The Chi-Square Diagnostic Test for Count Data Models

M. Manjón-Antolín and O. Martínez-Ibañez

QURE-CREIP Department of Economics, Rovira i Virgili University.

2012 *Spanish Stata Users Group Meeting*

(Universitat de Barcelona, September 12).

The Pearson chi-squared goodness-of-fit test is a diagnostic tests implemented in Stata as a post estimation command, **estat gof**, to be used after **logit**, **logistic**, **probit** and **poisson** commands.

The Pearson chi-squared goodness-of-fit test is a diagnostic tests implemented in Stata as a post estimation command, **estat gof**, to be used after **logit**, **logistic**, **probit** and **poisson** commands.

Moreover, the *group()* option yields the related Hosmer–Lemeshow test.

The Pearson chi-squared goodness-of-fit test is a diagnostic tests implemented in Stata as a post estimation command, **estat gof**, to be used after **logit**, **logistic**, **probit** and **poisson** commands.

Moreover, the *group()* option yields the related Hosmer–Lemeshow test.

However, the Pearson and Hosmer–Lemeshow tests assume that the estimated coefficients are known.

The Pearson chi-squared goodness-of-fit test is a diagnostic tests implemented in Stata as a post estimation command, **estat gof**, to be used after **logit**, **logistic**, **probit** and **poisson** commands.

Moreover, the *group()* option yields the related Hosmer–Lemeshow test.

However, the Pearson and Hosmer–Lemeshow tests assume that the estimated coefficients are known.

To control for the potential estimation error, Cameron and Trivedi (2009) suggest using the Chi-Square Diagnostic Test developed by Andrews (1988a, 1988b).

This Chi-Square Diagnostic Test compares the sample relative frequencies of the dependent variable with the predicted frequencies from the model using a quadratic form and an estimate of the asymptotic variance of the corresponding population moment condition.

This Chi-Square Diagnostic Test compares the sample relative frequencies of the dependent variable with the predicted frequencies from the model using a quadratic form and an estimate of the asymptotic variance of the corresponding population moment condition.

In contrast to the classical Pearson's test (or the Hosmer–Lemeshow test), the Chi-Square Diagnostic Test can be constructed from any regular, asymptotically normal estimator of the conditional expectation of the dependent variable.

This Chi-Square Diagnostic Test compares the sample relative frequencies of the dependent variable with the predicted frequencies from the model using a quadratic form and an estimate of the asymptotic variance of the corresponding population moment condition.

In contrast to the classical Pearson's test (or the Hosmer–Lemeshow test), the Chi-Square Diagnostic Test can be constructed from any regular, asymptotically normal estimator of the conditional expectation of the dependent variable.

However, to date this *m*-test is not available in Stata.

This paper discusses the implementation of the Chi-square Diagnostic Test of Andrews (1988a, 1988b) in count data models as a Stata post-estimation command.

This paper discusses the implementation of the Chi-square Diagnostic Test of Andrews (1988a, 1988b) in count data models as a Stata post-estimation command.

In particular, **chisqdt** can be used right after **poisson**, **nbreg**, **zip** and **zinb** commands.

This paper discusses the implementation of the Chi-square Diagnostic Test of Andrews (1988a, 1988b) in count data models as a Stata post-estimation command.

In particular, **chisqdt** can be used right after **poisson**, **nbreg**, **zip** and **zinb** commands.

The new command, **chisqdt**, reports the test statistic and its p-value.

Also, one may obtain a table with the actual, predicted and absolute differences between actual and predicted probabilities.

Let us consider a model given by $f(y|\mathbf{w}, \theta)$, the conditional density of the variable of interest (y) given a set of covariates ($\mathbf{w}$) and a vector of parameters (θ).

In particular, we are interested in the conditional density of the Poisson, Negative Binomial, Zero-Inflated Poisson and Zero-Inflated negative binomial models. Thus, $\mathbf{w} = \mathbf{x}$ in the Poisson and Negative Binomial models and $\mathbf{w} = \{\mathbf{x}, \mathbf{z}\}$ in the inflated versions

Also, let J be the number of (mutually exclusive) cells in which the range of the dependent variable y_i is partitioned ($i = 1, \cdots, N$).

Let us consider a model given by $f(y|\mathbf{w},\theta)$, the conditional density of the variable of interest (y) given a set of covariates ($\mathbf{w}$) and a vector of parameters (θ).

In particular, we are interested in the conditional density of the Poisson, Negative Binomial, Zero-Inflated Poisson and Zero-Inflated negative binomial models. Thus, $\mathbf{w} = \mathbf{x}$ in the Poisson and Negative Binomial models and $\mathbf{w} = \{\mathbf{x}, \mathbf{z}\}$ in the inflated versions

Also, let J be the number of (mutually exclusive) cells in which the range of the dependent variable y_i is partitioned ($i = 1, \cdots, N$).

Lastly, let $d_{ij}(y_i) = \mathbf{1}(y_i \in j)$ be an indicator variable that takes value one if observation i belongs to cell j and zero otherwise.

If the model is correctly specified, then

$$E[d_{ij}(y_i) - p_{ij}(\mathbf{w}_i, \theta)] = 0,$$

where $p_{ij}(\mathbf{w}_i, \theta)$ is the probability that observation i falls in cell j according to $f(y|\mathbf{w}, \theta)$.

If the model is correctly specified, then

$$E[d_{ij}(y_i) - p_{ij}(\mathbf{w}_i, \theta)] = 0,$$

where $p_{ij}(\mathbf{w}_i, \theta)$ is the probability that observation i falls in cell j according to $f(y|\mathbf{w}, \theta)$.

In particular, stacking all J moments in vector notation we obtain

$$E[d_i(y_i) - p_i(\mathbf{w}_i, \theta)] = 0.$$

Given a sample analog:

$$\hat{m}_N(\hat{\theta}) = \frac{1}{N}\sum_{i=1}^{N}[d_i(y_i) - p_i(\mathbf{w}_i, \hat{\theta})],$$

Given a sample analog:

$$\hat{m}_N(\hat{\theta}) = \frac{1}{N}\sum_{i=1}^{N}[d_i(y_i) - p_i(\mathbf{w}_i, \hat{\theta})],$$

the Chi-Square Diagnostic Test of Andrews (1988a, 1988b) is

$$chisqdt = N\hat{m}_N(\hat{\theta})\hat{V}^{-1}\hat{m}_N(\hat{\theta}).$$

where V is a variance-covariance matrix given by $\sqrt{N}\hat{m}_N(\hat{\theta}) \rightarrow N(0, V)$.

Under the null hypothesis that the moment condition holds, the **chisqdt** test is asymptotically χ^2-distributed with $rank[V]$ degrees of freedom.

Under the null hypothesis that the moment condition holds, the **chisqdt** test is asymptotically χ^2-distributed with $rank[V]$ degrees of freedom.

However, V may not be of full rank. Actually, the rank is usually $J-1$ because the sum of the probabilities over all J cells is one.

Under the null hypothesis that the moment condition holds, the **chisqdt** test is asymptotically χ^2-distributed with $rank[V]$ degrees of freedom.

However, V may not be of full rank. Actually, the rank is usually $J-1$ because the sum of the probabilities over all J cells is one.

Moreover, the computation of this variance-covariance matrix is often complicated.

This is why when using maximum likelihood estimation it is the outer product of the gradient form of the test what it is usually computed.

This is why when using maximum likelihood estimation it is the outer product of the gradient form of the test what it is usually computed.

This is just N times the (uncentered) R^2 of the following auxiliary regression:

$$1 = \hat{m}_i \delta + \hat{s}_i \gamma + u_i,$$

where 1 is a column vector of N ones, $\hat{m}_i$ includes $d_{ij}(y_i) - p_{ij}(\mathbf{w}_i, \hat{\theta}^{ML})$ for $j = 1, \ldots, J-1$ and $\hat{s}_i = \left. \frac{\partial \log f(y_i|\mathbf{w}_i, \theta)}{\partial \theta} \right|_{\theta = \hat{\theta}^{ML}}$ is the matrix of contributions to the score evaluated at the maximum likelihood estimate of θ.

In particular, it is easy to see that

$$chisqdt = N \times R^2 = 1'H(H'H)^{-1}H'1,$$

where $H_i = [\hat{m}_i, \hat{s}_i]$ is the $i - th$ row of matrix H.

This asymptotically equivalent version of (7) is the one used in the computation of **chisqdt**.

Notice that all is needed to compute the test are the predicted probabilities (p_{ij}) and the scores ($\hat{s}_i$). The paper provides detailed formulae; see also Greene (1994), Cameron and Trivedi (1998) and Cameron and Trivedi (2005).

In particular, it is easy to see that

$$chisqdt = N \times R^2 = 1'H(H'H)^{-1}H'1,$$

where $H_i = [\hat{m}_i, \hat{s}_i]$ is the $i - th$ row of matrix H.

This asymptotically equivalent version of (7) is the one used in the computation of **chisqdt**.

Notice that all is needed to compute the test are the predicted probabilities (p_{ij}) and the scores ($\hat{s}_i$). The paper provides detailed formulae; see also Greene (1994), Cameron and Trivedi (1998) and Cameron and Trivedi (2005).

Under the null hypothesis of correct specification of the model, this statistic asymptotically follows a χ^2 distribution with $J - 1$ degrees of freedom.

The syntax of the command is the following:

chisqdt, *cells(#) [prcount] [table]*

where *cells* is the number of (mutually exclusive) cells in which one partitions the range of the dependent variable to compute the test.

The syntax of the command is the following:

chisqdt, *cells(#) [prcount] [table]*

where *cells* is the number of (mutually exclusive) cells in which one partitions the range of the dependent variable to compute the test.

In principle, any partition of the dependent variable can be used.

For example, if one uses three cells the following partitions can be used: $\{0, 1, 2, 3\}$, $\{4, 5\}$ and $\{6, 7, \ldots, \infty\}$; $\{0, 1\}$, $\{2, 3, 4, 5\}$ and $\{6, 7, \ldots, \infty\}$; $\{0, 1, 2, 3, 4, 5\}$, $\{6\}$ and $\{7, 8, \ldots, \infty\}$; etc.

However, for simplicity **chisqdt** only considers partitions with single-value elements (except for the last cell).

However, for simplicity **chisqdt** only considers partitions with single-value elements (except for the last cell).

That is, **chisqdt** uses partitions like $\{0\}$ and $\{1, 2, 3, \ldots, \infty\}$; $\{0\}$, $\{1\}$ and $\{2, 3, \ldots, \infty\}$; $\{0\}$, $\{1\}$, $\{2\}$ and $\{3, 4, \ldots, \infty\}$; and so on.

However, for simplicity **chisqdt** only considers partitions with single-value elements (except for the last cell).

That is, **chisqdt** uses partitions like $\{0\}$ and $\{1, 2, 3, \ldots, \infty\}$; $\{0\}$, $\{1\}$ and $\{2, 3, \ldots, \infty\}$; $\{0\}$, $\{1\}$, $\{2\}$ and $\{3, 4, \ldots, \infty\}$; and so on.

In general, for *cells(J)*, the partition **chisqdt** uses is $\{0\}$, $\{1\}$, $\{2\}$, ..., $\{J-2\}$ and $\{J-1, \ldots, \infty\}$.

Options	Description
prcount	Uses **prcounts** to compute predicted probabilities; default is `direct` calculation.
table	A table with the actual, predicted and absolute differences between actual and predicted frequencies is reported.

The option `prcounts` refers to the way of computing the probability that, according to the model, a particular value of the dependent variable belongs to one of the defined cells.

The option `prcounts` refers to the way of computing the probability that, according to the model, a particular value of the dependent variable belongs to one of the defined cells.

By default the program calculates these predicted probabilities (or predicted frequencies) using the definition of the conditional density of the dependent variable (`direct`).

The option `prcounts` refers to the way of computing the probability that, according to the model, a particular value of the dependent variable belongs to one of the defined cells.

By default the program calculates these predicted probabilities (or predicted frequencies) using the definition of the conditional density of the dependent variable (`direct`).

However, one may alternatively compute these probabilities using the program **prcounts** of Long and Freese (2001, Stata *Journal* 1).

The option `prcounts` refers to the way of computing the probability that, according to the model, a particular value of the dependent variable belongs to one of the defined cells.

By default the program calculates these predicted probabilities (or predicted frequencies) using the definition of the conditional density of the dependent variable (`direct`).

However, one may alternatively compute these probabilities using the program **prcounts** of Long and Freese (2001, Stata *Journal* 1).

In general, results barely change when using one or the other.

Differences do arise, however, when the number of counts is high, particularly if the (zero-inflated) negative binomial model is used.

Differences do arise, however, when the number of counts is high, particularly if the (zero-inflated) negative binomial model is used.

In that case, one receives an error message informing that "Missing values encountered when "prcount" option is used (try "direct" option)".

Differences do arise, however, when the number of counts is high, particularly if the (zero-inflated) negative binomial model is used.

In that case, one receives an error message informing that "Missing values encountered when "prcount" option is used (try "direct" option)".

One also receives an error message when the statistic may not be computed for the (zero-inflated) negative binomial model because the α parameter is too small: "Problem with alpha prevents estimation of predicted probabilities (alpha too small)".

Differences do arise, however, when the number of counts is high, particularly if the (zero-inflated) negative binomial model is used.

In that case, one receives an error message informing that "Missing values encountered when "prcount" option is used (try "direct" option)".

One also receives an error message when the statistic may not be computed for the (zero-inflated) negative binomial model because the α parameter is too small: "Problem with alpha prevents estimation of predicted probabilities (alpha too small)".

Ultimately, both error messages arise because of the large numbers that the gamma function generates.

The option `table` produces a table with the actual, predicted and absolute differences between actual and predicted frequencies.

The option `table` produces a table with the actual, predicted and absolute differences between actual and predicted frequencies.

This can be useful in assessing the adequacy of the partition of the dependent variable we are using. As the examples will show, this may e.g. help detecting cells with too few observations.

The option `table` produces a table with the actual, predicted and absolute differences between actual and predicted frequencies.

This can be useful in assessing the adequacy of the partition of the dependent variable we are using. As the examples will show, this may e.g. help detecting cells with too few observations.

Also, the table may provide insights about the source of misspecification. In the **poisson** model, for example, big absolute differences in the zero value may indicate overdispersion.

We illustrate the use of the new command and the interpretation of its output we three examples.

We illustrate the use of the new command and the interpretation of its output we three examples.

The first example merely replicates results from chapters 5–6 of Cameron and Trivedi (1998). This is the one we report here.

The second and third examples replicate and extend results reported in chapter 17 of Cameron and Trivedi (2009).

We illustrate the use of the new command and the interpretation of its output we three examples.

The first example merely replicates results from chapters 5–6 of Cameron and Trivedi (1998). This is the one we report here.

The second and third examples replicate and extend results reported in chapter 17 of Cameron and Trivedi (2009).

In all the cases we report the output resulting from both the estimation command (**poisson**, **nbreg**, **zip** or **zinb**) and the new command (**chisqdt**).

In particular, in the first example we also report the table with the actual, predicted and absolute differences between actual and predicted frequencies (option `table`).

- **Exemple 1**.

Cameron and Trivedi (1998) analyse the determinants of takeover bids using a sample of 126 US firms that were taken over between 1978 and 1985.

The dependent variable is the number of bids received by the firm after the initial tender offer (`numbids`), while covariates include defensive actions taken by the management of the firm (`leglrest`, `realrest`, `finrest` and `whtknght`), firm-specific characteristics (`bidprem`, `insthold`, `size` and `sizesq`), and intervention by federal regulators (`regulatn`).

The relation between the dependent and explanatory variables is estimated using the Poisson regression model.

Results can be obtained by typing

```
. infile docno weeks numbids takeover bidprem insthold size
leglrest realrest finrest regulatn whtknght sizesq constant using
http://cameron.econ.ucdavis.edu/racd/racd5.asc, clear
(126 observations read)

. poisson numbids leglrest realrest finrest whtknght bidprem insthold size
sizesq regulatn, nolog
```

Results can be obtained by typing

```
. infile docno weeks numbids takeover bidprem insthold size
leglrest realrest finrest regulatn whtknght sizesq constant using
http://cameron.econ.ucdavis.edu/racd/racd5.asc, clear
(126 observations read)

. poisson numbids leglrest realrest finrest whtknght bidprem insthold size
sizesq regulatn, nolog
```

And the resulting output, including the Chi-square Diagnostic Test with $J = 6$, is

```
Poisson regression                                Number of obs   =        126
                                                  LR chi2(9)      =      33.25
                                                  Prob > chi2     =     0.0001
Log likelihood = -184.94833                       Pseudo R2       =     0.0825
```

numbids	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
leglrest	.2601464	.1509594	1.72	0.085	-.0357286	.5560213
realrest	-.1956597	.1926309	-1.02	0.310	-.5732093	.1818899
finrest	.0740301	.2165219	0.34	0.732	-.3503452	.4984053
whtknght	.4813822	.1588698	3.03	0.002	.170003	.7927613
bidprem	-.6776958	.3767372	-1.80	0.072	-1.416087	.0606956
insthold	-.3619912	.4243292	-0.85	0.394	-1.193661	.4696788
size	.1785026	.0600221	2.97	0.003	.0608614	.2961438
sizesq	-.0075693	.0031217	-2.42	0.015	-.0136878	-.0014509
regulatn	-.0294392	.1605682	-0.18	0.855	-.344147	.2852686
_cons	.9860598	.5339201	1.85	0.065	-.0604044	2.032524

```
. chisqdt, cells(6)
Chi-squared Test for Poisson Model =     48.66 (Prob>chi2 =  0.00)
```

Also, we can obtain the table the actual, predicted and absolute differences between actual and predicted probabilities by typing

```
. chisqdt, cells(6) table
Chi-squared Test for ZIP Model =     94.13 (Prob>chi2 =  0.00)
```

Counts	Actual	Predicted	Abs. Dif.
0	.6328	.6285	.0042
1	.1032	.0373	.0659
2	.0577	.0471	.0106
3	.0516	.0489	.0027
4	.0258	.0455	.0197
5 or more	.129	.1927	.0637

- **Exemple 1 (Continuation)**.

The second application we consider is their analysis of the determinants of the number of recreational boating trips to Lake Somerville, Texas, in 1980 (`trips`).

Covariates include a subjective quality index of the facility (`so`), a dummy variable to indicate practice of water-skiing at the lake (`ski`), the household income of the head of the group (`i`), a dummy variable to indicate whether the user paid a fee (`fc3`), dollar expenditure when visiting Lake Conroe (`c1`), dollar expenditure when visiting Lake Somerville (`educyr`), and dollar expenditure when visiting Lake Houston (`educyr`).

In their analyses Cameron and Trivedi (1998) discuss at length different models (including finite mixtures and hurdle-types of the Poisson and the negative binomial models) and goodness-of-fit measures (the G^2 statistic, the pseudo-R^2, etc.). However, we limit the reported results to the **poisson**, **nbreg** and **zip** estimates and the Chi-Square Diagnostic Test, **chisqdt**.

In their analyses Cameron and Trivedi (1998) discuss at length different models (including finite mixtures and hurdle-types of the Poisson and the negative binomial models) and goodness-of-fit measures (the G^2 statistic, the pseudo-R^2, etc.). However, we limit the reported results to the **poisson**, **nbreg** and **zip** estimates and the Chi-Square Diagnostic Test, **chisqdt**.

In particular, results can be obtained by typing

```
. infile trips so ski i fc3 c1 c3 c4 using http://cameron.econ.ucdavis.edu/racd
> /racd6d2.asc, clear
(659 observations read)
. poisson trips so ski i fc3 c1 c3 c4, nolog
. chisqdt, cells(6)
. nbreg trips so ski i fc3 c1 c3 c4, nolog
. chisqdt, cells(6)
. zip trips so ski i fc3 c1 c3 c4, inflate(so i) nolog
. chisqdt, cells(6)
```

```
Poisson regression                              Number of obs   =        659
                                                LR chi2(7)      =    2543.90
                                                Prob > chi2     =     0.0000
Log likelihood = -1529.4313                     Pseudo R2       =     0.4540
```

trips	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
so	.4717259	.0170905	27.60	0.000	.4382291	.5052227
ski	.4182137	.0571905	7.31	0.000	.3061224	.5303051
i	-.1113232	.0195885	-5.68	0.000	-.1497159	-.0729304
fc3	.8981652	.0789854	11.37	0.000	.7433567	1.052974
c1	-.0034297	.0031178	-1.10	0.271	-.0095405	.0026811
c3	-.0425364	.0016703	-25.47	0.000	-.0458102	-.0392626
c4	.0361336	.0027096	13.34	0.000	.0308229	.0414444
_cons	.2649934	.0937224	2.83	0.005	.0813009	.4486859

```
Chi-squared Test for Poisson Model =    252.57  (Prob>chi2 =   0.00)
```

```
Negative binomial regression                      Number of obs   =        659
                                                  LR chi2(7)      =     478.33
Dispersion     = mean                             Prob > chi2     =     0.0000
Log likelihood = -825.55758                       Pseudo R2       =     0.2246
```

trips	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
so	.721999	.0453323	15.93	0.000	.6331493	.8108487
ski	.6121388	.1504163	4.07	0.000	.3173282	.9069493
i	-.0260589	.0452342	-0.58	0.565	-.1147163	.0625986
fc3	.6691677	.3614399	1.85	0.064	-.0392415	1.377577
c1	.0480086	.0159516	3.01	0.003	.016744	.0792732
c3	-.092691	.0082685	-11.21	0.000	-.1088969	-.0764851
c4	.0388357	.0117139	3.32	0.001	.0158769	.0617945
_cons	-1.121936	.2208284	-5.08	0.000	-1.554752	-.6891205
/lnalpha	.3157293	.1060209			.1079321	.5235264
alpha	1.371259	.1453821			1.113972	1.68797

```
Likelihood-ratio test of alpha=0:  chibar2(01) = 1407.75 Prob>=chibar2 = 0.000
Chi-squared Test for NegBin Model =      23.54 (Prob>chi2 =  0.00)
```

```
Zero-inflated Poisson regression                  Number of obs   =        659
                                                  Nonzero obs     =        242
                                                  Zero obs        =        417

Inflation model = logit                           LR chi2(7)      =     622.01
Log likelihood  = -1180.795                       Prob > chi2     =     0.0000
```

trips	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
so	.0338331	.0239159	1.41	0.157	-.0130412	.0807073
ski	.4716906	.0581895	8.11	0.000	.3576412	.58574
i	-.0997796	.0207787	-4.80	0.000	-.1405052	-.059054
fc3	.6104876	.0794354	7.69	0.000	.4547972	.7661781
c1	.0023689	.0038282	0.62	0.536	-.0051343	.009872
c3	-.0376003	.002039	-18.44	0.000	-.0415966	-.033604
c4	.0252337	.0033666	7.50	0.000	.0186353	.0318321
_cons	2.099162	.1114393	18.84	0.000	1.880745	2.317579

```
(Inflated part omitted)
Chi-squared Test for ZIP Model =     94.13 (Prob>chi2 =  0.00)
```

Andrews, Donald W.K. (1988a): "Chi-square diagnostic tests for econometric models: Theory", *Econometrica* 56: 1419–1453.

Andrews, Donald W.K. (1988b): "Chi-square diagnostic tests for econometric models: Introduction and applications", *Journal of Econometrics* 37: 135–156.

Cameron, A.C. and Trivedi, P.K. (1998): *Regression Analysis of Count Data*, Cambridge University Press.

Cameron, A.C. and Trivedi, P.K. (2005): *Microeconometrics*, CUP.

Cameron, A.C. and Trivedi, P.K. (2009): *Microeconometrics Using* Stata, Stata Press.

Greene, William H. (1994): "Accounting for excess zeros and sample selection in Poisson and negative binomial regression models", WP 94-10, Stern School of Business, Department of Economics.

The Chi-Square Diagnostic Test for Count Data Models

M. Manjón-Antolín and O. Martínez-Ibañez

QURE-CREIP Department of Economics, Rovira i Virgili University.

2012 *Spanish Stata Users Group Meeting*

(Universitat de Barcelona, September 12).