MJA
The Medical Journal of Australia

Appendix

This appendix was part of the submitted manuscript and has been peer reviewed.
It is posted as supplied by the authors.

Appendix to: Taylor A. Distributions and what to do when they are non-normal. *Med J Aust* 2018; 208: 10-13. doi: 10.5694/mja16.01255.

Appendix

The dataset

The data consist of 506 cases drawn from a larger sample of people who responded to an online survey conducted by Babucarr Sowe (2016, personal communication) and are used with his permission. Only a small subset of variables is used here and, for the purposes of demonstration, the variables are not necessarily used in the way that Sowe used them. In addition, the values of the outcome were altered slightly for the purposes of demonstration.

The dependent or outcome variable in our examples is an index of harmful use of alcohol with a mean of .25 and a standard deviation of .45, with higher scores indicating a greater incidence of harmful behaviours. The distribution of the measure is positively skewed (index value of 3.1) and leptokurtic (index value of 12.1).

There are five independent variables: *age*, in years, *sex* (0=male, 1=female), and three categorical variables, *anx*iety (0, 1, 2), *drugs* (0, 1, 2, 3) and *relig*ion (0=non-religious, 1=former Christian, 2=always Christian). The higher-numbered categories of *anx* and *drugs* indicate greater anxiety and greater drug use respectively; the exact meaning of the categories is not material for our purposes.

Glossary

gamma regression – the gamma distribution can be used to model outcomes which have values of zero or greater. One of its properties is that the variance of its values increases with increases in the mean. Hardin and Hilbe note that when the outcome is positively skewed, gamma regression provides an option to carrying out an OLS regression on a log transformation of the outcome.

generalised linear models – models based on linear equations but which can allow for non-normally distributed outcomes. A linear relationship between the outcome and the regression equation is ensured by appropriate transformations.

incident rate ratio (IRR) – a ratio which compares the rates of occurrence of an outcome over different conditions. It is commonly used when interpreting analyses of count data in Poisson and negative binomial regression analyses, but is also applicable when these methods are applied to ordinal categorical data.

independent variable – a variable which may have an effect on, or an association with, an outcome or dependent variable. The term is most applicable to experiments, where the researcher manipulates the variable by means of random assignment (to experimental and control groups, for example), but is also applied in observational studies, where the conditions (e.g., exposed or not exposed) are not directly under the control of the researcher. The researcher may exercise some control by selection; e.g., by including people from a number of age groups.

In complex investigations, the distinction between dependent and independent variables may be blurred when a variable fills both roles.

indicator variables – also called dummy variables, indicator variables are 0/1 binary variables which are used to represent categorical variables in regression analyses. The number of 0/1 variables needed to represent a categorical variable is one fewer than the number of categories. For example, the original *anx* variable in the *alc_harm* example is coded 0, 1 and 2. For the purposes of the analyses described here, two dummy variables were needed to represent this variable. People who had a zero value on *anx* were coded zero on both dummies, those who had a value of 1 were coded 1 on the first dummy and zero on the second, and those who had a value of 2 were coded zero on the first dummy and 1 on the second.

With the above coding, category zero on *anx* is the reference category, which means that the regression coefficient for the first dummy variable shows the difference between the mean for those in category 0 and those in category 1 (mean for category 1 – mean for category 0) and the second dummy variable shows the difference between the mean for those in category 0 and those in category 2 (mean for category 2 – mean for category 0).

linear model – an equation which expresses an outcome as the sum of one or more independent variables, each weighted by a regression coefficient. In the example below, outcome Y is a function of X and Z, which are weighted by regression coefficients B_1 and B_2. B_0 is known as the intercept, which shows the value of Y when X and Z are zero (i.e., when the fitted line crosses the Y axis).

$$Y = B_0 + B_1X + B_2Z$$

ordinary least-squares regression – a linear model in which one or more independent variables are used to predict an outcome. The criterion for the best fit of the model to the data is that the sum of the squares of the differences between the actual and predicted values (the residuals) is as small as possible.

odds ratio (OR) – a ratio which compares the odds of an outcome occurring under one condition with the odds of it occurring under another condition. For example, if 10 of 50 females in a sample use alcohol in harmful way, the odds of females using alcohol harmfully are 10:40, which is equal to 10/40 or .25. If the equivalent figures for males in the sample are 20 out of 60, the odds for males are 20:40, which equals 20/40 or .5. To compare males with females, we compute the ratio of the odds, .5/.25 = 2. This shows that, in the sample, the odds of males using alcohol harmfully are twice those of females.

Odds ratios are often used to interpret the results of logistic regression analyses.

negative binomial regression – the Poisson distribution is used to model count data. One of its properties is that the variance is equal to the mean, which means that it often doesn't apply to data collected in studies, which is said to be 'over-dispersed'. Negative binomial regression adds a term which accounts for 'unobserved heterogeneity' and allows the method to be applied to distributions where the variance is greater than the mean.

regression coefficient – the weight applied to an independent variable in a regression equation. It shows the change in the outcome variable for a one-unit increase in the independent variable, with the other independent variables, if any, held constant. In generalised linear models, the coefficients may be translated into other quantities, such as odds ratios (ORs) and incident rate ratios (IRRs) to facilitate interpretation.

Regression coefficients are sometimes standardised to facilitate comparisons between the effects of independent variables which have been measured on different scales.

residuals - the difference between the value of an outcome measure and the value predicted by a model, regression models in our case. The difference is sometimes expressed as (observed – expected). The residuals give an indication of how well the model fits (or summarises) the data; their distribution indicates whether the data conform to the theoretical distribution assumed by the model.

standard error – the standard error of a regression coefficient shows how accurate it is as an estimate of the corresponding coefficient in the population. The larger the standard error, the less accurate the value based on a given sample is likely to be. With a given population, larger samples provide smaller standard errors.