WINPEPI PROGRAMS

POISSON

MANUAL

(Version 1.27)

Revised April 17, 2015

What POISSON does

POISSON is a WINPEPI program (Abramson 2004), part of the PEPI suite of computer programs for epidemiologists. ("PEPI" is an acronym for "Programs for EPIdemiologists".)

This program has a single module, which performs multiple Poisson regression analysis.

WORDS OF CAUTION

It is unwise to use a statistical procedure whose use one does not understand. This manual cannot supply this knowledge, and it is certainly no substitute for the basic understanding of statistics and epidemiological thinking that is essential for the wise choice of methods and the correct interpretation of their results.

FINDER.PDF (provided with this program) is an alphabetical index that identifies the modules (in all WinPepi programs) that deal with a specific procedure or kind of study. It is called up by pressing F9 or clicking on "*Finder*" in any WinPepi program, or on the FINDER icon, and can be printed for easy reference.

How to use POISSON

Running the program: See "**Instructions**", below. The program provides detailed on-screen instructions and help. POISSON can be run in any version of Windows except Windows 3**.**

Recalling results: Click on *"View"* in the top menu to display the current session's previous results

Pasting results: Results shown on the screen are automatically copied to the Windows clipboard, from which they can be pasted into a Microsoft Word or other text file at the site of the cursor (usually by pressing *Shift-Insert* or *Ctrl-V*. To ensure proper alignment of tabulated results, a Courier or similar font should be used in the text file. If the current session's previous results are recalled (by clicking on "*View*"), text can be marked (drag the mouse over it with button pressed) and copied to the clipboard (by pressing *Ctrl-Insert* or *Ctrl-C*) for pasting elsewhere.

Adding comments: Click on "*Note*" in the top menu if you wish to add explanatory comments to be placed in the clipboard, saved, or printed with the results.

Saving results: By default, all results of Pepi-for-Windows programs are saved in PEPI.TXT in the Winpepi folder, with a warning if it exceeds 500K. Results also go to PEPI.TMP (for display in the "*View"* option); this file may be overwritten unless it is renamed on quitting POISSON. Click on **"***Save***"** (in the top menu) to see the default procedure or to change it. Also, currently-shown results can be saved by clicking on "*Print or save*".]Results saved in earlier installations may be found in C:\PEPI.TXT]. TXT files can be combined with **JOINTEXT**, supplied with the Winpepi package.

Printing results: Click on *"Print".* If this fails, try switching the printer off and on again, and click on "*Print*" again. The "*Print*" button sends the results straight to the printer, and with some printers this does not work. A simple solution is to paste the currently-shown results (which have automatically been copied to the Windows clipboard) into a Microsoft Word or other text program, and print from there. To ensure proper alignment of tabulated results, a Courier or similar font should be used in the text file. Results can also be printed from one of the files in which they are automatically saved, e.g. PEPI.TXT (which can be accessed by clicking on "*Results*" in the Winpepi portal).

FINDING WHAT YOU WANT

FINDER.HLP (provided with this program) is an alphabetical index that identifies the modules (in all WinPepi programs) that deal with a specific procedure or kind of study. It is called up by pressing F9 or clicking on "*Finder*" in any WinPepi program, or on the FINDER icon, and can be printed for easy reference.

A DO-IT-YOURSELF THREESOME

1. The WinPepi suite of computer programs for epidemiologists, with their manuals. Can be downloaded free at www.brixtonhealth.com

2. "Research Methods in Community Medicine: Surveys, Epidemiological Research, Programme Evaluation, Clinical Trials" (J.H. Abramson and Z.H. Abramson), sixth edition. Wiley and Sons, 2008.

3. "Making Sense of Data: A Self-Instruction Manual on the Interpretation of Epidemiological Data" (J.H. Abramson and Z.H.Abramson), third edition. Oxford: Oxford University Press 2001.

HOW TO OBTAIN PEPI PROGRAMS

All WINPEPI (PEPI-for-Windows) and other PEPI programs can be downloaded free. The latest versions of WINPEPI programs – currently COMPARE2, DESCRIBE, ETCETERA, LOGISTIC, PAIRSetc, POISSON, and WHATIS – can be downloaded from www.brixtonhealth.com; and the latest release of Version 4 of PEPI, which contains over 40 DOS-based programs (which can be used in Windows) and WHATIS, can be downloaded from www.simtel.net/pub/pd/54632.html

COMPARE2, DESCRIBE, ETCETERA, LOGISTIC, PAIRSetc, and POISSON are distributed with PDFmanuals. A printed manual is available for the DOS-based programs and WHATIS (Abramson and Gahlinger 2001.).

POISSON REGRESSION ANALYSIS

This module performs Poisson regression analysis. It may be used in longitudinal surveys and trials concerned with events whose occurrence is believed to follow a Poisson random distribution. The outcome event may be one that can occur only once (such as death) or one that can happen more than once (e.g. accidents, disease spells, or doctor visits). The program appraises the effects of single variables or combinations of variables on the rate of occurrence of the outcome event (its incidence density), and permits control of confounding effects and appraisal of modifying effects. It is especially appropriate if the outcome event is a rare one. Used in cross-sectional studies, it gives accurate estimates of prevalence ratios, but the confidence intervals may be unduly wide (Barros and Hirakata 2003). For a discussion of the use of Poisson regression in epidemiological studies, see (*inter alia*) McNeill (1996: 170-185).

The program computes the **Poisson regression coefficients** (with their standard errors and *z*-scores) and the corresponding **incidence density rate ratios** (with confidence intervals). For this purpose, the values of the variables in the model are treated as separate nominal categories, using the lowest value as the reference category. In some instances alternative rate ratios are computed, for use in prevalence studies. Three **significance tests** are performed**: Wald, likelihood-ratio and score tests. Goodness of fit** with the Poisson model is appraised by computing the **deviance** and a **Pearson chi-square**, by a graphic display of the relationship between **standardized residua**ls and their expected normal scores, and by **Filliben's test.** The program can use the regression coefficients to compute a rate ratio expressing the contrast between two selected sets of values

Instructions

To use the module, first *enter the data* Data must initially be entered by pasting. The program can then create and store a *data file* (in the PR subfolder in the folder in which POISSON.EXE is situated) for future use.

Then either enter *information about the variables,* or load a *dictionary file* that contains it. The information must initially be entered at the keyboard, and the program can then save it in a dictionary file (in the PR subfolder) for future use. A dictionary file can be used only if it has previously been created and put in this folder by POISSON. The file contains information on the number, names, and sequence of variables, the names of the outcome and denominator variables, and (for each independent variable) the number of categories at the time the file was created, and their cut-points.

After entry of the data, alterations can be made to the variables' *categories,* by entering the desired cut-points.. By default, each value of a variable is treated as a separate

nominal category, to a maximum of 15; if there are more than 15 values, the higher values are placed in a single category.

The *confidence level* for the estimation of confidence intervals (which is set at 95% as a default) can also be altered.

The *model* is then entered (see below), and the program is run.

When the results have appeared, the model or options can be changed, or new data or a new dictionary file can be entered.

The data

To be available for pasting, the data must first be copied into the Windows clipboard (usually by pressing *Ctrl-Insert* or *Ctrl-C*) from the file in which they have been entered, e.g. a simple text file created by Notepad, or a spreadsheet (e.g. Excel).
The data can then be pasted into the empty data-entry box [by pressing *Shift-Insert* or *Ctrl-V* (if the data box is not empty, press *Esc*). Up to 56 kb of data may be pasted.

The data must be in the correct format (see below).

Data can be entered separately for each group for which a count of outcome events is available or, less typically, for each individual (if the number of events per individual is believed to have a Poisson distribution).

Groups or individuals with the same characteristics can be combined before entry, or entered separately. This will not affect the coefficients and significance tests, but is likely to affect goodness of fit. It has been suggested that similar groups should always be combined before entry, unless the number of unique patterns of covariates is small, say 20 or fewer (Sribney 1997).

The variables that must be entered for each group or individual are:

- the number of outcome events experienced by the group or individual;
- the denominator, i.e., the number of person-time units of exposure or, if all groups were at risk for the same period or the duration of exposure is irrelevant, the size of the group; if individual data are entered, the denominator is the individual's period at risk or, if all individuals were exposed for'the same period, 1;
- the independent variables, i.e. the characteristics whose associations with the outcome variable are of interest, or that may confound or modify these associations.

There may be other variables, not required in the analysis, such as case identity numbers.

All the values must be numerical.. The numbers may be continuous or discrete, and they may be numerical labels allotted to categories. Decisions on coding may be affected by awareness that the lowest values will be used as reference categories.

The data must be in the correct format., with the data for each group or individual in a separate line, and with no text or non-numerical entries. The values in each line (up to 40) must be in the same (arbitrarily chosen) order, and separated by spaces. The same number of values must be shown in each line; there can be no missing values .For example, the first two lines might look like this:

```
11  1  1  74  0  0  1.56  1  3  20  4  136
7   1   0  75  0  0  9.03  0  0  3   5  78
```

Exact alignment of the columns is not necessary.

The model

The form of the Poisson regression equation is

$$Y = \exp(a + bB + cC\ldots)$$

where Y is the incidence density rate of the outcome variable; a is a constant; B, C, etc. are the values of the independent variables; and b, c, etc. are their regression coefficients.

In this instance, the model to be entered in the program would be

B C

or, more specifically, the names of the selected independent variables would be entered, e.g.

AGE SMOKING

Optionally, first-degree interactions may be included, as in the model.

AGE SMOKING AGE*SMOKING

The program treats the values as nominal categories (by creating dummy variables), and displays coefficients for separate categories (e.g. for B_1, B_2, and B_3, if these are the categories of variable B, and for B_2C_3, one of the combinations of categories of variables B and C).

If the number of parameters (B_1, B_2, B_2C_3, etc.) in the model is the same as the number of groups for which data were entered, the model is a *saturated* one, with no degrees of freedom; goodness of fit annot be tested. The program displays a warning and offers appropriate suggestions.

Poisson regression coefficients

The program computes Poisson regression coefficients (with their standard errors and z-scores) and the corresponding odds ratios. Results are displayed for each category of every variable, except the first (reference) category. The z-scores (the coefficients

divided by their standard errors) may be used to compare the impact of different variables (Selvin 1996: 262).

The coefficients reflect the influence of the specified category or interaction (i..e., its effect on the log rate), in comparison with the reference category, when other influences (all the other covariates in the model) are held constant. The reference categories have coefficients (not shown) of 0.

Rate ratios

The rate ratios, which are the antilogarithms of the regression coefficients, are the ratios (according to the Poisson equation) of the rate (i.e., the incidence density rate of the outcome variable) in the specified category to the rate in the reference category, when other influences (all the other covariates in the model) are held constant. The rate ratios for the reference categories (not shown) are of course 1. Confidence intervals are displayed.

Crude rate ratios (not controlling for other variables) are displayed for comparison

If Poisson regression is used to appraise prevalence rate ratios in a cross-sectional study, the point estimates will be accurate, but the confidence limits are likely to be wider than those that would be provided by a binomial model. As a simple rough remedy, the program provides alternative rate ratios based on replacement of the Poisson scale parameter of 1 by a value derived from the Pearson chi-square for the model (Barros and Hirakta 2003), an adjustment that provides a degree of correction and may often yield a reasonable approximation (Breslow 1995). These alternative rate ratios are displayed only if there is evidence of underdispersion (see below) and the chi-square has more than 1 degree of freedom.

Significance tests

The P value shown for each coefficient is based on a *Wald test*, which may be over-conservative if the coefficient is large.

Two tests are applied to the total model: a *score test*, recommended for small samples (Breslow and Day 1980: 207- 208), and a *likelihood-ratio test*. A likelihood-ratio test is also applied to the last variable added to the model, in instances where a model is gradually incremented . The likelihood-ratio tests must be regarded as approximate, since they are influenced (although only slightly) by the use of an augmenting factor in the computation (see "*Methods*", below).

Goodness of fit

A Poisson model is not necessarily appropriate, and the results may have little validity if the fit with the observed data is poor.

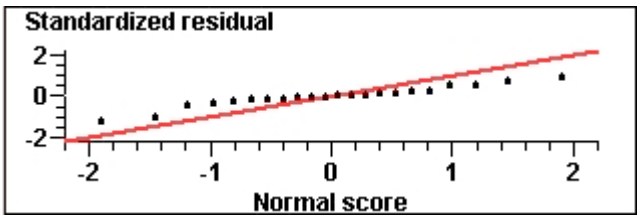
The *deviance*, which is a statistic that compares the model with a saturated model, provides one indication of goodness of fit, provided that the predicted counts are sufficiently large (McNeill 1996: 172); this may be an argument in favour of amalgamating groups with similar characteristics before entering the data (Sribney 1997). If the deviance exceeds its degrees of freedom, the observations vary more than might be expected in a Poisson distribution (*overdispersion* or extra-Poisson variation; the true variance is larger than the mean); and if the deviance is less than its degrees of freedom, there is less variation than might be expected (*underdispersion;* the true variance is smaller than the mean). Overdispersion may occur if relevant explanatory variables are not included in the model, or if there are excess zeroes, e.g. if the number of caries lesions follows a Poisson distribution, but only in children who are susceptible to caries.

A low P value for the deviance points to a poor fit with the Poisson model. But if there are groups with fewer than 5 predicted events, this test is questionable; the program then combines the small groups and repeats the test.. If there are above 50-100 degrees of freedom, the test result should be treated with caution (Sribney 1997).

The program reports the ratio of the deviance to its degrees of freedom if $P < 0.05$ for the deviance.

A *Pearson goodness-of-fit chi-square* is also computed..

Standardized residuals (the differences between the numbers of observed events and the counts predicted by the Poisson equation, divided by their estimated standard errors) are also computed. If there is a good fit, the standardized residuals for the groups (or, if data were entered for individuals, for the individuals) should have a normal distribution, and should coincide with their expected normal scores. This relationship is shown in a graph, where divergence from the red line indicates a poor fit. The graph is a Q-Q (quantum-quantum) graph in which the standardized residuals are plotted against their expected normal scores (i.e., their rankits, the scores expected if the distribution is normal). The graph is of no interest if the model is a saturated one, since all the residuals are then zero.



The graph can be printed, copied to the clipboard for pasting elsewhere, or saved in a bitmap (.BMP) file.

"Outlier" groups that have standardized residuals exceeding 1.96 are identified. It may be decided to omit these groups from a subsequent analysis.

Filliben's test appraises the normality of the standardized residuals by examining their relationship to their expected normal scores. The P value is reported as < 0.05, < 0.1, or > 0.1. A low P value indicates that the correlation coefficient is not sufficiently close to1, and points to a poor fit.

Rate ratios comparing various sets of values

Optionally, the program can use the regression coefficients to compute a rate ratio expressing the contrast between two selected sets of values of some or all of the variables in the model. A variable with the same value in both sets can affect the result only if it is involved in an interaction.

METHODS

POISSON uses the Poisson log-linear equation

$$Y = \exp(a + bB + cC\ldots)$$

(see "*The model*", above).

The analysis uses an approach suggested by Clayton (1988), who pointed out that to perform Poisson regression "any logistic regression program may be used ... by declaring the observed numbers of cases as binomial 'successes' from very large numbers of trials, which must be proportional to the person-years. Choosing a large enough binomial denominator ensures that the binomial likelihood approximates the Poisson..." POISSON uses an algorithm written by Daly (1989) for SAS for the above purpose, applying it in an unconditional logistic regression analysis based on an algorithm appropriate for grouped data, adapted from LOGRESS, which was written by Dan McGee (1986), who has kindly made his Fortran code available.

The augmenting factor used by POISSREG is exp(10), or 22,026. The program converts the information about each group or individual entered into two data records, with the same values for independent variables. In one record, the number of events is 1 and the denominator is the number of events entered (0, 1 or more); in the other, the number of events is 0 and the denominator is the denominator entered (usually person-time units) multiplied by the augmenting factor, with the number of events entered then subtracted. Adjustments are later made to compensate for the use of the augmenting factor; the constant in the Poisson equation, for example, is obtained by adding 10 to the computed logistic coefficient. The augmenting factor is incorporated in the data files created by POISSON, which renders them useless for other purposes.

The program occasionally fails to produce results, and instead reports a computational problem and suggests possible solutions.

Up to 1000 lines of data may be entered, and up to 56k can be pasted. The maximum number of parameters (including dummy variables) in a model is 60.

`

Rate ratios

The rate ratios are the antilogarithms of the regression coefficients.

The alternative confidence intervals for use in cross-sectional studies are estimated by multiplying the variances of the coefficients by a scale parameter derived from the Pearson goodness-of-fit chi-square for the model and its degrees of freedom (*df*):

$$\text{Scale parameter} = \text{chi-square} / df.$$

In an empirical comparison, Barros and Hirakati (2003) found that it was better to use the Pearson chi-square than the deviance for this purpose. If there are groups with fewer than 5 predicted events, the program uses a chi-square value computed after combining the small groups.

Significance tests

The *Wald tests* for the significance of the coefficients use the square of the *z*-score as chi-square, and may be over-conservative for a large coefficient (Hauck and Donner 1977).

Two tests are applied to the total model (Agresti 1996: 94-96): a *score test* and a *likelihood-ratio test*. The latter test, which is based on a comparison of G-statistics, is influenced by the use of an augmenting factor.

A likelihood-ratio test is also applied to the last variable added to the model, for use when a model is gradually incremented

Goodness of fit

The deviance (Agresti 1996: 96) is

$$-2(L_M - L_S)$$

where L_M = maximized log-likelihood value for the model
L_S = maximized log-likelihood value for the saturated model

Its degrees of freedom are the number of groups for which counts of events were entered (or, if data were entered for individuals, the number of individuals) minus the number of nonredundant parameters in the model.

The *Pearson chi-square* (which has the same degrees of freedom) is:

$$\Sigma\,[(O - E)^2 \,/\, E]$$

where O = observed count for group or individual i
E = expected count (based on the regression equation) for group or individual i

A *standardized residual* is the difference between the number of observed events and the count predicted by the Poisson equation, divided by the square root of the predicted count (the estimated standard error).

The graph is a Q-Q (quantum-quantum) graph (Gerson 1975; McNeill 1996: 180) in which the standardized residuals are plotted against their rankits. The rankits, which are the expected values of order statistics of the standard normal distribution corresponding to the data points in a manner determined by the order in which the data points appear, are computed by ranking the standardized residuals by their size and then applying the formula (Blom 1958: 71; Looney and Gulledge 1985).

$$\text{rankit} = (i - 0.375) \,/\, (n + 0.25)$$

where i = rank of the observation
n = the number of observations.

Filliben's test is described by Filliben (1975). The significance of the correlation coefficient between the standardized residuals and their rankits (at the 5% and 10% levels) is determined from a table of percentage points published by Looney and Gulledge (1985).

REFERENCES

Abramson JH (2004) WINPEPI (PEPI-for-Windows) computer programs for epidemiologists. Epidemiologic Perspectives & Innovations, 2004, 1:6 (available on the Internet at www.epi-perspectives.com/content/1/1/6).

Abramson JH, Gahlinger PM (2001) Computer programs for epidemiologists: PEPI version 4. Sagebrush Press: Salt Lake City.

Agresti A (1996) An introduction to categorical data analysis. New York: John Wiley & Sons.

Barros AJD, Hirakata VN (2003) Alternatives for logistic regression in cross-sectional studies: an empirical comparison of models that directly estimate the prevalence ratio. BMC Medical Research Methodology 2003, 3:21.

Blom G (1958) Statistical estimates and transformed beta variables. New York: John Wiley

Breslow NE (1996) Generalized linear models: checking assumptions and strengthening conclusions. Statistica Applicata 8: 23-41.

Breslow NE, Day NE (1980) Statistical methods in cancer research. vol. I. The analysis of case-control studies. Lyon: International Agency for Research on Cancer.

Clayton D (1988). The analysis of event history data: A review of progress and outstanding problems. Statistics in Medicine 7: 819-841.

Daly L (1989). Poisson regression macro & notes. Internet document http//jse.stat.ncsu.edu.70/0/software/sas/poisson.pak

Filliben JJ (1975) The probability plot correlation coefficient test for normality. Technometrics 17: 111-117.

Gerson M (1975) The techniques and uses of probability plotting. The Statistician 24:. 235-257.

Hauck WW, Donner A (1977) Wald's test as applied to hypotheses in logit analysis. Journal of the American Statistical Association 82 :371-386.

Looney SW, Gulledge TR Jr (1985) Use of the correlation coefficient with normal probability plots. The American Statistician 389: 75-7.

McGee DL (1986) A program for logistic regression on the IBM PC. American Journal of Epidemiology 124: 702-705.

McNeill D (1996). Epidemiological Research Methods. New York: John Wiley and Sons.

Selvin S (1996) Statistical analysis of epidemiologic data, 2nd edn. New York: Oxford University Press.

Sribney B (1997) FAQ: Stata 5: Goodness-of-fit chi-squared test reported by poisson, Internet document www.stata.com/support/faqs/stat/poisson.gof.html