WINPEPI PROGRAMS

LOGISTIC

MANUAL

(Version 1.54)

Revised April 19, 2015

What LOGISTIC does

LOGISTIC (previously named MULTI) 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 logistic regression analysis and (optionally) propensity-score 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.

How to use LOGISTIC

Running the program: Click on "*Start*", and then follow the on-screen instructions. (But first select the"propensity-score analysis" option if this is required). if To return to the main menu, click on the "Back to main menu" button. LOGISTIC cannot be run in Windows 3.

Entry of data: Use of **data files**: see p. 6.
Optionally, data can be **pasted** into entry boxes (see next page).
If entries are required in different boxes, pressing *Enter* or *Tab* after entering a number will generally take you to the next box; pressing *Escape* will clear the entry.
If several entries are required in the same box, press *Enter* or *Space* after each entry.

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 LOGISTIC. Click on ***"Saving"*** (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 by using **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).

PASTING DATA

If the data are available in a text file (e.g. a text file created by Notepad) or in a spreadsheet, they can be copied to the Windows clipboard [usually by pressing *Ctrl-Insert* or *Ctrl-C*], and then "pasted" into a data-entry box [usually by pressing *Shift-Insert* or *Ctrl-V*]. This can simplify data entry in boxes that require a number of entries (in rows or columns). [Also, data can be copied from a data-entry box and pasted to a text file for future re-use; press *Ctrl-A* to mark it for copying.]

Precautions:

- The data must be pasted into the box as a single block, and not piecemeal.
- The data must be in the format required in the box, with spaces between the numbers; exact alignment of the columns is not necessary. For example

  45 66 1
  20 3 132
  53 11 44
- If a defined number of rows is required, this number must be entered first, e.g. in the "Number of strata" or "Number of categories" box.

- If row numbers are shown on the left (1, 2, etc.), ensure that the"1" is visible before pasting.
- The cursor must be in the top left corner of the box when the "paste" keys are pressed.

FINDING WHAT YOU WANT

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.

A DO-IT-YOURSELF THREESOME

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

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

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

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 PDF manuals. A printed manual is available for the DOS-based programs and WHATIS (Abramson and Gahlinger 2001.).

MULTIPLE LOGISTIC REGRESSION

J. H. Abramson and Eduardo L. Franco

This program performs multiple logistic regression analysis and (optionally) propensity-score analysis (see below). It may be used in cohort studies and trials that examine the occurrence or non-occurrence of a disease or other outcome, in case-control studies of risk or protective factors associated with a disease or other disorder, in studies that aim to determine how diagnostic or prognostic criteria can be combined to appraise the probability that a disease is present or likely to occur, and for other purposes. The procedure measures the effects of single variables or combinations of variables, and permits control of confounding effects and appraisal of modifying effects.

A number of **CHOICES** are offered: the analysis can be **unconditional or conditional**; **individual or grouped** data can be analysed; data can be entered by **pasting** or by loading a **data file** or at the **keyboard**; the names and sequence of the variables can be entered at the keyboard or by loading a **dictionary file** previously created by this program; the logistic model can include first-degree **interactions**; variables can be treated as **simple** (continuous) or **categorical** (with nominal or ordered categories); variables can be **centered**; values can be defined as **missing**; the analysis can be performed on a **restricted sample;** the **confidence level** can be altered; and an **automatic mode** can be selected, to create a set of models. each with a different main variable but the same covariates. Also, (optionally) **computations using the regression coefficients** can be performed after the analysis..

To use the module, first enter the data, by pasting or by loading a data file, or at the keyboard. Then enter a description of the variables, or load a dictionary file. If required, one or more of the choices listed above can then be made. A **model** (the list of independent variables and interactions to be included in the analysis) is then entered, or specifications can be entered for the automatic creation of a set of models. The program is then run. When the results have appeared, the model or options can be changed, or new data or a new dictionary file can be entered. At all stages, detailed instructions and help are provided on the screen.

The program computes the **logistic regression coefficients** (with their standard errors), the corresponding **odds ratios** (with 90, 95 or 99% confidence intervals), ***Z*-scores**, **crude odds ratios**, the **log-likelihood** for the model and for the null model, and the ***G* statistic** for the model. **Wald, likelihood-ratio** and **score tests** are performed,and the **Hosmer-Lemeshow goodness-of-fit test, Brier skill scores,** indicators of the **discriminatory** power of the model (**Tjur's coefficient of discrimination**, **Hedges' *g***, the ***c*-index**, **gamma**,and the **overlap coefficient**), and **other indicators of the aptness of the logistic model** are provided (namely, the Pearson **correlation coefficient** between the observed values of the dependent variable and the probabilities predicted by the *logistic equation; an estimate of the **proportion of explained variation**; **Darlington's**

logistic regression fit index; and the **pseudo R-squared** value). Optionally (after running the analysis) the logistic regression coefficients can be used to calculate the **probability of the outcome, odds ratios comparing different sets of values.** and the **risk ratio**, **risk difference**, and n**umber needed to treat** (expressing the effect of a selected variable, e.g. treatment or exposure to a risk or protective factor) or re-run to provide **risk ratios** or **prevalence ratios** instead of odds ratios.

If the **propensity-score analysis** option is selected {to examine the effect of a causal variable, e.g. treatment, on a dependent variable) the program reports the overall **average causal effec**t and a **risk ratio** and **odds ratio** based on weighting by inverse propensity scores, the **average causal effect for the exposed**, and a **logistic regression coefficient** and **odds ratio** expressing the effect of the causal variable when controlling for propensity scores. The data must be entered by pasting or at the keyboard.

Instructions

1. A data file can be used only if it has previously been put in an LR subfolder in the folder containing LOGISTIC.EXE. If this subfolder does not exist, it can be created by opening the module and clicking on "Find data file". A test file (LA.DAT) is provided with this program, with its corresponding dictionary file (LA.DIC); to use them, first copy them to the LR folder.

2. After opening the module, either click on "Find data file" (and follow the on-screen instructions), or paste the data. Then click on "OK".

3. Either enter a dictionary file (which supplies the names of the variables and specifies the dependent variable) or describe the variables (i.e., enter their names and specify the dependent variable). Then click on "OK".

4. Optionally, select one or more of the "choices" listed on the screen, to alter the default settings (many of which are determined by the content of the dictionary file). You can decide whether variables are to be treated as simple (continuous) or categorical (with nominal or ordered categories). You can also specify the use of unconditional or conditional analysis and of individual or grouped data, restrict the analysis to a specified subgroup of the sample, define missing values, center variables, alter the confidence level, and change the dependent variable.

5. Optionally, click on "Update/create dictionary" to create a dictionary file (which is automatically placed in the LR folder) or update an existing one. This is advisable if information about the variables has been entered at the keyboard, or if "choices" have been made.

6. Either enter a model, or choose the "Automatically create a set of models" option. To enter a model, just enter each independent variable and interaction (e.g. AGE*SEX), in

any order. The dependent variable and constant need not be specified. The model will be displayed on the screen; e.g. AGE SEX AGE*SEX. If the "automatic" option is selected, a series of models will be created, each including a different "main" variable and the same covariates ("control" variables), all of which must be specified.
+
7. Click on "RUN".

8. When the results have appeared, click on "NEXT" to change the model or options, or to enter new data or a new dictionary file, or (if the "automatic" option was chosen) to run the next model in the series. Optionally, click on "Use results" to calculate the probability of the outcome (except in case-control studies) for a specified set of values, or odds ratios expressing the contrast between two specified sets of values, or the risk ratio, risk difference, and number needed to treat..

At each stage, detailed instructions and help are shown on the screen. Click on "Variables" to view a list of the variables and information about the variables, on "Values" to view information about a specific variable and its values, or on "Numbers" to display descriptive statistics (sample size, numbers of missing values, and the distribution of the dependent variable, for each independent variable or category).

Variables

The variables must be numerical. If a REC file is used (see "Data files", below) the program will translate "Y" and "N" to 1 and 0 respectively, "F" and "M" to 1 and 0, other entries in single-letter fields to 9, and blank fields to 9.

The *dependent variable* may be the occurrence of a disease or other outcome, or (in case-control studies) caseness, i.e. membership of the case or the control group. It is *Y* in the logistic regression equation

log-odds($Y = 1$) = $a + bB + cC + d(B*C)$,

where $Y=1$ refers to a specific category of *Y*, and it may have only two values, 0 and 1. In a cohort study or trial, occurrence of the disease or other outcome should be coded 1; in a case-control study, cases should be coded 1. If other codes are used, the program will report this and offer options for their exclusion or their conversion to 1 or 0 for the purposes of the analysis.

The *independent (explanatory) variables* whose associations with the dependent variable are under study, or that may confound or modify these associations, e.g. B and C in the above equation, may have any numerical values (e.g. continuous or discrete numbers, or numerical labels representing categories).

If grouped data are to be analysed, there must be a *frequency variable* that specifies the number of subjects with identical specified data. If study findings displayed in a cross-tabulation are to be entered in the program, each cell represents a group, and the number in the cell is the frequency variable; zero cells can be ignored.

If conditional logistic regression analysis is to be used, a *matching variable* is required. This variable identifies the matched sets, the same number being allotted to each member of the set. For convenience a decimal point and an extra digit may be added to identify individual members (e.g. 34.1, 34.2, etc. for members of set 34), but this extra digit will be ignored in the analysis.

There may be other numerical variables, e.g. a case identity number, that are not used in the analysis.

Data files

The data file may be a *text (ASCII) file*, containing numbers only, prepared by a word processor or a data-entry or statistical program, or a *REC file* (which specifies the variables' names also) created by EpiData (Lauritsen and Bruus 2003-2004) or Epi Info version 6 (Dean AG *et al*. 1996) after direct entry of the data or after importing a data file created by SPSS or another program.

In a *text file*, the data for each subject or (for grouped data) each group must be in a separate line. The values in each line must be in a defined order, separated by spaces or commas. Vertical alignment is not essential. For example, the first two lines might look like this:

```
1  1  74  0  0  1.56  1  3  96  1  1
2  1  0  75 0  0  9.03  0  0  0  0  0
```

A value must be shown for each variable, with no blanks (a "missing value" code should be used instead). If the data file contains headers or comments before or after the numerical records, they should be deleted. The names and sequence of the variables must be supplied separately, either by entry at the keyboard or by loading a dictionary file previously created by this program.

If a *REC file* is used, it supplies the names and sequence of the variables, and a dictionary file is not required. But it is advisable, however, to afterwards use LOGISTIC s option for the creation of a dictionary file, so that other information about the variables will be available for future use. When LOGISTIC uses a REC file, it translates “Y” and 'N' to “1” and “0” respectively, “F” and “M” to “1” and “0”, other entries in single-letter fields to 9's, and blank fields to 9's. Data in other non-numeric formats, date formats, and records marked as “deleted” are not used. A REC file may not exceed 49kb in size.

Dictionary files

LOGISTIC can create dictionary files, which record information about the variables; it places them in the C:\PEPI\LR folder. The dictionary file stores the names and sequence of the variables, and specifies the dependent variable, the frequency variable (if data are grouped), the matching variable (if data are matched), whether variables are categorical and (if so) the number of categories, the cut-points, and the way the categories are to be

handled in the analysis. LOGISTIC can create a new dictionary file at any time, recording the current set-up of the variables.

Pasting data

Instead of loading a data file, data that have been copied to the Windows clipboard from a text file (created by Notepad or another word processor) or a spreadsheet (e.g. Excel) [usually by clicking on "*Copy*" or by pressing *Ctrl-Ins* or *Ctrl-C*] can be pasted into the data box provided for this purpose, by pressing *Ctrl-V* or *Shift-Ins*. The data box should first be emptied if necessary, by pressing *Esc*. [In older versions of Windows there may be a limit, e.g. 32Kb or 64Kb, to the amount of data.]

Only numerical data should be pasted. The data for each subject or (for grouped data) each group must be in a separate line. The values in each line must be in a defined order, separated by spaces.
Vertical alignment is not essential. For example, the first two lines might look like this:

```
1  1  74  0  0  1.56  1  3  96  1  1
2  1  0  75  0  0  9.03  0  0  0  0  0
```

A value must be shown for each variable, with no blanks (a "missing value" code should be used). The names and sequence of the variables must be supplied separately, either by entry at the keyboard or by loading a dictionary file previously created by this program. After entry in the box, the data can be modified if necessary, and copied and pasted to a text file for future re-use; to mark all the data in the box for copying, press *Ctrl-A*; copy to the clipboard by pressing *Ctrl-Ins* or *Ctrl-C*.

The model

The logistic regression equation is :

$$\text{log-odds}(Y = 1) = \text{constant} + aA + bB + cC$$

where Y is a dichotomous dependent variable, and *A*, *B*, etc. are independent variables. In addition, it may include first-degree interactions (terms involving two variables, such as *A***B*.

In this program, the term "model" is taken to mean the list of independent variables and interactions, e.g. *A B C A**C, using the names of the variables, e.g.

```
AGE  SEX  WEIGHT AGE*WEIGHT.
```

Unconditional or conditional analysis

The program does both unconditional analyses (for unmatched data) and conditional analyses (which are appropriate for matched or finely-stratified data; Selvin 1996: 298-310), with easy switching between modes to permit comparison of the results.

Conditional multiple logistic regression requires a matching variable that identifies matched sets.

Conditional analyses may abort if the data are very extensive.

Categorical variables

Each independent variable can be treated either as simple (continuous) or as a categorical variable with up to 10 categories defined by the user (by specifying cutting-points); the program creates the required dummy variables. A category can contain a single value or more than one value.

Optionally, the categories can be treated as nominal or ordinal. *Nominal categories* can be handled in two ways in the analysis: the reference category can be either the first (baseline) category (this is the default), or the preceding category. Contrasts with the preceding category (Walter *et al.* 1987) permit close scrutiny of the effects of successive levels, e.g. in trials using different doses. Variables with *ordinal categories* are treated as simple variables, the successive categories (1, 2, 3 etc.) constituting its levels.

Missing values

Missing values should preferably be denoted by an unattainably high number, preferably 9, 99, 999 etc., but other numbers can be defined as codes for missing values.

Records that contain variables with missing values, or with categories that include missing values, are omitted from analyses.

Restriction of sample

Optionally, the analysis can be restricted to a specified subgroup of the sample, e.g women of a certain age.

Centering

The program can center simple independent variables by subtracting the mean from each observation. This reduces the effect of collinearity (highly correlated independent variables), and may be especially useful if both a variable (e.g. age) and its quadratic term (age-squared, or age*age) are included in the model (Selvin 1996: 256-259; Breslow and Day 1980: 233-236).

Logistic regression coefficients

The program provides logistic regression coefficients (with their standard errors) and the corresponding *odds ratios* (with 90, 95 or 99% confidence intervals). Results are displayed for simple variables and for all categories (except the first) of categorical variables.

Each coefficient reflects the influence of the relevant variable or interaction (i..e., its effect on the log-odds) when other influences (all the other covariates in the model) are held constant. For a simple variable, the coefficient and corresponding odds ratio express the effect of a change in magnitude of one unit. For a categorical variable, they express the contrast with the reference category (baseline or preceding) or (if the categories are ordinal) the effect of a rise from one level to the next.

Crude odds ratios (not holding other covariates constant) are displayed for comparison.

Z-scores (the coefficients divided by their standard errors) are displayed for use in comparing the impact of different variables (Selvin 1996: 262).

Statistical tests

The significance of each coefficient is tested by the *Wald test*, which uses 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: a *score test* (recommended for small samples; Breslow and Day 1980: 207- 208)) and a *likelihood-ratio test* (based on a comparison of *G*-statistics). For a factor that is treated as ordinal, these are tests of trend.

A likelihood-ratio test is also applied to the last variable added to the model, for use when a model is gradually incremented. For "automatic" models, likelihood-ratio tests are applied to each total model, and to each "main effect".

Log-likelihood

The log-likelihood is displayed for the total model and for the null model (for the constant only), and the *G* statistic (-2 times the log-likelihood) is displayed for the model.

Hosmer-Lemeshow goodness-of-fit test

Goodness of fit of the model is appraised by the Hosmer-Lemeshow test; a low P value suggests that the logistic model is unsuitable, i.e. that there is a poor fit with the observed data. A good fit does not necessarily mean that the results of the logistic analysis are valid, but a poor fit points to low validity. If the sample is very small, the test has a low power for detecting a poor fit; and if the sample is very large, a very small deviation from

the expected values may be highly significant. The test is not done if conditional analysis is used.

For this test, the subjects are arranged in a rising sequence of probabilities of the outcome (dependent variable = 1, or "yes") as predicted by the model, and split into equally-sized "deciles of risk", in which observed and expected status (for both "yes" and "no") are compared. The test is replicated, because if there are tied probabilities (e.g. if grouped data were entered), tied individuals may be split among adjacent deciles in different ways, and their arrangement in the observed data may affect the findings. The test is therefore done 100 times, after randomly shuffling the subjects 20 times before each test, and the median P value is displayed.

A warning is shown if there are deciles with expected numbers (of "yes" or of "no") that are less than 5; the program may then repeat the test (once) after combining deciles so as to avoid these low expected numbers.

Brier skill score

The *Brier skill score* (Steyerberg *et al*. 2010) measures the accuracy of predictions, such as those based on the logistic coefficients for a specific model. It is derived from the Brier score (Brier 1950), which is based on a comparison of the prediction for each subject with the observed occurrence of the disease (or other attribute).

The Brier skill score expresses the accuracy of these predictions as a percentage. In this program it is based on a comparison with accurate predictions based solely on knowledge of the frequency of the disease in the observed subjects. The score ranges from 100%, indicating perfect predictions, through 0%, which indicates that the test provides no improvement over predictions based solely on the frequency of the disease, to negative values, which indicate that the test provides worse predictions than those based on the frequency of the disease (Mason 2004). The prevalence of the disease in the target population (as opposed to its prevalence in the observed subjects) is not taken into account.

The score is not appropriate in case-control studies.

Indicators of discriminatory power

The program provides several indicators of the performance of the logistic model in discriminating between the two outcome groups, i.e. between subjects with a dependent variable of 0 or 1.

Tjur's coefficient of discrimination, proposed by Tjur (2009) as a standard measure of a model's explanatory power, is the arithmetic difference between the mean predicted probabilities of the outcome (based on each subject's combination of the independent variables in the model, weighted by the logistic regression coefficients) in the two groups of subjects.

Hedges's g (Hedges 1981) is a measure of *effect size,* standardized by dividing the difference between the mean predicted probabilities in the two outcome groups by an estimate of the pooled standard deviation. This index is almost identical to Cohen's *d,* which uses a slightly different estimate of the pooled standard deviation (Hedges and Olkin 1985). It is a biased estimator of the population effect size, but correction of the bias has no practical significance (Royston and Altman 2010). Assuming a normal distribution, approximate 95% confidence limits for *g* are computed from its standard error.

The *c-index* is a measure of concordance between the observed outcomes and the predictions. It ranges from 0.5 (performance no better than chance) to 1 (perfect discrimination). It expresses the probability, across all subjects, that the model will be correct in predicting that any one subject has a higher probability of the outcome than any other subject (Royston and Altman 2010). It is numerically equivalent to the area under the ROC curve. As a rough guide, a value area of 0.97 or more indicates excellent discriminatory power, a value of at least 0.92 very good discriminatory power, and a value of over 0.75 good discriminatory power (Simon 2004). The index and its approximate 95% confidence limits are derived from Hedges's *g* and its confidence limits. The *c*-index is not reported if computational difficulties are encountered.

The *gamma* index, suggested by Harrell *et al.* (1982) as being easier to understand than the *c-index*, is a rescaled *c-index* ranging from 0 ("no better than a coin-flip") to 1 (perfect discrimination). It expresses the probability that the prediction when comparing a pair of subjects will be correct, minus the probability that it will be incorrect. The *gamma*-index is not reported if computational difficulties are encountered.

The *overlap* (*OVL*) *coefficient* indicates the degree of overlap between the frequency distributions of the probabilities computed from the logistic regression coefficients in the two outcome groups. It can range from 0 to 1; the better the discrimination, the lower the overlap. On the assumptions that the distributions are normal and have a common variance (Rom and Hwang 1996, Royston and Altman 2010), the coefficient is derived from Hedges's *g,* and its approximate 95% confidence limits (reported if no computational difficulties are encountered) from those of Hedges's *g.*

Other indicators of the suitability of the logistic model

The program provides several other indicators of the suitability of the logistic model: the *Pearson correlation coefficient* between the observed value of the dependent variable (0 = "no", 1 = "yes") and the probability (of "yes") predicted by the logistic equation; the *square of the correlation coefficient,* which is an estimate of the proportion of explained variation (Mittlboeck and Schemper 1996); and Darlington's *logistic regression fit index* (Darlington 1990: 449) and the *pseudo R-squared* value (Selvin 1996: 266), both of which are based on a comparison of likelihood statistics based on the full model and on the null model, and are not direct measures of goodness of fit.

Probability of the outcome

Optionally, the program can use the regression coefficients to compute the probability of the outcome (dependent variable = 1), for given values of the variables in the model. Confidence intervals are displayed. This option is not appropriate in case-control studies.

Odds ratios comparing different sets of values

Optionally, the program can use the regression coefficients to compute an odds ratio expressing the contrast between two given sets of values of some or all of the variables in the model. Confidence intervals are displayed. When entering the values it should be kept in mind that a variable with the same value in both sets will not affect the result, unless it is involved in an interaction.

The program can also compute an odds ratio expressing the effect of a given difference between two values of a variable. Confidence intervals are displayed. This option is not appropriate for categorical variables.

Risk ratio, risk difference, and NNT

Since the odds ratios provided by a logistic regression analysis may give a misleading impression of the strength of associations, the program offers estimates of risk ratios and risk differences (prevalence ratios and differences, in a cross-sectional study) as well. These estimates express the association between a selected simple "yes-no" variable (e.g. treatment, in a trial, or exposure to a risk or protective factor) and the outcome, controlling for the variables included in the logistic model. This option is available only if all the covariates are simple "yes-no" variables or their interactions (i.e., not categorical variables). The option is not appropriate in case-control studies, where the constant coefficient is not validly estimated.

Adjusted risk ratios, together with adjusted risk differences and the *NNT* (the number needed to treat or to be exposed for one person to benefit or be harmed) are estimated by a method described by Austin (2010a), based on the predicted probabilities of the outcome in each subject in the sample (using the logistic regression coefficients). These probabilities are averaged, and the mean probability when a "yes" value is assumed for the independent variable under consideration is compared with the mean probability when a "no" value is assumed. In a trial, these results express the "average treatment effect" (*ATE*) based on a comparison between the estimated risks if all subjects were treated and if all subjects were not treated. The procedure may underestimate the risk ratio if the effect of exposure is large and the outcome has a high incidence (Knol *et al.* 2012).

Following the lines suggested by Bender et al. (2007), separate similar analyses are also performed on subjects who have "no" and "yes" values for the independent variable (i.e., unexposed and exposed, or untreated and treated). In the analysis of subjects with "no" values, the risk ratio, risk difference and NNT (the *NNE*, the number needed to be

exposed) express the effect of exposure in unexposed persons. In the analysis of subjects with "yes" values, the risk ratio, risk difference and NNT (the *EIN*, the exposure impact number) express the effect of removing exposure in exposed persons; this is the *ATT*, or the average treatment effect for the treated. As stressed by Austin (2010b) different measures of treatment effect are appropriate in different situations.

Risk ratios or prevalence ratios

Relative risks (risk ratios or prevalence ratios) can be reported instead of odds ratios , using a procedure described by Diaz-Quijano (2012). This is done by re-running the logistic regression analysis after modifying the data by adding duplicates of the cases (those with a positive outcome), but defining them as non-cases. [If this option is selected, the program makes the modification after the usual analysis, and places the modified data-base in a file from which it can be copied and pasted as a new entry.]

This procedure is analogous to the analysis of case-cohort studies, in which the cases are compared, not with non-cases, but with a sample of the entire cohort. In a computer simulation of data concerning dichotomous variables, the estimates of relative risks provided by this procedure were shown to be similar to those estimated by more accurate methods (binomial regression or Cox regression with robust variance). Their confidence intervals are somewhat wider than those provided by the other methods, especially if the outcome is frequent; and P values are higher. Significance tests for the difference of the odds ratio from 1 apply to the risk ratios also.

Propensity-score analysis

The aim of propensity-score analysis is to control for selection bias in an observational or nonrandomized study of the effect of a causal Yes-No factor (e.g., treatment or some exposure) on a Yes-No dependent variable (outcome), where there may be differences between the prior characteristics or experiences of the treated and untreated, or of the exposed and unexposed, that may affect the result of the study. There is generally no substantial difference between the results of propensity-score analyses and multivariable regression analyses (Shah et al. 2005, Sturmer et al. 2006). Propensity-score analysis may be especially useful in studies where the outcome is too rare to permit the inclusion of enough covariates in a logisticc regression model (Braitman abd Rosenbaum 2002).

The variables to be entered are the causal variable and the variables that may affect it - i.e., those that may differ in the treated and untreated or in the exposed and unexposed groups and (optionally) variables that may affect the dependent variable. Interactions may be included. The explanatory variables must be simple ones (i.e., without predefined categories), and the program must be set for unconditional logistic regression. Missing values must be defined before entry of the model (the relevant records are excluded from the analysis).

Experts vary in their advice on the selection of explanatory variables for inclusion in the propensity score model. Some advocate using only those variables that predict exposure (treatment assignment, or allocation to the exposed or unexposed group), others suggest including all variables potentially related to the outcome, and others advocate including only variables that are associated with both exposure and outcome. According to Austin et al.(2007), inclusion of variables that are not related to the outcome is not helpful. Newgard et al .(2004) point out that the exclusion of even weak confounders nay produce bias. D'Agostino (2007) recommends that only variates that occur prior to exposure should be included. According to Williamson et al. (2011), the model should include all confounders, and the inclusion of variables related only to the exposure or only to the outcome does not create bias.

The program first reports the results of a logistic regression analysis showing the effects on the dependent variable of the causal variable and all the explanatory variables entered in the mode. It then calculates propensity scores by performing a logistic regression analysis that examines the effect of the explanatory variables on the causal variable. This yields an odds log for each subject, from which the probability of treatment or exposure is then calculated. These probabilities are the propensity scores (which are not reported).

Two propensity-score analysis methods are then used. First, the inverse probability of treatment weighting (IPTW) procedure, which uses weights based on the propensity scores, and yields counterfactual estimates of what the rate of the dependent variable would be if all subjects were exposed to the causal factor, and what it would be if no subjects were exposed to the causal factor. (Controlling, of course, for the propensity score). The absolute difference between these two estimates is the "average causal effect" (ACE; sometimes called the "average treatment effect", ATE); it is the average difference between the two possible responses for each participant. The ratio of these two rates can be regarded as a risk ratio, and an odds ratio can be derived from the rates. A similar method is used to calculate the average causal effect among those exposed to the causal factor (which usually means among those exposed to treatment).

Secondly, a logistic-regression analysis is conducted, regressing the outcome variable on an indicator variable denoting exposure status and the estimated propensity score, and yielding an odds ratio expressing the effect of the exposure on the outcome, adjusted for the effect of the propensity score.

By using the “NEXT” button, other models (using the same data and the same causal and dependent variables) can be tested. When the program is closed, a ‘runtim error”message may then appear; this does not affect the results, and can apparently be ignored.

METHODS

Basic statistical techniques for multiple logistic regression analysis are described by Breslow and Day (1980: 182-227). Variance formulae for use in the estimation of confidence intervals are provided by Hosmer and Lemeshow (1989: 103-106).

The basic computation is based on MULTLR (Campos-Filho and Franco 1989), which uses adapted algorithms from LOGRESS (McGee 1986) and PECAN (Lubin 1981), respectively, for unconditional and conditional maximum likelihood estimation of the logistic coefficients. We are grateful to Dan McGee for providing us with the Fortran code for LOGRESS and to Dr A. Negassa for his helpful comments.

The maximum number of parameters (including dummy variables) in a model is 33.

Pseudo R-squared is defined as

$$(LLN - LLM) / LLN$$

and Darlington's *logistic regression fit index 1* as

$$\{\exp[(LLN - LLM) / N] - 1\} / [\exp(-LLN / N) - 1]$$

where LLM = the log-likelihood for the model
LLN = the log-likelihood for the null model
N = sample size

The formula for the goodness-of-fit test (Hosmer and Lemeshow 1989: 140-145; Selvin 1996: 264- 266) is

$$\text{chi-sq} = \Sigma [(O - E)^2 / E]$$

where O = observed number
E = expected number, counted separately for "Yes" and "No" categories in each decile (20 cells altogether).
The test is done 100 times, each time shuffling the observations 20 times, using the *ran0* function of Press *et al.* (1989), and the median P value is displayed.

The alternative probabilities of the outcome in each individual subject (the probability if the subject were treated or exposed, and the probability if the subject were not treated or exposed) are computed by formulae 2 and 3 of Austin (2010a). The adjusted *risk ratio* is the ratio of the mean probabilities, and the adjusted *risk difference* is their difference. The *NNT* is the reciprocal of the mean difference.

Brier skill score

The general formula for the Brier score (from which the Brier skill score is derived) is

$$\Sigma (P_i - D_i)^2 / N$$

where N = number of observations
P_i = prediction for subject i (probability, between 0 and 1)
D_i = observed status of subject i (0 or 1)

The formula for the Brier skill score (Steyerberg *et al.* 2010) is

$$100(1 - B) / F$$

where B = Brier score
$F = P * (1 - P)$
P = proportion with the disease
If the calculated value of the Brier skill score is less tha 0%, it is reported as 0%; and if it exceeds100%, it is reported as 100.

Confidence intervals are estimated by substituting the confidence limits of the Brier score for *B*.

Indicators of discriminatory power

Tjur's coefficient of discrimination is the arithmetic difference between the mean predicted probabilities of the outcome in the two outcome groups (Tjur 2009).

Hedges's g is the arithmetic difference between the mean predicted probabilities of the outcome in the two outcome groups, divided by the pooled standard deviation, computed by the formula provided by Royston

and Altman (2010). Its approximate confidence intervals are based on the standard error formula derived by Royston and Altman from their formula A1.

The *c-index* is computed by Royston and Altman's formula A4, the *gamma* index is $2(c\text{-}index - 0.5)$ (Harrell *et al*. 1982), and the *overlap coefficient* is computed by the formula of Rom and Hwang (1996: 1493), cited by Royston and Altman (2010: appendix A, section A.2). The *c-index* and *gamma* index are not provided if computing difficulties are encountered.

Two test files may accompany this program. LA.DAT is an ASCII data file containing 315 individual records published by Breslow and Day (1980, Appendix III), from a study of endometrial cancer in Los Angeles (Mack *et al*. (1976), and LA.DIC is the corresponding dictionary file. They must be placed in an LR subfolder in the folder in which LOGISTIC.EXE is situated. Missing values are coded 9 or 99.

Risk ratios or prevalence ratios

The method is described by Diaz-Quijano (2012). This is done by re-running the logistic regression analysis after modifying the data by adding duplicates of the cases (those with a positive outcome), but defining them as non-cases.

Logistic coefficients are then computed in the usual way, but the "odds ratios" that are obtained are considered direct estimates of risk or prevalence ratios.

Propensity-score analysis

The inverse probability of treatment weighting (IPTW) procedure uses formula 4 of Hirano and Imbens (2001) [also formula 8 of Lunceford and Davidian 2004) to calculate the average causal effect in the total sample. The average causal effect among those exposed to the causal factor is calculated by formula 20 of Schafer and Kang (2008). The risk ratio is *Z1 / Z2*, and the odds ratio is *[Z1* / (1 - *Z1*]) / *[(Z2* / (1 - *Z2*)], where *Z*1 and *Z2* are the hypothetical rates of the dependent variable if all subjects were (respectively) exposed or not exposed to the causal factor.

In the final logistic regression analysis, the outcome variable is regressed on the causal variable and the estimated propensity score.

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.

Austin PC (2010a) Absolute risk reductions, relative risks, relative risk reductions, and numbers needed to treat can be obtained from a logistic regression model. Journal of Clinical Epidemiology 63: 2-6.

Austin PC (2010b) Different measures of treatment effect for different research questions. Journal of Clinical Epidemiology 63 9-10.

Austin PC, Grootendorst P, Anderson GM (2007) A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: a Monte Carlo study. Statistics in Medicine 26 :734–753.

Bender R, Kuss O, Hildebrandt M, Gehrmann U (2007) Estimating adjusted NNT measures in logistic regression analysis. Statistics in Medicine 26: 5586-5595.

Braitman LE, Rosenbaum PR (2002) Rare outcomes, common treatments: analytic strategies using propensity scores. Annals of Internal Medicine 137: 693-696.

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.

Brier GW (1950) Verification of forecasts expressed in terms of probability. Monthly Weather Review 78: 1–3.

Campos-Filho N, Franco EL (1989) A microcomputer program for multiple logistic regression by unconditional and conditional maximum likelihood methods. American Journal of Epidemiology 129:439-444.

Darlington RB (1990) Regression and Linear Models. New York, McGraw-Hill.

D'Agostino RB Jr (2007) Propensity scores in cardiovascular research. Circulation 115: 2340-2343.

Dean AG, Dean JA, Coulombier D, Brendel KA, SmithDC, Burton AH, Dicker RC, Sullivan K, Fagan RF, Arner, TG. Hathcock,L. Epi Info, Version 6: a word processing, database, and statistics program for public health on IBM-compatible microcomputers. Centers for Disease Control and Prevention, Atlanta, Georgia, U.S.A., 1996. Available on the Internet at http://www.cdc.gov/epiinfo/Epi6/ei6.htm

Diaz-Quijano FA (2012) A simple method for estimating relative risk using logistic regression. BMC Medical Research Methodology 12:14.

Harrell FE Jr, Califf RM, Pryor DB, Leo KL, Rosati RA (1982) Evaluating the yield of medical tests. JAMA 247: 2543-2546.

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

Hedges LV (1981) Distribution theory for Glass's estimator of effect size and related estimators. Journal of Educational Statistics 6: 107-128.

Hedges LV, Olkin I (1985). Statistical Methods for Meta-Analysis. Orlando: Academic Press.

Hirano K, Imbens GW (2001) Estimation of causal effects and propensity score weighting: An application to data on right heart catheterization .Health Services and Outcomes Research Methodology 2001 2: 259-278.

Hosmer DW Jr, Lemeshow S (1989) Applied Logistic Regression. New York: John Wiley & Sons.

Knol MJ, Le Cessie S, Algra A, Vandenbroucke JP. Groenwold RHH (2012) Overestimation of risk ratios by odds ratios in trials and cohort studies: alternatives to logistic regression. Canadian Medical Association Journal. Available on the Internet at CMAJ 2012,DOI:10.1503/cmaj.101715

Lauritsen JM, Bruus M (2003-2004) EpiData (version 3). A comprehensive tool for validated entry and documentation of data. The EpiData Association, Odense, Denmark. (Available on the Internet at http://www.epidata.dk/

Lubin JH (1981) A computer program for the analysis of matched case-control studies. Computers and Biomedical Research 14: 138-143.

Lunceford JK, Davidian M (2004) Stratification and weighting via the propensity score estimation of causal treatment effects: a comparative study. (Statistics in Medicine 204, 23: 2937-2960

Mack TM, Pike MC, Henderson, BE, Pfeffer RI, Gerkins VR, Arthur BS, Brown SE (1976) Estrogens and endometrial cancer in a retirement community. New England Journal of Medicine 294: 1262-1267.

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

Mittlboeck M, Schemper M (1996) Explained variation for logistic regression. Statistics in Medicine 15: 1987.

Newgard CG, Hedges JR, Arthur M, Mullins RJ (2004) Advanced statistics: The propensity score – a method for estimating treatment effect in observational studies. Academic Emergency Medicine 11: 953-961.

Press WH, Flannery BP, Teukolsky SA, Vetterling WT (1989) Numerical recipes in Pascal: The art of scientific computing. Cambridge: Cambridge University Press.

Rom DR, Hwang E (1996) Testing for individual and population equivalence based on the proportion of similar responses. Statistics in Medicine 15: 1489-1505.

Royston P, Altman DG (2010) Visualizing and assessing discrimination in the logistic regression model. Statistics in Medicine. Published online 13 May 2010.

Schafer JL. Kang J (2008). Average causal effects from nonrandomized studies: a practical guide and simulated example. Psychological Methods 13: 279-313.

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

Shah BR, Laupacis A, Hux JE, Austin PC (2005) Propensity score methods gave similar results to traditional regression modelling in observational studies: a systematic review. Journal of Clinical Epidemiology 58: 550-559.

Simon S (2004) Ask Professor Mean: ROC curve http://www.cmh.edu/stats/ask/roc.asp

Steyerberg EW, Vickers AJ, Cook MR et al.(2010) assessing the performance of prediction models. Epidemiology 21: 128-138.

Tjur T (2009) Coefficients of determination in logistic regression models – a new proposal: the coefficient of discrimination. The Americal Statistician 63: 366-372.

Sturmer T, Joshi M, Glynn RJ, Avorn J, Rothman KJ, Schneeweiss S (2006) A review of the application of propensity score methods yielded increasing use, advantages in specific settings, but not substantially different estimates compared with conventional multivariable methods. Journal of Clinical Epidemiology 59: 437 – 447

Walter SD, Feinstein AR, Wells CK (1987) Coding ordinal independent variables in multiple regression analyses. American Journal of Epidemiology 125: 319-23.

Williamson E, Morley R, Lucas A, Carpenter J (2012) Propensity scores: From naive enthusiasm to intuitive understanding. Statistical Methods in Medical Research 21: 273-293.