riskplot: A graphical aid to investigate the effect of multiple categorical risk factors

Milena Falcaro
Biostatistics Group
University of Manchester, UK
milena.falcaro@manchester.ac.uk

Andrew Pickles
Biostatistics Group
University of Manchester, UK
andrew.pickles@manchester.ac.uk

Abstract. In this article, we describe a command, riskplot, aiming to provide a visual aid to assess the strength, importance, and consistency of risk factor effects. The plotted form is a dendrogram that branches out as it moves from left to right. It displays the mean of some score or the absolute risk of some outcome for a sample that is progressively disaggregated by a sequence of categorical risk factors. Examples of the application of the new command are drawn from the analysis of depression and fluid intelligence in a sample of elderly men and women.

Keywords: gr0044, riskplot, pathways, risk factors, graphs

1 Introduction

Getting a rounded picture of the contributions of individual risk factors and their combination and interaction on the risk of some disease is not easy. While we may have moved on from a narrow focus on significance levels to present confidence intervals for effect estimates obtained under particular models, we rarely progress to isolate the breadth of the sample evidence base for each effect or fully describe the likely importance of the effect on population risk.

Pickles and colleagues (Quinton et al. 1993; Hill et al. 2001) have used manually constructed diagrams to illustrate the impact of sequential risks and risk pathways to study outcome. They considered the effect of a “turning point”, the gaining of a supportive partner, on reducing the risk of antisocial behavior (Quinton et al. 1993), and also how such supportive love relationships may be able to moderate the effects of childhood sexual abuse on the risk of depression in adulthood (Hill et al. 2001). The latter extended the diagram in several ways by 1) plotting estimates of risk and risk factor prevalences that were obtained after adjustment for the sample design by the use of weights and 2) attaching to each risk factor combination a label that can contain additional quantitative information, such as the outcome risk under some model or the population attributable fraction for that pathway. In this article, we describe a new command, riskplot, that generates plots such as these in a straightforward way.

The riskplot command is not statistically sophisticated but is a visual aid intended to give an audience access to the data to assess the strength, importance, and consistency of risk factor effects. It displays the mean of some score or the risk of some outcome, commonly the proportion affected, for a sample that is progressively disaggregated by a
sequence of categorical risk factors. If the risk factors are considered in a chronological order of exposure, then it becomes natural to think of the progressive disaggregation as being a risk factor pathway and the plot can be used to highlight particular pathways that are especially common or lead to unusually high or low outcome risks. As such, they would seem a natural visual aid in the fields of development and lifecourse epidemiology, highlighting the impact of accumulating risk factors.

2 The riskplot command

2.1 Description

The riskplot command draws a plot that shows the effect of a set of categorical risk factors on some outcome of interest, $Y$. The plotted form is a dendrogram that branches out progressively as it moves from left to right, and in which the $y$ axis is the expected value of the outcome and the $x$ axis is a sequence of categorical risk factors, with the vertical width of the branches representing the relative frequency of the risk factor combination.

On the right-hand side next to each branch end, the user can decide to display additional information such as the observed and expected frequencies and the “pathway labels” (composite strings obtained by concatenating the numeric values of the risk factor levels).

For binary risks, the plot is easily interpretable for up to three or four risk factors, though the setting of a threshold frequency such that the rare risk combinations are trimmed from the sample and omitted from the plot often helps. The riskplot command also offers scope for the use of color to distinguish, for example, high from low risk pathways. This also helps highlight circumstances where interaction and moderation take place, such that the same level of one risk factor may either increase or decrease risk depending upon the level of a prior risk factor. The use of weights is also allowed and permits, for example, adjustment via probability weights for stratified sampling and attrition. However, when the computation of expected frequencies is required, only frequency weights may be specified.

2.2 Syntax

\[ \text{riskplot depvar [xvars] [if] [in] [weight] [, all path observed expected c(colorstyle) thick(#) trim(#) extended(#) twoway_options]} \]

\text{depvar} represents the outcome variable. \text{xvars} is a list of categorical risk factors whose values must be integers between 0 and 9. No more than 20 risk-factor–level combinations may be displayed in the same plot. Observations that have missing values for either \text{depvar} or any of the \text{xvars} are disregarded. Weights are allowed, but when the \text{all} or \text{expected} option is specified, only \text{fweight} may be used. The user is responsible for ensuring the validity of what is produced when \text{aweights} or \text{iweights} are used.
2.3 Options

all displays pathway labels as well as observed and expected frequencies and percentages. This is equivalent to simultaneously specifying the path, observed, and expected options.

path displays the pathway labels at the end of each branch.

observed displays observed frequencies and percentages.

expected displays expected frequencies and percentages (i.e., frequencies and percentages that we would observe if the risk factors were independent).

c(colorstyle) specifies the colors of the branches. colorstyle is a list of strings defining the colors to be used for each level of the risk factors. The first string refers to 0s, the second to 1s, and so on. In general, the ith string \((i \geq 1)\) in \(c()\) represents the color for value \(i - 1\) of the risk factors. The user may also specify a dot when the default color (i.e., black) is required. For example,

\[
c(\text{red green blue blue}) \text{ implies } 0 = \text{red}, 1 = \text{green}, 2 = \text{blue}, 3 = \text{blue},
\text{ others} = \text{black}
\]

\[
c(\text{red . blue}) \text{ implies } 0 = \text{red}, 1 = \text{black}, 2 = \text{blue}, \text{ others} = \text{black}
\]

\[c(\text{red}) \text{ implies } 0 = \text{red}, \text{ others} = \text{black}\]

thick(\#) increases the line thickness of the branches. The number specified must be between 0 and 25, where 20 is twice as thick as 10. The default is thick(10).

trim(\#) prevents the display of branches with relative frequency smaller than the percentage specified. \# must be a number between 0 and 100.

xextended(\#) provides additional space for labels on the right-hand side of the graph.

twoway_options allow control of graph titles, legends, additional lines, text, etc.

2.4 Refining and repositioning graph labels

When using riskplot, some plotted pathways may end up being very close to each other, and the corresponding information displayed on the right side of the graph (e.g., path labels and frequencies) may easily overlap. The Graph Editor allows the user to change the appearance of the graph and, among other things, to add and move text. We recommend using this new graphical device to solve overlapping text problems. To use the Graph Editor, right-click on the graph and select “Start Graph Editor”. For more details, see the Stata 11 Graphics Reference Manual (StataCorp 2009) and A Visual Guide to Stata Graphics (Mitchell 2004).
3 Examples

Here we illustrate the application of `riskplot` using data from the Steel-Dyne Cognitive Ageing cohort of elderly men and women. The study was established in 1982, and it involves the follow-up of over 6,000 normal healthy individuals aged 50 years and over (for more details, see Rabbitt et al. [2004]). In this article, we focus on the subsample of subjects who were assessed for depression in 1991 and in 1995. We graphically explore the effect of a set of risk factors on cognitive decline and depression as measured in 1995 by the AH4 intelligence test (Heim 1970) and the Yesavage geriatric depression scale test (Yesavage et al. 1982), respectively. More specifically, the AH4 test consists of two 65-item parts and yields a total score that is used as a scale for grading fluid intelligence, which is the ability to reason abstractly and to solve new problems. The Yesavage geriatric depression scale test is a screening instrument of 30 items with a yes/no format and a scale ranging between 0 (no depression) and 30 (severe depression).

The risk factors we focus on are sex (0 = male, 1 = female), social class (sclass: 0 = high, 1 = low), and depression (depr1991: 0 = no depression, 1 = mild depression, 2 = severe depression), as measured in 1991.

We start by considering a simple example (figure 1) illustrating the association of social class and previous depression status with AH4 (fluid intelligence) in 1995.

```
. riskplot AH4 sclass depr1991, path ytitle(Fluid intelligence 1995)
```

![Figure 1. Simple risk plot for assessing the impact of social class and depression in 1991 on fluid intelligence in 1995. Pathway labels are displayed by specifying the `path` option.](image-url)
The labels distinguish the risk combinations; for example, “12” indicates low social class (1) and severely depressed (2). Substantial differences in 1995 AH4 mean score are shown for both social class and depression, the consistent effect of depression being very apparent within each social class. We can also introduce *sex* as a risk factor, but here it is better to omit the branches with low relative frequency, say, less than 5%. This risk plot is displayed in figure 2 and is generated by the following command:

```
.riskplot AH4 sex sclass depr1991, path obs trim(5) ytitle(Fluid intelligence 1995)
```

![Risk plot](image)

Figure 2. Risk plot with pathway labels and observed frequencies displayed using the `path` and `observed` options. Branches with relative frequency less than 5% are omitted by specifying `trim(5)`.

The plot shows that men with high social class and no depression at baseline tend to have better fluid intelligence performances than the other subjects in the sample.

Let’s now consider the depression score in 1995 (*depr1995*) as the outcome of interest and explore the effect of *sex* and social class for subjects who were found to be depressed in 1991. Let *dep91* denote a binary variable that takes on the value 1 when depr1991 is equal to 1 or 2, corresponding to subjects with mild (*depr1991 = 1*) or severe (*depr1991 = 2*) depression at baseline.

A clearer picture of the effect of risk factors on the outcome of interest can be obtained by specifying, for example, options for path colors and thickness. Observed frequencies and those expected under independence can also be added (figure 3). The `expected` option also generates some tabulated output in the standard results window. As the length of the labels increases, it may be necessary to move the right-hand margin of the plot with the `xextended(#)` option.
As very often happens in longitudinal studies, the Steel-Dyne Cognitive Ageing cohort is affected by the presence of selective dropout. Sampling weights can be estimated by modeling the probability of dropout via a logistic or probit regression and incorporated into the model to adjust for attrition.
4 Conclusion

The \texttt{riskplot} command described in this article is a graphical aid to the investigation of the contributions of risk factors on outcomes of interest. In form, this plot has much in common with dendrograms derived from cluster analysis. For each of the risk factor combinations, it is possible to highlight differences in prevalence and to display additional information such as labels, and observed and expected frequencies and percentages. It might also prove useful in illustrating independence between two variables conditional upon a third and for examining possible mediation of effects.

5 Acknowledgments

This work was supported by the Economic and Social Research Council research grant RES-000-22-2921. We would like to thank Neil Pendleton and Mike Horan (Department of Geriatric Medicine, University of Manchester, UK) for allowing us access to the cognitive aging cohort data used in this article.
6 References


About the authors

Milena Falcaro is a research fellow in biostatistics at the University of Manchester. Her research interests include structural equations modeling, longitudinal data analysis, and genetic epidemiology. The context of much of her work has been longitudinal behavioral and developmental psychopathological studies.

Andrew Pickles is a professor of epidemiological and social statistics at the University of Manchester. He shares similar research interests to Milena, has long used Stata in teaching and research, and together with Sophia Rabe-Hesketh and Anders Skrondal formulated the generalized linear latent and mixed model (\texttt{gllamm}) in Stata.