The Stata Journal

The *Stata Journal* publishes reviewed papers together with shorter notes or comments, regular columns, book reviews, and other material of interest to Stata users. Examples of the types of papers include 1) expository papers that link the use of Stata commands or programs to associated principles, such as those that will serve as tutorials for users first encountering a new field of statistics or a major new technique; 2) papers that go "beyond the Stata manual" in explaining key features or uses of Stata that are of interest to intermediate or advanced users of Stata; 3) papers that discuss new commands or Stata programs of interest either to a wide spectrum of users (e.g., in data management or graphics) or to some large segment of Stata users (e.g., in survey statistics, survival analysis, panel analysis, or limited dependent variable modeling); 4) papers analyzing the statistical properties of new or existing estimators and tests in Stata; 5) papers that could be of interest or usefulness to researchers, especially in fields that are of practical importance but are not often included in texts or other journals, such as the use of Stata in managing datasets, especially large datasets, with advice from hard-won experience; and 6) papers of interest to those who teach, including Stata with topics such as extended examples of techniques and interpretation of results, simulations of statistical concepts, and overviews of subject areas.

The *Stata Journal* is indexed and abstracted by *CompuMath Citation Index*, *Current Contents/Social and Behavioral Sciences*, *RePEc: Research Papers in Economics*, *Science Citation Index Expanded* (also known as *SciSearch*, *Scopus*, and *Social Sciences Citation Index*.

For more information on the *Stata Journal*, including information for authors, see the webpage

http://www.stata-journal.com

Subscriptions are available from StataCorp, 4905 Lakeway Drive, College Station, Texas 77845, telephone 979-696-4600 or 800-STATA-PC, fax 979-696-4601, or online at

http://www.stata.com/bookstore/sj.html

Subscription rates listed below include both a printed and an electronic copy unless otherwise mentioned.

U.S. and Canada		Elsewhere	
Printed & electronic		**Printed & electronic**	
1-year subscription	$ 98	1-year subscription	$138
2-year subscription	$165	2-year subscription	$245
3-year subscription	$225	3-year subscription	$345
1-year student subscription	$ 75	1-year student subscription	$ 99
1-year institutional subscription	$245	1-year institutional subscription	$285
2-year institutional subscription	$445	2-year institutional subscription	$525
3-year institutional subscription	$645	3-year institutional subscription	$765
Electronic only		**Electronic only**	
1-year subscription	$ 75	1-year subscription	$ 75
2-year subscription	$125	2-year subscription	$125
3-year subscription	$165	3-year subscription	$165
1-year student subscription	$ 45	1-year student subscription	$ 45

Back issues of the *Stata Journal* may be ordered online at

http://www.stata.com/bookstore/sjj.html

Individual articles three or more years old may be accessed online without charge. More recent articles may be ordered online.

http://www.stata-journal.com/archives.html

The *Stata Journal* is published quarterly by the Stata Press, College Station, Texas, USA.

Address changes should be sent to the *Stata Journal*, StataCorp, 4905 Lakeway Drive, College Station, TX 77845, USA, or emailed to sj@stata.com.

The Stata Journal (2013)
13, Number 4, pp. 876–883

A score test for group comparisons in single-index models

Paulo Guimarães
Universidade do Porto
Porto, Portugal
pguimaraes@fep.up.pt

Abstract. In this article, I derive a score test for the equality of one or more parameters across groups of observations following estimation of a single-index model. The test has a wide array of applications and nests Pearson's chi-squared test as a particular case. The postestimation command `scoregrp` implements the test and works with `logit`, `logistic`, `probit`, `poisson`, or `regress` (see [R] **logit**, [R] **logistic**, [R] **probit**, [R] **poisson**, and [R] **regress**). Finally, I show some applications of the test.

Keywords: st0321, scoregrp, score test, logit, logistic, probit, Poisson, regress

1 Introduction

In many practical situations after estimation of a regression model, there is interest in performing a test for equality of one or more parameters across groups of observations. The analysis of variance (ANOVA) and analysis of covariance models are probably the better-known examples, but many other situations fall under this general description. For example, one may want to implement a test to decide whether to include a factor variable or an interaction with a factor variable as a regressor. Other generic examples are tests of structural change when one wants to decide whether to impose a single model to the pooled data or estimate the model in each subsample separately. Yet another example is the situation where one wants to decide whether a panel-data estimator or even a mixed model is more appropriate. Some goodness-of-fit tests are also based on the comparison of estimated parameters across groups of observations (for example, the goodness-of fit test for the logistic regression proposed by Tsiatis [1980]).

For models estimated by maximum likelihood, three asymptotically equivalent tests may be used for hypothesis testing: the likelihood-ratio test (LRT), the Wald test, or the score (or Lagrange multiplier) test. The score test has the advantage of requiring only estimation of the restricted model, that is, estimation of the model under the null hypothesis. This advantage is particularly relevant in situations when it becomes computationally expensive to estimate the unrestricted model. The score test has better small-sample properties than the Wald test (Boos 1992; Fears, Benichou, and Gail 1996) and is effective relative to the LRT (Godfrey 1981). However, in practice, the score test is rarely used because it lacks a general estimation command such as Stata's `lrtest` (see [R] **lrtest**) for the LRT or `test` (see [R] **test**) for the Wald test.

As we will see, a score test for the equality of parameters across groups following estimation of single-index models is easy to implement. Moreover, I will also show that for some particular situations, this test is identical to Pearson's chi-squared test applied to individual-level data. In the following section, I derive the test and show its relation with Pearson's chi-squared test. Next I present the Stata command `scoregrp`, which implements the test after estimation with `logit`, `logistic`, `probit`, `poisson`, or `regress` (see [R] **logit**, [R] **logistic**, [R] **probit**, [R] **poisson**, and [R] **regress**). Finally, I illustrate the use of `scoregrp` in some examples.

2 Score tests for group effects

2.1 The score test

Suppose that we have specified a probability model for a dependent variable Y and have a collection of n independent and identically distributed observations. Further, admit that the observations of Y may be classified into G mutually exclusive groups, each group with n_g observations and $g = 1, \ldots, G$. Assume that for the ith observation of group g, the expected value of Y is a known function of μ_{ig}; that is, $E(y_{ig}) = g(\mu_{ig})$. The index μ_{ig} is a linear combination of covariates; that is, $\mu_{ig} = \mathbf{x}'_{ig}\theta$, where $\mathbf{x}_{ig}$ is a vector of the observed covariates for the ith observation on group g, and $\boldsymbol{\theta'} = [\theta_1, \theta_2, \ldots, \theta_k]$ is a $k \times 1$ vector of unknown parameters associated with the $\mathbf{x}$ covariates.

If we let the known density function for Y be represented by $f(y;\boldsymbol{\theta})$, then we can write the likelihood function as

$$L(\boldsymbol{\theta};\mathbf{Y}) = \prod_{g=1}^{G}\prod_{i=1}^{n_g} f(\boldsymbol{\theta}; y_{ig}) \tag{1}$$

where y_{ig} is the ith observation of Y on group g. The maximum likelihood estimates are the values of $\boldsymbol{\theta}$ that maximize (1). They are obtained by solving the k equations that result from differentiating the logarithm of the likelihood function with respect to $\boldsymbol{\theta}$. Thus the maximum likelihood estimates $\widehat{\boldsymbol{\theta}}$ are those values of $\boldsymbol{\theta}$ such that

$$\mathbf{s}(\boldsymbol{\theta}) = \sum_{g=1}^{G}\sum_{i=1}^{n_g} \mathbf{s}_{ig}(\boldsymbol{\theta}) = \sum_{g=1}^{G}\sum_{i=1}^{n_g} \frac{\partial \ln f(\boldsymbol{\theta}; y_{ig})}{\partial \mu}\mathbf{x}_{ig} = \mathbf{0} \tag{2}$$

For $\widehat{\boldsymbol{\theta}}$ to be a maximum likelihood estimate, the matrix of second derivatives of the log-likelihood function, the Hessian matrix, evaluated at $\widehat{\boldsymbol{\theta}}$, must be negative definite. This matrix equals

$$\mathbf{H} = \sum_{g=1}^{G}\sum_{i=1}^{n_g} \mathbf{H}_{ig} = \sum_{g=1}^{G}\sum_{i=1}^{n_g} \frac{\partial^2 \ln f\left(\widehat{\boldsymbol{\theta}}; y_{ig}\right)}{\partial \boldsymbol{\mu}^2}\mathbf{x}_{ig}\mathbf{x}'_{ig} \tag{3}$$

The vectors $\mathbf{s}$ and $\mathbf{s}_{ig}$ have a dimension of $k \times 1$. To refer to the element of the vector $\mathbf{s}_{ig}$ that is associated with a specific coefficient, say, coefficient θ_j, we will use the generic notation $s_{\theta_j,ig}$. Similarly, $\mathbf{H}$ and $\mathbf{H}_{ig}$ are $k \times k$ matrices, and the notation $h_{\theta_j\theta_l,ig}$ will refer to the specific element of matrix $\mathbf{H}_{ig}$ that corresponds to the coefficients θ_j and θ_l. At times, I will give a different interpretation to a subscripted matrix, but the intended meaning should be clear from the context.

Suppose now that one wants to test the equality of a subset of the parameters (say, a total of k_1 parameters) across groups of observations. Without loss of generality, admit that $\boldsymbol{\theta}' = [\boldsymbol{\alpha}', \boldsymbol{\beta}']$ and that $\boldsymbol{\alpha}$ is a vector containing all the parameters to be tested. Our null hypothesis is then

$$H_o : \boldsymbol{\alpha}_1 = \boldsymbol{\alpha}_2 = \cdots = \boldsymbol{\alpha}_G$$

Implementing Rao's score test for this hypothesis leads to the statistic

$$T = \mathbf{s}\left(\widehat{\boldsymbol{\vartheta}}\right)' \left[-\mathbf{H}\left(\widehat{\boldsymbol{\vartheta}}\right)\right]^{-1} \mathbf{s}\left(\widehat{\boldsymbol{\vartheta}}\right) \tag{4}$$

where $\mathbf{s}(\widehat{\boldsymbol{\vartheta}})'$ is a score vector calculated with respect to all the coefficients implied by the alternative hypothesis but evaluated at the maximum likelihood solution obtained under the null hypothesis. Thus $\boldsymbol{\vartheta}' = [\boldsymbol{\alpha}_1', \boldsymbol{\alpha}_2', \ldots, \boldsymbol{\alpha}_G'; \boldsymbol{\beta}']$ is the "expanded" set of coefficients that is consistent with the alternative hypothesis. Under the null hypothesis, the score test in (4) is asymptotically approximated by a chi-squared distribution with $k_1(G-1)$ degrees of freedom. Partitioning the score vector and Hessian matrix in (4) with respect to the two sets of coefficients, $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, we can rewrite (4) as

$$T = -\begin{bmatrix} \mathbf{s}_\alpha \\ \mathbf{s}_\beta \end{bmatrix}' \begin{bmatrix} \mathbf{H}_{\alpha\alpha} & \mathbf{H}_{\alpha\beta} \\ \mathbf{H}_{\beta\alpha} & \mathbf{H}_{\beta\beta} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{s}_\alpha \\ \mathbf{s}_\beta \end{bmatrix}$$

The second set of score values evaluated at the restricted estimates is 0; thus $\mathbf{s}_\beta(\widehat{\vartheta}) = \mathbf{0}$. Hence, using the well-known result on the inverse of partitioned matrices, we can rewrite (4) as

$$T = -\mathbf{s}_\alpha' \left[\mathbf{H}_{\alpha\alpha} - \mathbf{H}_{\alpha\beta}\left[\mathbf{H}_{\beta\beta}\right]^{-1}\mathbf{H}_{\beta\alpha}\right]^{-1} \mathbf{s}_\alpha \tag{5}$$

The important thing to note is that all matrices in (5) are easily obtained following estimation of the restricted model. The matrix $\mathbf{H}_{\beta\beta}$ is the Hessian matrix of the restricted model obtained by excluding the rows and columns corresponding to the parameters in $\boldsymbol{\alpha}$. The other matrices are obtained as partial sums of the observation-level components of the Hessian and gradient vectors shown in (2) and (3). The expression for the score test in (5) may be presented in an alternative way, which will prove useful in subsequent analysis. Using a known result on matrix identities (see, for example, Demidenko [2004, 651]), we can restate the test statistic on (5) as

$$T = -\mathbf{s}_\alpha' \left[\mathbf{H}_{\alpha\alpha}^{-1} + \mathbf{H}_{\alpha\alpha}^{-1}\mathbf{H}_{\alpha\beta}\left[\mathbf{H}_{\beta\beta} - \mathbf{H}_{\beta\alpha}\mathbf{H}_{\alpha\alpha}^{-1}\mathbf{H}_{\alpha\beta}\right]^{-1}\mathbf{H}_{\beta\alpha}\mathbf{H}_{\alpha\alpha}^{-1}\right]\mathbf{s}_\alpha$$

or more succinctly as

$$T = -\mathbf{s}_\alpha'\mathbf{H}_{\alpha\alpha}^{-1}\mathbf{s}_\alpha + \Delta \tag{6}$$

where

$$\Delta = -\mathbf{s}'_{\alpha}\mathbf{H}_{\alpha\alpha}^{-1}\mathbf{H}_{\alpha\beta}\left[\mathbf{H}_{\beta\beta} - \mathbf{H}_{\beta\alpha}\mathbf{H}_{\alpha\alpha}^{-1}\mathbf{H}_{\alpha\beta}\right]^{-1}\mathbf{H}_{\beta\alpha}\mathbf{H}_{\alpha\alpha}^{-1}\mathbf{s}_{\alpha}$$

The matrix $\mathbf{H}_{\alpha\alpha}$ is block diagonal; it is thus easily invertible regardless of the number of groups because it only requires the inversion of the diagonal matrices that have dimension k_1. The other matrix that needs to be inverted has the dimension of $\boldsymbol{\beta}$ (that is, a dimension equal to the number of covariates not tested). In practical applications, it may be simpler to define a matrix $\mathbf{G}$ with dimensions $n \times G$, where the gth column is a vector with elements that take the value 1 if the observation belongs to group g and 0 otherwise. Letting $\mathbf{X}$ be a matrix containing all covariates in the model and $\mathbf{M}$ be a diagonal matrix with generic element $h_{\alpha\alpha,ig}$, then we can write all the matrices that go into the formula for the test as $\mathbf{H}_{\alpha\alpha} = \mathbf{G}'\mathbf{M}\mathbf{G}$, $\mathbf{H}_{\alpha\beta} = \mathbf{G}'\mathbf{M}\mathbf{X}$, and $\mathbf{H}_{\beta\beta} = \mathbf{X}'\mathbf{M}\mathbf{X}$.

2.2 Relationship with the Pearson χ^2 statistic

To further explore the relation with the Pearson χ^2 statistic, let us now consider the situation where one wants to test whether the constant of a regression model differs across groups. In this case, $k_1 = 1$ and we can rewrite (6) as

$$T = -\sum_{g=1}^{G} \frac{s_{\alpha,\bullet g}^2}{h_{\alpha\alpha,\bullet g}} + \Delta \tag{7}$$

where the symbol "$\bullet$" is used to represent a summation across all elements of i. The above expression makes obvious the relationship between our test and Pearson's χ^2 test. Without covariates, $\Delta = 0$, and the score test for the equality of the intercept across groups of observations becomes the Pearson χ^2 test.

Poisson regression

Consider a typical Poisson regression model with expected value

$$\lambda_{ig} = \exp(\alpha + \mathbf{x}'_{ig}\boldsymbol{\beta})$$

To implement the score test in the Poisson regression model, we need to note that the generic elements for the score vector are $s_{\alpha,ig} = y_{ig} - \widehat{\lambda}_{ig}$ and for the $\mathbf{M}$ matrix are $h_{\alpha\alpha,ig} = -\widehat{\lambda}_{ig}$.

If the regression model has no covariates, then $\widehat{\lambda}_{ig} = \overline{y}$ and $\Delta = 0$. If we plug these values into (7), then we obtain the well-known Pearson χ^2 test for count data.

$$T = \sum_{g=1}^{G} \frac{n_g(\overline{y}_g - \overline{y})^2}{\overline{y}}$$

Logit regression

Consider now a typical logistic regression with binary dependent variable:

$$\text{Prob}(y_{ig} = 1|\mathbf{x}) = \Lambda_{ig} = \frac{\exp(\alpha + \mathbf{x}'_{ig}\boldsymbol{\beta})}{1 + \exp(\alpha + \mathbf{x}'_{ig}\boldsymbol{\beta})}$$

Now the generic elements for the score vector are $s_{\alpha,ig} = y_{ig} - \widehat{\Lambda}_{ig}$ and for the $\mathbf{M}$ matrix are $h_{\alpha\alpha,ig} = -\widehat{\Lambda}_{ig}(1 - \widehat{\Lambda}_{ig})$. If we let p denote the proportion of 1s in the total sample and let p_g denote the proportion of 1s in each subgroup, then in a model without covariates, the test simplifies to

$$T = \sum_{g=1}^{G} \frac{n_g(p_g - p)^2}{p(1-p)} \tag{8}$$

which is the known Pearson χ^2 test for binary data.

Linear regression

Finally, let us consider a typical linear regression model such as

$$y_{ig} = \alpha + \mathbf{x}'_{ig}\boldsymbol{\beta} + u_{ig}$$

where u_{ig} is normally independent and identically distributed with 0 expected value and variance equal to σ^2. The elements of the score vector are $s_{\alpha,ig} = (y_{ig} - \widehat{y_{ig}})/\sigma^2$ and those of the Hessian are $h_{\alpha\alpha,ig} = -\sigma^{-2}$. Without covariates, $\widehat{y_{ig}} = \overline{y}$ and the test simplifies to

$$T = \sum_{g=1}^{G} \frac{n_g(\overline{y}_g - \overline{y})^2}{\sigma^2} \tag{9}$$

which is identical to the one-way ANOVA formula. However, in this circumstance, the test will not produce the same result as the usual ANOVA because the test uses the maximum likelihood estimate of σ^2. As a curiosity, I note that when applied to binary data, (9) produces the same results as the Pearson test for binary data in (8).

3 The scoregrp command

The `scoregrp` command is a user-written command for Stata that implements the test described above after estimation with the commands `logit`, `logistic`, `probit`, `poisson`, or `regress`. It is partially implemented in Mata and requires installation of the user-written command `matdelrc`, programmed by Nicholas J. Cox. Because of the way `scoregrp` is programmed, the command should work well in situations when the number of groups is very large. Additionally, incorporating other single-index models into `scoregrp` should be a straightforward task requiring only the coding of the score and Hessian for the new models.

3.1 Syntax

The command has a very simple syntax:

`scoregrp` [*indepvars*] , `group(`*varname*`)` [`nocons`]

The argument *indepvars* consists of a list of the variables whose coefficients we want to test. By default, it is assumed that the constant is included among *indepvars*, but we can exclude it with the `nocons` option.

3.2 Options

`group(`*varname*`)` specifies the variable that identifies the group. `group()` is required.

`nocons` specifies that the constant not be included among the coefficients to be tested.

4 Examples

To illustrate the use of `scoregrp`, let us use `union.dta`, downloaded from the Stata website. After reading in the data, we start by implementing Pearson's χ^2 to test whether the proportion of unionized individuals remains constant over time.

```
. webuse union
(NLS Women 14-24 in 1968)
. tabulate union year, nofreq chi2
          Pearson chi2(11) = 107.8144   Pr = 0.000
```

The same result is obtained if we run a logit regression without covariates and test for differences in the constant term across years.

```
. quietly logit union
. scoregrp, group(year)
Score test for logit regression
Test result is chi(11) = 107.8144  Pr = 0.0000
```

Next let us consider a logit regression with three covariates, `age`, `grade`, and `black`, and again test for differences in the constant term across years.

```
. quietly logit union age grade black
. scoregrp, group(year)
Score test for logit regression
Test result is chi(11) = 97.6887   Pr = 0.0000
```

The results clearly reject the null hypothesis, and thus we include yearly dummy variables in the logit regression. In the following, we check whether to include an interaction between the variables `grade` and `black`.

```
. quietly tabulate year, generate(y)
. quietly logit union age grade black y1-y11
. scoregrp grade, group(black) nocons
Score test for logit regression
Test result is chi(1) = 0.6684    Pr = 0.4136
```

The hypothesis that the coefficient on the interaction is 0 is not rejected. The following test is akin to a test of permanence of structure and compares whether all coefficients in the two subsamples defined by the variable `south` are identical.

```
. quietly logit union age grade black y1-y11
. scoregrp age grade black y1-y11, group(south)
Score test for logit regression
Test result is chi(15) = 789.5522  Pr = 0.0000
```

Finally, we use `scoregrp` to test whether one should account for unobserved heterogeneity across individuals.

```
. quietly logit union age grade black y1-y11
. scoregrp, group(idcode)
Score test for logit regression
Test result is chi(4433) = 1.46e+04  Pr = 0.0000
```

The results suggest that the data have substantial unobserved heterogeneity.

5 Conclusion

In this article, I derived a score test to check whether one or more coefficients differ across groups of observations following estimation of a single-index model. The user-written command `scoregrp` is a Stata implementation of the test. The present version of the command works after `logit`, `logistic`, `probit`, `poisson`, or `regress` and may be easily extended to other single-index models.

For many practical applications, `scoregrp` offers no computational advantage and can be slower than existing alternatives based on LRT or Wald tests. But with large datasets and particularly when the unrestricted model is complex (for example, a random-effects or mixed model), then `scoregrp` is likely to be the faster approach. Researchers may also want to use `scoregrp` in situations when an LRT or a Wald test is not an option. Consider the cases of panel-data estimators for logit and Poisson regression with fixed effects. In these cases, a Wald or an LRT test to check whether one should include the fixed effect is not possible, because the alternative model is estimated by conditional maximum likelihood. As shown earlier, implementation of the test with `scoregrp` is straightforward.

6 Acknowledgment

I thank João Santos Silva for helpful comments on a previous version of this article.

7 References

Boos, D. D. 1992. On generalized score tests. *American Statistician* 46: 327–333.

Demidenko, E. 2004. *Mixed Models: Theory and Applications*. New York: Wiley.

Fears, T. R., J. Benichou, and M. H. Gail. 1996. A reminder of the fallibility of the Wald statistic. *American Statistician* 50: 226–227.

Godfrey, L. G. 1981. On the invariance of the Lagrange multiplier test with respect to certain changes in the alternative hypothesis. *Econometrica* 49: 1443–1455.

Tsiatis, A. A. 1980. A note on a goodness-of-fit test for the logistic regression model. *Biometrika* 67: 250–251.

About the author

Paulo Guimarães is an associate professor in the Department of Economics at the University of Porto in Portugal.