is a comparable number if we average *response_var* by *group_var*, or `mpg` by `manu` in our example. It is the theoretical upper bound of a regression of manufacturer-averaged `mpg` on characteristics of manufacturers. Why would we want to collapse our 36-observation dataset into a 9-observation dataset of manufacturer averages? Because the 36 observations might be a mirage. When General Motors builds cars, do they sometimes put a Pontiac label and sometimes a Chevrolet label on them, so that it appears in our data as if we have two cars when we really have only one, replicated? If that were the case, and if it were the case for many other manufacturers, then we would be forced to admit that we do not have data on 36 cars; we instead have data on nine manufacturer-averaged characteristics.

Stored results

`loneway` stores the following in `r()`:

Scalars

`r(N)`	number of observations	`r(rho_t)`	estimated reliability
`r(rho)`	intraclass correlation	`r(se)`	asymp. SE of intraclass correlation
`r(lb)`	lower bound of 95% CI for rho	`r(sd_w)`	estimated SD within group
`r(ub)`	upper bound of 95% CI for rho	`r(sd_b)`	estimated SD of group effect

Methods and formulas

The mean squares in the `loneway`'s ANOVA table are computed as

$$\mathrm{MS}_\alpha = \sum_i w_{i\cdot}(\overline{y}_{i\cdot} - \overline{y}_{\cdot\cdot})^2/(k-1)$$

and

$$\mathrm{MS}_\epsilon = \sum_i \sum_j w_{ij}(y_{ij} - \overline{y}_{i\cdot})^2/(N-k)$$

in which

$$w_{i\cdot} = \sum_j w_{ij} \quad w_{\cdot\cdot} = \sum_i w_{i\cdot} \quad \overline{y}_{i\cdot} = \sum_j w_{ij} y_{ij}/w_{i\cdot} \quad \text{and} \quad \overline{y}_{\cdot\cdot} = \sum_i w_{i\cdot}\overline{y}_{i\cdot}/w_{\cdot\cdot}$$

The corresponding expected values of these mean squares are

$$E(\mathrm{MS}_\alpha) = \sigma_\epsilon^2 + g\sigma_\alpha^2 \qquad \text{and} \qquad E(\mathrm{MS}_\epsilon) = \sigma_\epsilon^2$$

in which

$$g = \frac{w_{\cdot\cdot} - \sum_i w_{i\cdot}^2/w_{\cdot\cdot}}{k-1}$$

In the unweighted case, we get

$$g = \frac{N - \sum_i n_i^2/N}{k-1}$$

As expected, $g = m$ for the case of no weights and equal group sizes in the data, that is, $n_i = m$ for all i. Replacing the expected values with the observed values and solving yields the ANOVA estimates of σ_α^2 and σ_ϵ^2. Substituting these into the definition of the intraclass correlation

$$\rho = \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_\epsilon^2}$$

yields the ANOVA estimator of the intraclass correlation:

$$\rho_A = \frac{F_{\mathrm{obs}} - 1}{F_{\mathrm{obs}} - 1 + g}$$

F_{obs} is the observed value of the F statistic from the ANOVA table. For no weights and equal n_i, ρ_A = roh, which is the intragroup correlation defined by Kish (1965). Two slightly different estimators are available through the `mean` and `median` options (Gleason 1997). If either of these options is specified, the estimate of ρ becomes

$$\rho = \frac{F_{\text{obs}} - F_m}{F_{\text{obs}} + (g-1)F_m}$$

For the `mean` option, $F_m = E(F_{k-1,N-K}) = (N-k)/(N-k-2)$, that is, the expected value of the ANOVA table's F statistic. For the `median` option, F_m is simply the median of the F statistic. Setting F_m to 1 gives ρ_A, so for large samples, these different point estimators are essentially the same. Also, because the intraclass correlation of the random-effects model is by definition nonnegative, for any of the three possible point estimators, ρ is truncated to zero if F_{obs} is less than F_m.

For no weighting, interval estimators for ρ_A are computed. If the groups are equal sized (all n_i equal) and the `exact` option is specified, the following exact (assuming that the y_{ij} are normally distributed) $100(1-\alpha)$% confidence interval is computed:

$$\left\{ \frac{F_{\text{obs}} - F_m F_u}{F_{\text{obs}} + (g-1)F_m F_u}, \frac{F_{\text{obs}} - F_m F_l}{F_{\text{obs}} + (g-1)F_m F_l} \right\}$$

with $F_m = 1$, $F_l = F_{\alpha/2,k-1,N-k}$, and $F_u = F_{1-\alpha/2,k-1,N-k}$, $F_{\cdot,k-1,N-k}$ being the cumulative distribution function for the F distribution with $k-1$ and $N-k$ degrees of freedom. If `mean` or `median` is specified, F_m is defined as above. If the groups are equal sized and `exact` is not specified, the following asymptotic $100(1-\alpha)$% confidence interval for ρ_A is computed,

$$\left[\rho_A - z_{\alpha/2}\sqrt{V(\rho_A)}, \rho_A + z_{\alpha/2}\sqrt{V(\rho_A)} \right]$$

where $z_{\alpha/2}$ is the $100(1-\alpha/2)$ percentile of the standard normal distribution and $\sqrt{V(\rho_A)}$ is the asymptotic standard error of ρ defined below. This confidence interval is also available for unequal groups. It is not applicable and, therefore, not computed for the estimates of ρ provided by the `mean` and `median` options. Again, because the intraclass coefficient is nonnegative, if the lower bound is negative for either confidence interval, it is truncated to zero. As might be expected, the coverage probability of a truncated interval is higher than its nominal value.

The asymptotic standard error of ρ_A, assuming that the y_{ij} are normally distributed, is also computed when appropriate, namely, for unweighted data and when ρ_A is computed (neither the `mean` option nor the `median` option is specified):

$$V(\rho_A) = \frac{2(1-\rho)^2}{g^2}(A+B+C)$$

with

$$A = \frac{\{1+\rho(g-1)\}^2}{N-k}$$

$$B = \frac{(1-\rho)\{1+\rho(2g-1)\}}{k-1}$$

$$C = \frac{\rho^2\{\sum n_i^2 - 2N^{-1}\sum n_i^3 + N^{-2}(\sum n_i^2)^2\}}{(k-1)^2}$$

and ρ_A is substituted for ρ (Donner 1986).

The estimated reliability of the group-averaged score, known as the Spearman–Brown prediction formula in the psychometric literature (Winer, Brown, and Michels 1991, 1014), is

$$\rho_t = \frac{t\rho}{1+(t-1)\rho}$$

for group size t. `loneway` computes ρ_t for $t = g$.

The estimated standard deviation of the group effect is $\sigma_\alpha = \sqrt{(\mathrm{MS}_\alpha - \mathrm{MS}_\epsilon)/g}$. This deviation comes from the assumption that an observation is derived by adding a group effect to a within-group effect.

The estimated standard deviation within group is the square root of the mean square due to error, or $\sqrt{\mathrm{MS}_\epsilon}$.

Acknowledgment

We thank John Gleason of Syracuse University (retired) for his contributions to improving `loneway`.

References

Donner, A. 1986. A review of inference procedures for the intraclass correlation coefficient in the one-way random effects model. *International Statistical Review* 54: 67–82.

Gleason, J. R. 1997. sg65: Computing intraclass correlations and large ANOVAs. *Stata Technical Bulletin* 35: 25–31. Reprinted in *Stata Technical Bulletin Reprints*, vol. 6, pp. 167–176. College Station, TX: Stata Press.

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

Marchenko, Y. V. 2006. Estimating variance components in Stata. *Stata Journal* 6: 1–21.

Winer, B. J., D. R. Brown, and K. M. Michels. 1991. *Statistical Principles in Experimental Design*. 3rd ed. New York: McGraw–Hill.

Also see

[R] **anova** — Analysis of variance and covariance

[R] **icc** — Intraclass correlation coefficients

[R] **oneway** — One-way analysis of variance