INJURY TECHNICAL PAPERS

Australian Government
Australian Institute of
Health and Welfare

A guide to statistical methods for injury surveillance

Jesia Berry

James Harrison

FLINDERS UNIVERSITY
ADELAIDE AUSTRALIA

ISBN 1 74024 477 X

A guide to statistical methods for injury surveillance

Jesia Berry
James Harrison

The Australian Institute of Health and Welfare is Australia's national health and welfare statistics and information agency. The Institute's mission is *better health and wellbeing for Australians through better health and welfare statistics and information*.

Please note that as with all statistical reports there is the potential for minor revisions of data in *A guide to statistical methods for injury surveillance* over its life. Please refer to the online version at <www.nisu.flinders.edu.au>.

Injury Technical Paper Series
Number 5

A guide to statistical methods for injury surveillance

Jesia G. Berry

James E. Harrison

Research Centre for Injury Studies

Flinders University, SA

2005

Australian Institute of Health and Welfare
Canberra

AIHW cat. no. INJCAT 72

This publication is part of the Australian Institute of Health and Welfare's Injury Technical Paper Series. A complete list of the Institute's publications is available from Business Promotion and Media, Australian Institute of Health and Welfare, GPO Box 570, Canberra ACT 2601, or via the Institute's web site (www.aihw.gov.au). Electronic copies of publications in this series can be downloaded from the Research Centre for Injury Studies web site (http://www.nisu.flinders.edu.au).

ISSN 1446-778X
ISBN 1 74024 477 X

Suggested citation

Berry JG, Harrison JE. A guide to statistical methods for injury surveillance. Injury Technical Paper Series No 5. (AIHW cat. no. INJCAT 72). Adelaide: AIHW.

♦ Reissued with corrections: 4 October 2006

Any enquiries about or comments on this publication should be directed to:

Jesia Berry
Research Centre for Injury Studies
Flinders University of South Australia
GPO Box 2100,
Adelaide 5001, South Australia

Phone: (08) 8201 7602
email: jesia.berry@flinders.edu.au

Published by Australian Institute of Health and Welfare

Proofreading and Layout by Stacey Avefua

Contents

Preface

The AIHW National Injury Surveillance Unit (NISU) undertakes analysis and reporting on mortality and morbidity data for injury in Australia. The process of analysis presents many technical issues which require careful consideration. For example, what is the best way to make comparisons between population groups and measure trends in mortality and morbidity over time? How should we deal with random variation when case numbers are small? What is the best way to express uncertainty for an estimate? These are methodological questions for descriptive epidemiology. Most textbooks of epidemiology do not focus on the descriptive aspect of the discipline. We have found that the literature addressing these questions is dispersed in many different sources. In addition, the specific ways in which this information can be applied in injury surveillance is not always obvious.

This report was compiled in response to the need to provide a readily accessible guide to NISU staff on appropriate statistical methods for reporting injury. The aim is to provide a systematic guide on how to report injury in a methodologically robust manner. The report draws on a wide range of data sources and draws them together with practical examples pertinent to the injury field. It is written assuming the reader has only a basic knowledge of statistics and therefore complicated technical theory and use of mathematical notation has been kept to a minimum. The reader is recommended to consult the references for a more comprehensive coverage of statistical theory or for further clarification on any of the methods described.

This is the first methods handbook that we have developed. It does not cover all of the issues that might have been included, and there is potential to expand on the treatment of those that are covered. This document might be developed further or separate documents might be written to cover other topics. These issues that arise when we undertake descriptive epidemiology of injury at NISU are likely to arise when similar work is done elsewhere, on injury or other topics. This is why we are publishing what began as a guide envisaged as being only for internal use. We would welcome feedback on it.

James Harrison

Director, Research Centre for Injury Studies

Acknowledgment

Thankyou to Dr John Condon from the Menzies School of Health Research and Ken Tallis from the Australian Institute of Health and Welfare for their valuable contributions to the content of this report. Thanks also to Stacey Avefua for her assistance with editing and layout.

Abbreviations

ABS	Australian Bureau of Statistics
AIHW	Australian Institute of Health and Welfare
ASGC	Australian Standard Geographical Classification
BAC	Blood Alcohol Concentration
CI	Confidence Interval
CMF	Comparative Mortality Figure
DSR	Directly Standardised Rate
HR	Hazard Ratio
IRR	Incidence Rate Ratio
ISR	Indirectly Standardised Rate
NBD	Negative Binomial Distribution
NISU	National Injury Surveillance Unit
OLS	Ordinary Least Squares
OR	Odds Ratio
RCIS	Research Centre for Injury Studies
RR	Relative Risk
SE	Standard error
SIR	Standardised Incidence Ratio
SMR	Standardised Mortality Ratio
Var	Variance
ZIP	Zero Inflated Poisson

List of tables

List of figures

1 Confidence intervals

National morbidity and mortality statistics are usually based on counts of all cases of a condition or cause of death occurring in a specified population during a specified time. Such values are subject to errors in the registration process e.g. missing or incomplete data or poor data collection, but are not subject to the sampling error that arises when calculating estimates based on a random sample of cases.[1]

Rates and percentages based on a count of all cases (i.e. a census of cases) are subject to random variation.[1] A rate observed in a particular year can be considered an estimate of the true or underlying rate—the rate observed is one of a large series of possible results that could have arisen under the same circumstances. The time periods used to group cases (e.g. calendar year) are arbitrary, and use of another period (e.g. financial year) would result in different rates.

To interpret a point estimate, we need some indication of its precision. This is provided by the confidence interval (CI) which states the probability of containing the true value. Most are calculated as 95% CIs—there is a 95% chance that the confidence interval covers the true value; less frequently 90% or 99% limits may be used.

When viewing data in a graph, it may be tempting to draw conclusions about the statistical significance of differences between group means by looking at whether their confidence intervals overlap. For example, if the CIs around the age-adjusted rates for two States overlap, it may seem reasonable to conclude the rates are not significantly different, and vice versa. Although this may be a good approximation to a statistical test, it is not equivalent to one.[1] A confidence interval is constructed by taking into account the population size and variance in the group (e.g. State) for which it was constructed.[1] A proper statistical test will take into account the larger pooled size of the two groups when evaluating the difference between their means, and therefore can provide different results.[1] This means that in some cases an appropriate statistical test will indicate a statistically significant difference exists even though the CIs of two groups overlap.[1] However, if two CIs do not overlap, an appropriate statistical test will always indicate there is a significant difference between them.[1] While providing a convenient indication of precision, CIs are not appropriate to be used for testing for differences between groups or trends over time, which requires specific statistical tests (see Section 3).

Random variation can be substantial when the measure, such as a rate or percentage, has a small number of events in the numerator (e.g. less than 20), especially when the denominator (population) is also relatively small. As the population size increases, the standard error becomes smaller and hence the Cl becomes narrower. When interpreting CIs, it is important to include a commentary about the probability that differences or trends observed may be due to random variation due to small numbers (and resulting wider confidence intervals). Methods for calculating CIs for crude rates, age-specific rates, and age-adjusted rates for small and large population sizes are outlined in Section 2.8–2.10.

Many reports from agencies such as the Australian Institute of Health and Welfare (AIHW) display graphs which depict age-standardised mortality rates over time. However, few of these graphs ever use CIs around the mean. This pattern is also apparent in many peer-reviewed journal articles.

Many graphs follow this format:

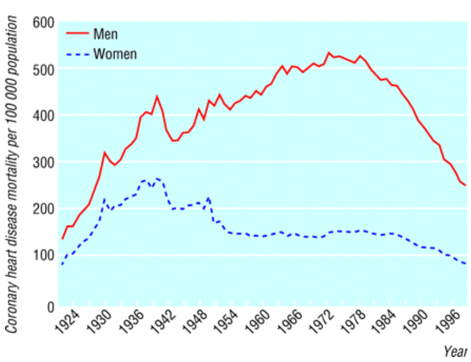

Figure 1.1: Secular trends in age-standardised mortality per 100,000 population from coronary heart disease for men and women, 1921–98, England and Wales[2]

Few graphs use confidence intervals:

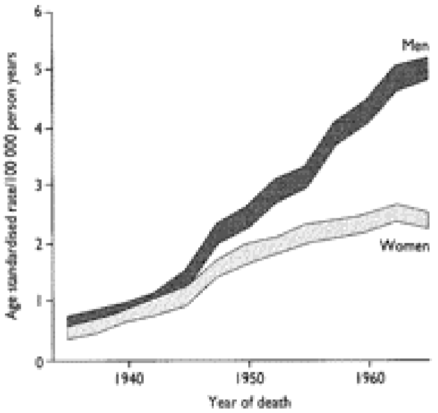

Figure 1.2: 95% Confidence bands around age-standardised five year average mortalities per 100,000 person-years from malignant melanoma: Australia 1960–4 to 1990–4[3]

Line graphs artificially connect the rates at consecutive time-points, and impose an interpretation on the reader of the pattern of the age-standardised rate by year. If more time-points were collected and graphed, the shape of the line would be subject to more variation although the overall trend would remain unchanged. At each time-point, 95% CIs can be estimated for the age-standardised rate. Separate 'I' bands can be displayed for each observation, or the upper CIs can be joined together, as can the age-standardised rate and the lower CIs, to give a series of three connected lines. Another option is to shade in the area between the upper and lower CIs to give a 95% confidence band.

Is there any statistical reason why CIs are often not displayed in publications? Or is it appropriate to use CIs for these type of graphs in future National Injury Surveillance Unit (NISU) reports? Including CIs would provide a measure of the variability of the data, and indicate that the mortality profile of two groups differ when the CIs of the two groups do not overlap.

1.1 Justification for use

A search of the medical literature and other publications[4, 5] identified the following considerations when graphing data. Cleveland[6], Kosslyn[7], and Gillan *et al* [8] advocate the use of error bars to show variability in the data being graphed. This involves placing an 'I' bar on each plotting symbol in a line graph or on the topmost horizontal of a bar in a bar graph. Bryant[9] suggests that if the author wishes to show the variation of the group mean, then it is better to use a CI bar to show this, as opposed to a mean ± standard error bar. In this context, standard error bars usually depict 95% confidence limits.[6]

1.2 Common problems with use

Bryant[9] noted that a common problem with error bars is that they are often undefined in the text and can add clutter to the graph. Cleveland[6] suggests that error bars should be unambiguous and clearly explained in the text or in the caption. A recent experimental study showed that reader comprehension of CIs was improved when a definition was included with the graph.[5] The explanation of the CI used was that they 'indicate statistical uncertainty about each value on the graph. Longer intervals mean more uncertainty. When two intervals overlap there is more uncertainty that the groups are really different.'[4]

1.3 Strategies for best presentation

Confidence intervals work well in a graph when there are only 2–6 data points, but when there are many points, they can obscure the data and detract from the depiction of any trends. If the graph is too cluttered, CIs can be put in a table accompanying the graph or in the text of the results section.

In line, bar or column graphs that use error bars to show variability, do not make the error bars thick and dark relative to the data point markers on the graph.[8] Figure 1.3 shows a line graph where the error bars overlap and are cluttered so that the reader cannot distinguish the error bars for data at the same level of the independent variable on the x axis.

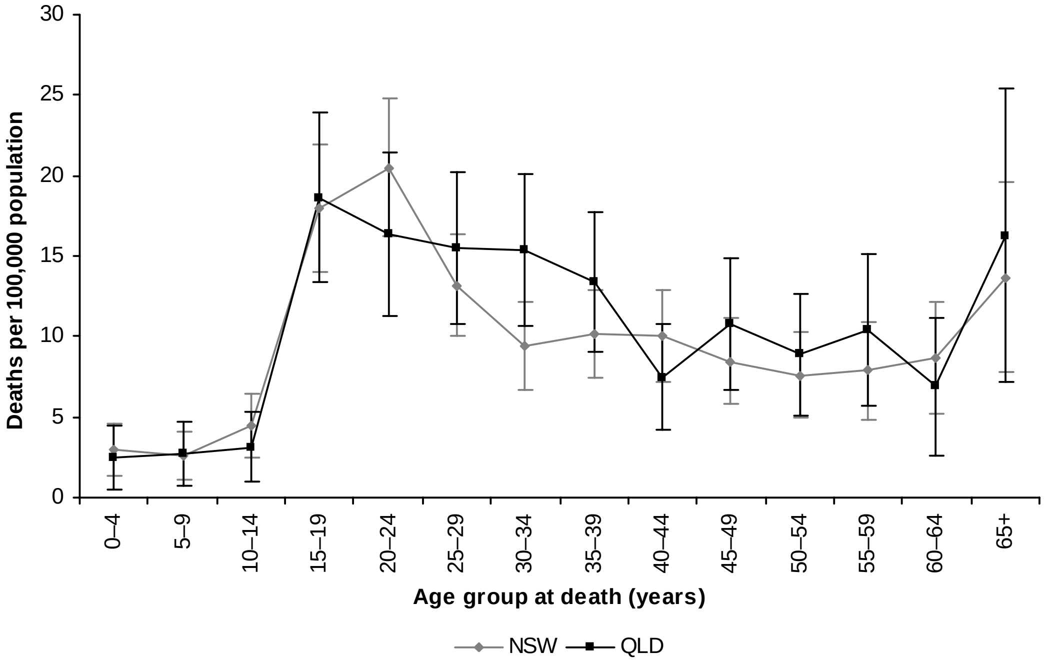

Figure 1.3: Line graph—age-standardised mortality rates due to unintentional motor vehicle traffic accidents, NSW and QLD, 2000

To avoid this situation, the following suggestions are recommended:

1 Display the data in a bar or a column graph—the bar indicators for data at the same level of the independent variable are not vertically aligned (as in Figure 1.3), so the error bars won't overlap (see Figure 1.4).[8] However, for most graphs of trends, line graphs are easier to follow than bar graphs.

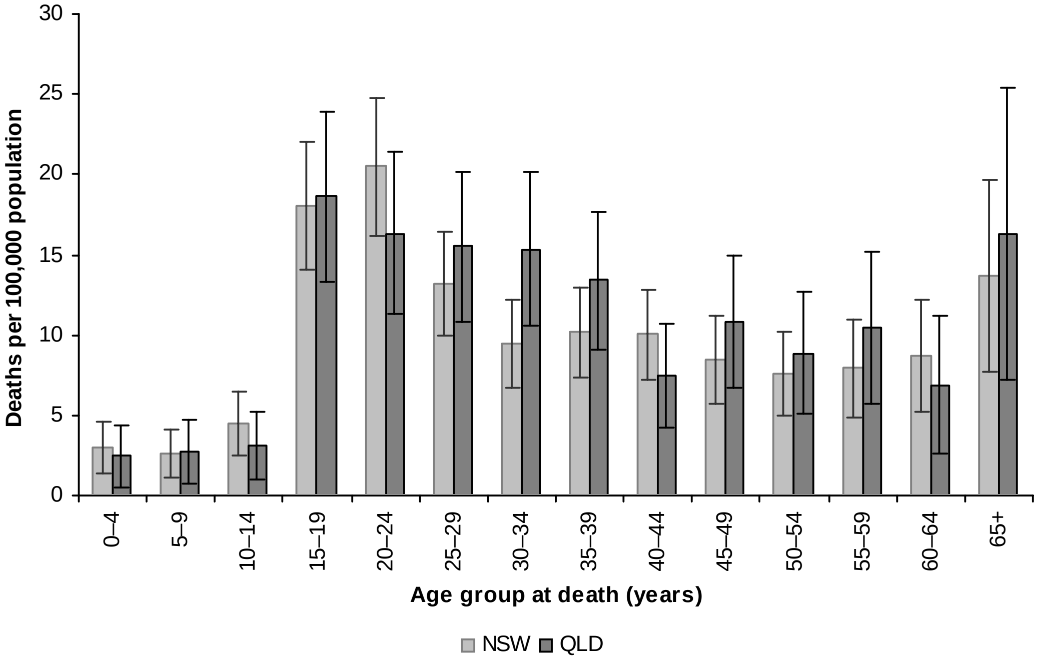

Figure 1.4: Bar graph—age-standardised mortality rates due to unintentional motor vehicle traffic accidents, NSW and QLD, 2000

2 Show only the top half of the error bar on the upper line and the bottom half of the error bars on the bottom line (see Figure 1.5).[8]

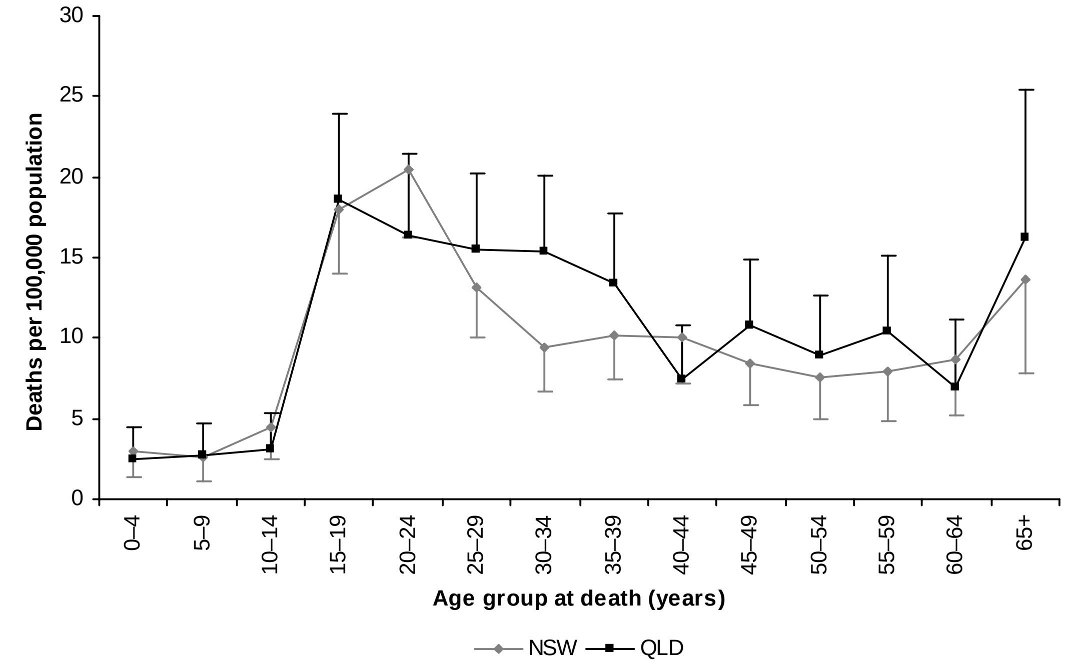

Figure 1.5: Line graph with half 'I' bars—age-standardised mortality rates due to unintentional motor vehicle traffic accidents, NSW and QLD, 2000

Kosslyn supports the use of half 'I' bars in all bar and line graphs, though recommends full 'I' bars in scatter plots.[7] However, CIs are not always symmetrical about the point estimate, which is often the case in the analysis of population health data and therefore half 'I' bars should be used judiciously. Many readers find half 'I' bars difficult to comprehend even when CIs are symmetrical, for this reason it may be preferable to avoid using half 'I' bars altogether.

3 If there are too many data points (see Figure 1.6), such that the graph becomes unintelligible, then a summary plot using just the mean values is advisable (see Figure 1.7).[9]

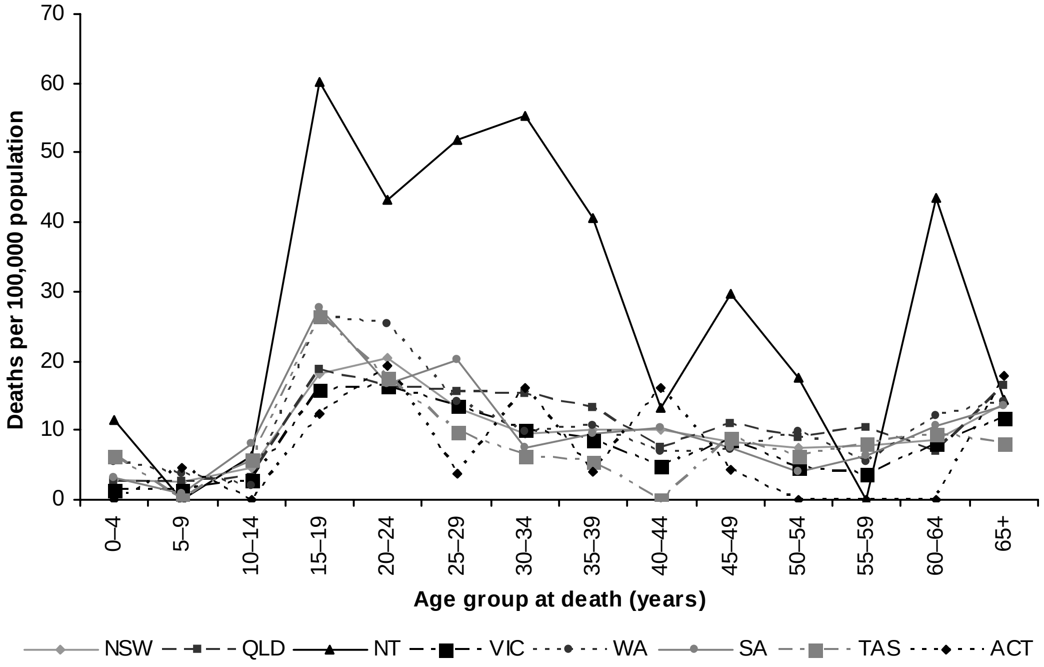

Figure 1.6: Line graph—age-standardised mortality rates due to unintentional motor vehicle traffic accidents, all States, 2000

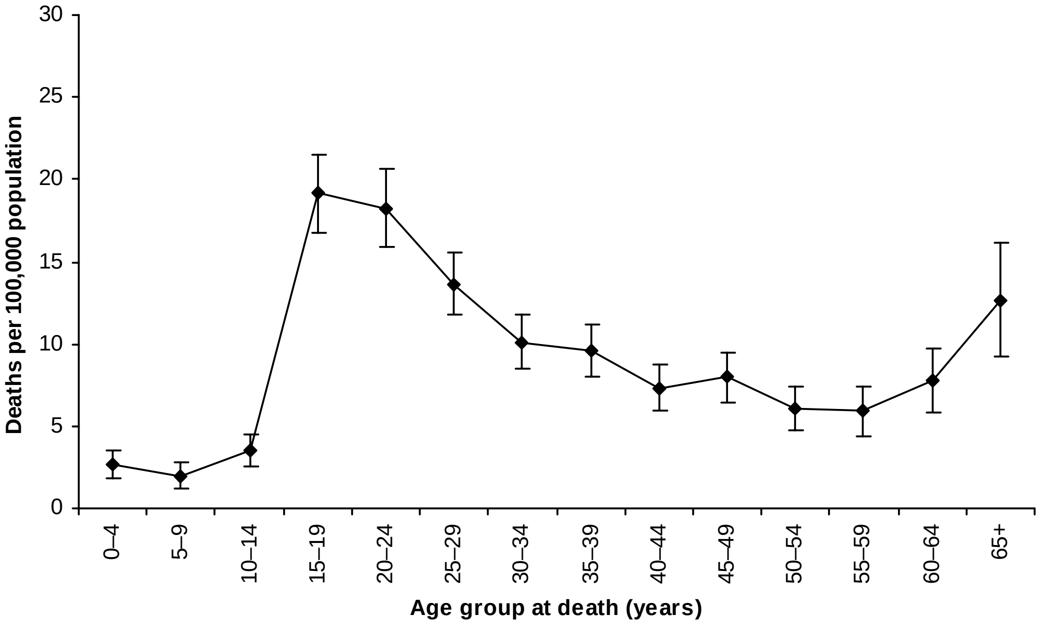

Figure 1.7: Summary plot—age-standardised mortality rates due to unintentional motor vehicle traffic accidents, all Australia, 2000

The summary plot must present a reasonable summary of the trends seen from individual plots and should not hide interesting effects because of the averaging process.[9] In Figure 1.6, the Northern Territory had a much higher traffic accident mortality rate than the other States and Territories. This fact is hidden if Figure 1.7 is the only graph presented. In this instance, it may be preferable to present the data for each of the states as a series of individual plots using identical scales.[9] Examining interstate differences for injury risk is often considered important, therefore it is likely to be worthwhile including a series of individual plots. Alternatively, the average mortality rate can be graphed and the state differences described in the text or tabulated.

1.4 Issues of independence

A basic assumption in constructing CIs is that the observations are independent.[10] That is, the errors associated with one observation are not correlated with the errors of any other observation. In general, we would expect mortality data to meet the assumption of independence between events, except in epidemics.[11] However, there may be instances where the assumption of independence between events does not hold. An example is yearly trends in hospital separation rates for a disease where individuals may be hospitalised many times e.g. asthma. A person who is hospitalised once for asthma is more likely to be readmitted for asthma than a person who has never been hospitalised for asthma. This may make it likely that errors for observations between adjacent time periods are more highly correlated than for observations more separated in time; known as autocorrelation.[10] Most commonly, autocorrelation leads to standard errors that are too small, so that the CIs are too narrow.[12]

For any NISU graphs, it is worthwhile considering whether autocorrelation is plausible (e.g. hospital separation data for a disease with multiple readmissions). Testing for autocorrelation would require statistical analysis of the data. Of the methods described in Section 3 for measuring trends over time, least squares, logistic and Poisson regression assume that errors in the modelled observations are independent (uncorrelated).[12] Yearly counts of deaths data and hospital separations are likely to follow a Poisson distribution, but if there is greater variability due to autocorrelation which can be determined using a goodness-of-fit test for the Poisson model, the Negative Binomial Distribution can be used to account for overdispersion (see Section 3). Time series analysis is another statistical method available which assumes errors are correlated.[12] However, it requires a particular data format and advanced technical expertise that makes it impractical for routine NISU surveillance.

1.5 Conclusion

For future reports for NISU which depict age-standardised mortality or hospital separation rates over a period of time, it is acceptable to report CIs as 'I' bands, connected upper and lower lines around the mean or as shaded bands unless the graph is too cluttered, in which case, plotting just the mean values is advisable. Confidence intervals are not a test for differences between means or for detecting trends over time, rather appropriate statistical tests must be applied. A consideration before constructing CIs is whether the observations meet the assumption of independence. There are instances where the assumption of independence between the numbers of occurrences of the event in disjoint time intervals may not hold for mortality and morbidity data.

1.6 References

1. Washington State Department of Health. Guidelines for using confidence intervals for public health assessment; 2002. Available at http://www.doh.wa.gov/data/guidelines/WordDocs/CI_guidelines.pdf (accessed Jul 2004).
2. Lawlor DA, Ebrahim S, Davey Smith G. Sex matters: secular and geographical trends in sex differences in coronary heart disease mortality. BMJ 2001;323(7312):541–5.
3. Giles GG, Armstrong BK, Burton RC, Staples MP, Thursfield VJ. Has mortality from melanoma stopped rising in Australia? Analysis of trends between 1931 and 1994. BMJ. 1996;312(7039):1121–5.
4. The Hunter Valley Research Foundation. Best practice in graph design. Volume 1: Literature reviews (final draft). Maryville, NSW; Hunter Valley Research Foundation, 2004.
5. The Hunter Valley Research Foundation. Best practice in graph design. Volume 11: Experimental Study (final draft). Maryville NSW; Hunter Valley Research Foundation, 2004.
6. Cleveland WS. The elements of graphing data. Murray Hill NJ; AT and T Bell Laboratories, 1994.
7. Kosslyn SM. Elements of graph design. New York; WH Freeman and Company, 1994.
8. Gillan DJ, Wickens CD, Hollands JC, Carswell CM. Guidelines for presenting quantitative data in HFES publications. Human Factors 1998; 40(1): 28–41.
9. Bryant TN. Presenting graphical information. Pediatr Allergy Immuno 1999;10: 4–13.
10. Chen X, Ender PB, Mitchell M, Wells C. Regression with Stata [Web book]. Available at http://www.ats.ucla.edu/stat/stata/webbooks/reg/default_short.htm (accessed Sept 2004).
11. Campbell MJ. Statistics at Square Two: Understanding Modern Statistical Applications in Medicine. London: BMJ Books, 2001.
12. Rosenberg D. Trend analysis and interpretation. Maryland, USA: Maternal and Child Health Bureau; 1997.

2 Age-standardisation and small case numbers

The term ‘standardisation’ refers to a procedure which facilitates the comparison of a summary measure (commonly mortality or morbidity rates) across groups. Most commonly, rates are standardised by adjusting for age, because death or morbid events occur with different frequencies among groups of different ages. However, the same principle can be applied to standardise groups according to other variables that may differ between them e.g. sex or remoteness of usual residence. Rates can be standardised for two or more variables simultaneously. An example is age-sex standardisation, which determines what the rates in a particular group would be if the group had the same sex and age make-up as the standard population.

Age-standardisation is a procedure for adjusting rates to minimise the effects of differences in age composition, thus enabling valid comparison of rates for populations that have differing age compositions. Age-standardisation is used to compare risks of two or more populations at one point in time (e.g. populations in different geographical areas) or one population at two or more points in time.

2.1 Mortality rates

Mortality data can be used to provide information on the health of populations relative to one another and to assess changes in mortality over time. To do this, the data must meet two criteria, 1) mortality rates should relate the number of events to the population at risk, and 2) since many health outcomes vary by age, the effect of the population’s age distribution must be taken into account.

The simplest death rate is the crude mortality rate. This is defined as the total number of deaths divided by the mid-year population, and is usually expressed as a rate per 1,000 or 100,000 population. For individual age cohorts (e.g. 0–4 years), the crude mortality rates are called age-specific mortality rates. The age-specific mortality rates are defined as the ratio of the number of deaths in a given age group to the population of that age group, and are usually expressed per 1,000 or 100,000 population.

The crude mortality rate does relate the number of events to the population, but does not take into account the age distribution of the population and therefore is not suitable to be used to compare differences between population groups or for assessing change in mortality over time. Age-specific mortality rates between population groups can be compared because if the age range is narrow, age will have little effect on mortality. A summary index (e.g. age-standardised rates) of two populations are more easily compared than an entire table of age-specific mortality rates, which can overwhelm the intended audience. However, sometimes standardising mortality rates can mask important patterns in death rates, and therefore standardisation should not be viewed as a substitute for a careful examination of age-specific rates.

2.2 Rate denominators

For the purposes of population-based statistics, the aggregate population is treated as 'person-time' (usually expressed as person-years). Person-years can be defined as the sum of the number of years that each member of a population is at risk of dying (or developing a certain illness). The mid-year population is often used as an estimate of person-years. There are limitations to this approach. It assumes that the mid-year population is a good estimate; there can be instances when this is not the case. An example would be when you have a population that has seasonal fluctuations in numbers e.g. a tourist resort town. At mid-year, the population may be much lower than at other times of the year, and the use of the mid-year population as a denominator may not be appropriate. For population-based statistics, there needs to be careful consideration of what population denominator is the best choice. The Australian Bureau of Statistics provides the estimated resident population for each year for each of the States and Territories. This breaks the population down by age, sex and geographical location. More detailed breakdown of the population (e.g. income, Indigenous status, country of birth) is usually only provided for census years. This can constrain analyses—for example, if you want to look at trends over time for a health measure by Indigenous status, you may not be able to easily derive the population denominators you need.

There may be instances where the choice of population denominator is not straightforward. For example, population denominators based on the usual place of residence may not be the right choice if you are looking at mortality among the homeless, or if the population of interest were migratory, especially if movement is related to the phenomena being studied (e.g. elderly people moving from the country to the city to be closer to medical treatment).

2.3 Why age-standardise mortality rates?

Analysis of cohort data typically involves comparing the mortality rates observed in the study group with the rates for the general population. A common application of adjustment to remove confounding is the age adjustment of mortality rates. This permits comparison of mortality risk for various groups free from the distortion introduced by one group having a different age distribution than another.[1]

Directly and indirectly standardised mortality rates are two basic methods commonly used. The corresponding comparative measures are known as the age-standardised rate ratio or comparative mortality figure (CMF) and the standardised mortality ratio (SMR). Any method of standardisation carries the risk of oversimplification and important information may be lost through use of these approaches to data analysis. A single combined measure (CMF or SMR) may obscure what is going on in each age group. The researcher should always compare age-specific rates to see whether the contrasts between study populations vary greatly with age. If the age-specific deaths rates do not have a consistent relationship in the study populations being compared, standardisation should be avoided as it will not indicate that these differences exist, instead it tends to mask the differences.[2]

2.4 Direct standardisation

Obtained by applying the age-specific study population rates to the age distribution of the standard population. Comparison of directly standardised rates between different groups is intended to eliminate the differences that are observed in the crude rates solely by virtue of one group having a different age structure (or a different distribution of some other variable, such as sex or remoteness) from another.[3]

2.4.1 Choice of standard population

It is important to note that in order to compare two age-standardised rates, the same standard population must be used. The age-standardised rates should be viewed as relative indexes rather than actual measures of risk. Other uses of age-standardised rates such as comparing results to studies that used a different standard population for age-standardisation are invalid.

There are two basic types of standard populations, internal or external. Internal standards are the total pooled population of the study groups to be compared. Internal standards are commonly used, but a limitation is that rates standardised to a specific study population are not as readily compared to age-standardised rates from other studies.

External standards are standard populations drawn from sources outside the analysis. Choice of an external standard is arbitrary, depending on the purposes of the study, but conventions apply. For studies that have an international focus, a standard population that is commonly used is 2000 World Standard Population.[4] In Australia, the convention followed by the AIHW and the Australian Bureau of Statistics (ABS) publications is to use the most recent census for a year ending in one as the standard population. That is, undertake age-standardisation using 2001 Census data until data become available from the 2011 Census. An advantage of choosing a commonly used standard population is that it allows comparisons of age-standardised rates with other published studies.

The general consensus of the scientific literature is that selection of the standard population should not affect relative comparisons, although it will affect the absolute values of the standardised rates.[5] Ideally the standard population selected should reflect a distribution not greatly different from that of the populations being studied.[5] If the age-specific rates in the study populations have a roughly consistent relationship, the choice of standard population should not substantially affect comparisons, but if the age-specific rates are not consistent, comparisons will depend on the standard population selected.[5] When the study populations are small, it can be difficult to decide whether there is a consistent relationship in age-specific rates, due to the large random variation associated with small numbers.

2.4.2 Directly standardised rate

The directly standardised rate for a study population is written as the weighted average of the age-specific rates. Age groups are most often specified as five-year bands (e.g. 0–4 years, 5–9 years) up to some limit (e.g. 80–84 years, 85 years and older).

$$\mathrm{DSR} = \sum \mathrm{w}_{\mathrm{si}}\ \mathrm{r}_{\mathrm{i}} \qquad (1)$$

$\mathrm{r}_{\mathrm{i}} =$ age-specific mortality rates for each age band in the cohort group = $\frac{\mathrm{D}_{\mathrm{i}}}{\mathrm{P}_{\mathrm{i}}}$ (the number of deaths in an age interval, divided by the mid-year population in the age-interval). Usually expressed as per 1,000 or 100,000 population.

$\mathrm{w}_{\mathrm{si}} = \frac{\mathrm{P}_{\mathrm{si}}}{\sum \mathrm{P}_{\mathrm{si}}}$ (standard population in each of the age bands, divided by the total standard population).

The width of the age bands and the starting age for the oldest age group are often constrained by the data available. For example, estimates for a population might be available only in 10 year bands to 60 years and older, or in single year bands to 100 years and older. In many circumstances, case data would be very sparse if divided into single year age groups. Five-year age groups are often used. With the ageing of the Australian population, the number of people (and cases of many conditions) at older ages has increased, and it is becoming common to use '95 years and older' as the oldest group in place of '75 years and older' or '85 years and older'. Small denominator values can introduce instability into directly standardised rates and rate ratios (Section 2.5) and this should be considered when choosing an age band width and an oldest age group.

It is most common to apply adjustment to data for all ages, but the same methods can be used to adjust a narrower age range. This may be useful where age structure within the age range of interest differs between groups for which rates are to be compared, and rates of the condition of interest vary with age. These conditions apply, for example, to trends in mortality due to falls by older people in Australia. Directly age-standardised rates for, say, ages '65 years and older' can be calculated using as weights $(\mathrm{w}_{\mathrm{si}})$ the age group specific proportions of the part of a standard population aged '65 years and older'.

Table 2.1: Mortality from all causes of injury for two fictitious towns

		Alton			Newton	
Age	Deaths	Person-years		Deaths	Person-years	
(yrs)	D_i	P_i	r_i (x1,000)	D_i	P_i	r_l (x1,000)
0–24	3	1,542	1.9	6	1,831	3.3
25–59	351	43,522	8.1	468	48,902	9.6
60+	532	30,265	17.6	52	2,889	18.0
Total	**886**	**75,329**	**11.8**	**526**	**53,622**	**9.8**

	Standard population (All State)				
Age (yrs)	Deaths	Person-years			
	(D_{si})	(P_{si})	r_{si} (x1,000)	w_{si}	w_i^2
0–24	1,142	176,131	6.5	0.046	0.002
25–59	25,799	3,021,675	8.5	0.787	0.619
60+	11,519	641,842	17.9	0.167	0.028
Total	**38,460**	$(\sum P_{si})$ **3,839,648**	(r_s) **10.0**	**1.000**	**0.649**

Crude death rate in Alton = 11.8 per 1,000

Crude death rate in Newton = 9.8 per 1,000

Directly age-standardised rates:

Alton = [(1.9)0.046 +(8.1)0.787 + (17.6)0.167] = 9.4 per 1,000

Newton = [(3.3)0.046 + (9.6)0.787 + (18.0)0.167] = 10.7 per 1,000

The overall crude death rate is higher in Alton even though Newton has higher mortality rates for each age group (Table 2.1). Why is this so? Older age is the major contributor to all external causes of injury mortality. Alton has an older age structure than Newton, and therefore its overall mortality is heavily weighted by high rates in the oldest age group. The population of the whole State is used as a standard to adjust for differences in age composition between the two towns. The result is that when adjusting for age, Newton has a higher injury mortality rate compared to Alton (the opposite pattern to the crude rates).

2.4.3 Standard error

Rates calculated from groups of limited size may be based on a relatively small number of cases. A weakness of direct standardisation is that the a-priori weights w_i take no account of the precision with which component rates are estimated. The data for a single age interval may make a major contribution to the standard error if the corresponding rate is based on a small denominator yet might be given a large weight.[3] The standard error is therefore useful as a measure of the statistical precision with which the rate is determined.

The calculation of the standard error (the square root of the variance) of the age-standardised rate is shown below.[3, 5]

$$\mathrm{Var(DSR)} = \sum w_{\mathrm{si}}^{2}\ \mathrm{var}(\mathrm{r_i}) \qquad (2)$$

$$\mathrm{SE(DSR)} = \sqrt{\sum \mathrm{w}_{\mathrm{si}}^{2}\ \mathrm{var}(\mathrm{r_i})} \qquad (3)$$

$$\mathrm{var}(\mathrm{r_i}) = \frac{\mathrm{D_i}}{\mathrm{P_i^2}} = \frac{\mathrm{r_i^2}}{\mathrm{D_i}} \qquad (4)$$

Using the example in Table 2.1, the variances for the age-standardised rates are:

Alton = $[(1.9^2/3)0.002 + (8.1^2/351)0.619 + (17.6^2/532)0.028] = 0.134$

Newton = $[(3.3^2/6)0.002 + (9.6^2/468)0.619 + (18.0^2/52)0.028] = 0.299$

The standard errors for the age-standardised rates are:

Alton = $\sqrt{0.134} = 0.367$

Newton = $\sqrt{0.299} = 0.548$

2.4.4 Comparative Mortality Figure (CMF)

The Comparative Mortality Figure (CMF) is a summary measure of the incidence or mortality rate ratios between the study and standard population that accounts for possible confounders such as age.[3]

The simplest way to view this measure is as a rate ratio of two directly standardised rates (which have been derived using the same standard population).[3]

For example, the rate ratio of the directly standardised rates for Newton and Alton is 10.7/9.4 = 1.14.

The formula for the CMF is more complicated, but is included to further clarify the logic behind this measure:[6]

$$\mathrm{CMF} = \frac{\sum \mathrm{P_{si}} \frac{\mathrm{D_i}}{\mathrm{P_i}}}{\sum \mathrm{P_{si}} \frac{\mathrm{D_{si}}}{\mathrm{P_{si}}}} = \frac{\text{Expected deaths (in standard population)}}{\text{Observed deaths (in standard population)}} \qquad (5)$$

$$\text{Alton CMF} = \frac{[176131(3/1542) + 3021675(351/43522) + 641842(532/30265)]}{[176131(1142/176131) + 3021675(25799/3021675) + 641842(11519/641842)]}$$

$$\text{Newton CMF} = \frac{[176131(6/1831) + 3021675(468/48902) + 641842(52/2889)]}{[176131(1142/176131) + 3021675(25799/3021675) + 641842(11519/641842)]}$$

For Alton, the CMF is 0.94

For Newton, the CMF is 1.07

The ratio of CMFs for Newton and Alton is 1.07/0.94 = 1.14 (the same result can be achieved by simply calculating the rate ratio of the two town's directly age-standardised rates).

The CMF can be expressed as the ratio of the mortality rate that would be expected for the whole State (for example) if it had the mortality experience of the study population, and the mortality rate that the whole State actually has. The CMF is often multiplied by 100 for expression as a percentage. A CMF of over 1.0 (or over 100%, depending on whether it is converted to a percentage) represents an unfavourable mortality experience.[6]

2.4.5 Standard error

The standard error of the CMF is:

$$\mathrm{SE(CMF)} = \frac{\sqrt{\left(\sum \mathrm{P}_{\mathrm{si}}^2 \frac{\mathrm{D_i}}{\mathrm{P_i}^2}\right)}}{\sum \mathrm{P_{si}} \frac{\mathrm{D_{si}}}{\mathrm{P_{si}}}} \qquad (6)$$

$$\text{Alton SE(CMF)} = \frac{\left(\sqrt{176131^2(3/1542^2)} + \sqrt{3021675^2(351/43522^2)} + \sqrt{641842^2(532/30265^2)}\right)}{38460}$$

$$\text{Newton SE(CMF)} = \frac{\left(\sqrt{176131^2(6/1831^2)} + \sqrt{3021675^2(468/48902^2)} + \sqrt{641842^2(52/2889^2)}\right)}{38460}$$

Using the example in Table 2.1, the SE(CMF) for Alton is 0.052 and for Newton is 0.083.

Because of the skewed distribution of the CMF it is necessary to transform it to the log scale. The approximate standard error for the transformed CMF is:

$$\mathrm{SE(logCMF)} = \frac{\mathrm{SE(CMF)}}{\mathrm{CMF}} \qquad (7)$$

The log transformed standard error is 0.055 for Alton and 0.077 for Newton.

Methods for calculating confidence intervals for the DSR and the CMF are outlined in Section 2.9.

2.5 Direct standardisation and small numbers

2.5.1 Stability of the CMF

A major disadvantage of the CMF is its instability when the component rates are based on small numbers of deaths. This problem is illustrated using a hypothetical example below:[3]

Table 2.2: Fictitious data used to illustrate the instability of the CMF

	Cohort		Standard population	
Age stratum (yrs)	**Deaths**	**Person-years**	**Deaths**	**Person-years**
45–64	10	10,000	140	150,000
65–84	9	3,000	290	70,000
85+	1	1	30	210
Totals	**20**	**13,001**	**460**	**220,210**

Table 2.9 in Breslow and Day, p73.[3]

The cohort CMF is:

$$\frac{150{,}000(10/10{,}000) + 70{,}000(9/3{,}000) + 210(1/1)}{460} = 1.24$$

However, if the single member in the 85+ age group were to survive instead of die, the same calculation gives

$$\frac{150{,}000(10/10{,}000) + 70{,}000(9/3{,}000) + 210(0/1)}{460} = 0.78$$

Thus, a change in only one death has made a large difference in the comparative analysis.

2.5.2 Modelling the stability of the CMF

The precision of the CMF can be explored by modelling curves of various combinations of the numerator and denominator for the 85+ age stratum. Figure 2.1 depicts the curves which result from increasing the denominator and numerator for the 85+ age group in a step-wise manner to see what size of a denominator achieves only a minor change in the CMF. In Figure 2.1, case numbers range from 0–5 and the denominator ranges from 0–75. When the case number in the 85+ age group is 1 and the denominator is 1 (as outlined in the worked example) the CMF is 1.24, whereas if the case number is 0, the CMF is 0.78. However, the CMF becomes more stable when the denominator increases to around 25. As case numbers increase in the 85+ age group, the denominator needs to be larger for the CMF to attain little change compared to when there are 0 cases in the 85+ age group. The choice of cut-off is arbitrary and depends on what degree of statistical precision the researcher would like for the CMF, but a rule of thumb—when the denominator is around 30 or greater in each age group, the CMF is fairly stable.

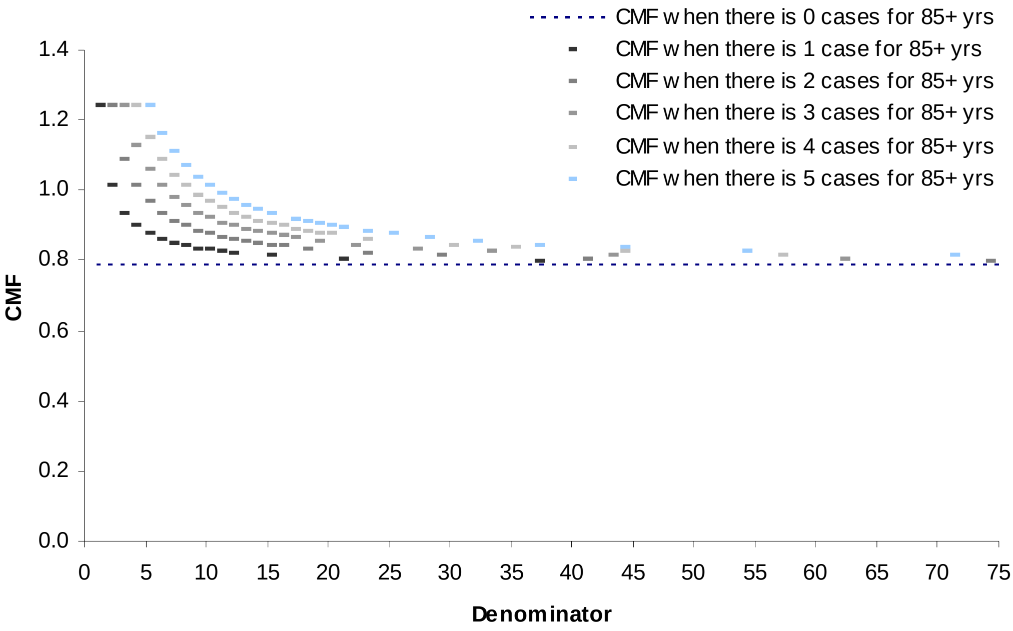

Figure 2.1: Size of denominator needed to reduce instability when the numbers in an age-stratum are small when using direct standardisation

2.5.3 Instability when numbers are small

The precision of the CMF is defined by the variance of the age-standardised rate. As the denominator increases, the variance of the age-standardised rate decreases. Although an increase in the numerator will elevate the variance of the age-standardised rate, the denominator is more important for the stability of the CMF [see equation (4)].

For most common causes of mortality (e.g. cancer, coronary vascular disease) the numerator and denominator in each age-specific stratum will be large enough so that there is no reason to suspect instability in the CMF when using direct standardisation. For rare causes of mortality (e.g. lymphomas, spinal cord injuries) the numerator may be small (only several cases may occur in an age-specific stratum), but the denominator will be large enough for the CMF to be stable. The stability of the CMF is likely to be a problem in the following instances:

1. When age-strata are small because of a decline in population numbers (e.g. 85+ years).
2. When age-strata are stratified (e.g. gender, Indigenous status, remoteness zone) leading to small numbers.
3. When the cause of mortality is common in a small age-strata (i.e. the numerator is large relative to the denominator). Although it is not shown in Figure 2.1, the CMF will be unstable when there are 10 cases for a denominator of 30.

For injury mortality, the first two instances may occur, but the third is unlikely to be a consideration as injury mortality is usually a rare occurrence, and unlikely to be a major cause of death in any particular age-strata.

2.5.4 Strategies to reduce instability

The following strategies are suggested for data in which there are small numbers in an age-stratum. Look at age-strata and use the principles of numerator/denominator combinations in Figure 2.1 as a guide to determine which age-strata may be unstable for direct standardisation. Then there are the following options:

1. Combine age-strata that are small due to a decline in population numbers (e.g. if there were less than 30 people in an 85+ age-stratum, you might collapse the category down to 80+ years or 75+ years etc).
2. When stratification by other variables (in addition to age stratification) results in small denominators for any one stratum, you might consider collapsing categories to increase the numbers within (e.g. combining the remote and very remote zones).
3. Use indirect standardisation to obtain the standardised mortality ratio (SMR) for the strata of interest and compare against the CMF to see whether there are any discrepancies which might suggest that using direct standardisation is problematic. If the findings are dissimilar for the SMR and CMF then it is best to look more closely at the underlying strata to see if the bias lies in the SMR or CMR (see Section 2.7).[3]
4. If a preliminary examination of the case numbers and denominators in the age-strata leads you to think that using direct standardisation will be problematic, then use indirect standardisation instead.

2.6 Indirect standardisation

2.6.1 Indirectly standardised rate

Indirect standardisation is less commonly used than direct standardisation, but has been purported as being useful when age-specific numbers of deaths for the cohort are unavailable. [5] However, the situations in which the age and sex of those to whom an event occurs (such as death or hospitalisation) are not known, are rare, and most injury datasets are likely to contain information on age and sex.[6] Indirect standardisation may be preferable under some circumstances, such as when age-specific denominators are small.[5] Indirect standardisation applies the age-specific rates from the standard population to the age-distribution of the study population. The indirect method calculates how many deaths would be expected in each group if the age-specific rates of the standard population were applicable.[5]

$$\mathrm{ISR} = \frac{\mathrm{r_s\ D_i}}{\sum \mathrm{r_{si}\ P_i}} \quad \text{or} \quad \frac{\mathrm{D_i}}{\sum \mathrm{r_{si}\ P_i}} \cdot \mathrm{r_s} \qquad (8)$$

$\mathrm{r_s}$ = crude rate for the standard population. Usually expressed as per 1,000 or 100,000 population.

$\mathrm{D_i}$ = total number of deaths in the study population

$\mathrm{r_{si}}$ = age-specific mortality rate in each five-year age band in the standard population

$= \frac{\mathrm{D_{si}}}{\mathrm{P_{si}}}$ (the number of deaths in an age interval, divided by the mid-year population in the age-interval). Usually expressed as per 1,000 or 100,000 population.

$\mathrm{P_i}$ = the population of each five-year age band in the study population.

Using the example in Table 2.1, the indirectly age-standardised rates are:

Alton = 886 x 1000 / [(6.5)1542 + (8.5)43522 + (17.9)30265] x 10.0 = 9.6 per 1,000

Newton = 526 x 1000 / [(6.5)1831 + (8.5)48902 + (17.9)2889] x 10.0 = 11.0 per 1,000

The indirect standardised rates are similar to those obtained by the direct method in Section 2.4.2.

2.6.2 Standardised Mortality Ratio (SMR)

More frequently, the ratio of observed deaths to expected deaths is presented. This ratio is called the Standardised Mortality Ratio (SMR). If incidences are used instead of deaths, then the ratio is called the Standardised Incidence Ratio (SIR).

$$\text{SMR} = \frac{\text{observed deaths (in study population)}}{\text{expected deaths (in study population)}} = \frac{D_i}{\sum r_{si} \; P_i} \qquad (9)$$

This SMR (or SIR) is usually expressed as a percentage by multiplying by 100.

Using the data in Table 2.1, the SMR for Alton (886/921.7) is 0.96 and for Newton (526/479.3) is 1.10. The ratio of the SMRs (1.10/0.96) gives 1.14, the same as the ratio of the CMFs obtained in Section 2.4.4.

A comparison of the formulae [equation (5) and (9)] for the SMR and CMF reveal two differences. Firstly, the CMF is the quotient of expected over observed deaths with reference to the standard population, whereas the SMR is the quotient of observed over expected deaths with reference to the study population.[6] This makes no difference to their interpretation—if both the study and standard population had the same distribution, then the CMF and SMR would give identical results. [6] Secondly, the age and sex sub-group weights used in the denominator of the SMR depend on the characteristics of the study population, whereas those used in the denominator of the CMF do not.[6] If the SMRs for several study populations are compared, the weights used to create the weighted sum of sub-group specific mortality rates will differ, and therefore the relative importance assigned to deaths in different sub-groups will differ between study populations. [6] This means the SMRs are not standardised with each other, although these use the same 'standard' population to provide their expected rates. This means that two study populations can only be compared via their SMRs if they have identical population structures (which is highly unlikely). In contrast, the CMFs from several study populations will have the same denominator - the observed number of deaths from the standard population, allowing direct comparisons of their CMFs.[6]

One advantage of the SMR over the CMF is that age-specific numbers of deaths are not required for its calculation. It suffices to know only the total number of deaths in the study population.[3] This can be useful for published data, where details on the numbers of death by cause, subgroup and age may be left out preventing the application of the CMF.[3] However, in most cases this data is available.

2.6.3 Standard error

The variance of the SMR is:

$$\mathrm{Var(SMR)} = \frac{\sum \mathrm{D_i}}{\left(\sum \mathrm{r_{si}}\ \mathrm{P_i}\right)^2} \tag{10}$$

Using the example in Table 2.1, the Var(SMR) for Alton $886/(921.7)^2$ is 0.001 and for Newton $526/(479.3)^2$ is 0.002.

The standard error of the SMR is:

$$\mathrm{SE(SMR)} = \frac{\sqrt{\sum \mathrm{D_i}}}{\sum \mathrm{r_{si}}\ \mathrm{P_i}} = \frac{\sqrt{\mathrm{Observed}}}{\mathrm{Expected}} \tag{11}$$

Using the example in Table 2.1, the SE(SMR) for Alton $(\sqrt{886}/921.7)$ is 0.032 and for Newton $(\sqrt{526}/479.3)$ is 0.048.

As with the CMF, it is usual practice to use the log transformed SMR to account for its skewed distribution. The approximate standard error for the transformed SMR is:

$$\mathrm{SE(logSMR)} = \frac{\mathrm{SE(SMR)}}{\mathrm{SMR}} = \frac{1}{\sqrt{\mathrm{Observed}}} \tag{12}$$

The log transformed standard error is 0.034 for Alton and 0.044 for Newton. The SMR tends to have a smaller standard error than the CMF (see Section 2.4.5) as it is the maximum likelihood estimate.[7]

2.7 Indirect standardisation and small numbers

2.7.1 Stability of the SMR

The SMR is a weighted average of the ratios of age-specific rates for the cohort and standard population.[3]

From the hypothetical example in Table 2.2, the expected number of deaths for the cohort is determined as follows:[3]

$$10{,}000(140/150{,}000)+3{,}000(290/70{,}000)+1(30/210)=21.90$$

$$\text{The SMR}=\frac{20}{21.90}=0.91$$

This indicates a slightly lower death rate among the cohort members as opposed to the general population. However, if the single member in the 85+ age group were to survive instead of die:

$$\text{The SMR}=\frac{19}{21.90}=0.87$$

This is only relatively minor change compared to that observed earlier with the CMF in Section 2.5.1.[3] The SMR tends to be less sensitive to numerical instabilities in one or two of the age-specific rates.[3]

2.7.2 Limitations of the SMR

The standard error of the SMR depends only on fluctuations in the total number rather than in the age-specific number of deaths, therefore it is generally smaller than the CMF. [3] The SMR weights the ratios optimally, in inverse proportion to their statistical precision, whereas the weights associated with unstable ratios may be much larger with the CMF. [3] This means the SMR is more appropriate when the study population size is small. However, there are also statistical disadvantages to using the SMR. Indirect standardisation is sometimes incorrectly used to compare the mortality experience of different study populations. The SMRs from study populations can only be legitimately compared with the standard (e.g. 1.00 or 100) and not with each other because different weighting is used to generate each SMR (the weights depend on the age distribution of the study population). This inability to compare the mortality experience of study populations is the major disadvantage of using indirect standardisation, and provides the motivation for using direct standardisation, which is not limited by this constraint.

A problem with indirect standardisation is that the ratio of two SMRs determined by pooling observed and expected deaths across age groups may sometimes lie completely outside the range of the age-specific rate ratios, as shown in Table 2.3 below.[3]

Table 2.3: Example of misleading ratios of SMRs

		Age range (years)		
Cohort		20–44	45–64	Total (20–64)
I	Deaths no.	100	1,600	1,700
	Expected no.	200	800	1,000
	SMR_1 (%)	50	200	170
II	Deaths no.	80	180	260
	Expected no.	120	60	180
	SMR_2 (%)	67	300	144
	SMR_1/SMR_2	75	67	118

Table 2.13 in Breslow and Day, p73.[3]

The overall SMR for each cohort is the weighted average of the two age-specific observed/expected deaths ratio, the weights being proportional to the expected number of deaths. Since Cohort I has more older people, the high observed/expected deaths ratio for the 45–64 year age interval is weighted more heavily (the overall SMR is 170%), whereas in Cohort II much more emphasis is given to the lower observed/expected deaths ratio in the 25–44 year age interval (the SMR is lower, 144%). The overall result is a change in sign in the apparent effect, from excess deaths in Cohort II (compared to Cohort 1) on an age-specific basis to an apparent excess in Cohort I when the data are pooled.[3] The CMF is not subject to this problem as the ratio of two CMFs are the ratio of directly standardised rates (which are the weighted average of the age-specific rate ratios). However, in practice, the CMF and SMR usually provide numerical results that are similar. In cases in which they differ, it is not necessarily true that the CMF is more nearly 'correct'.[3] It is important to determine whether the bias is due to extreme sensitivity to small numbers for the CMF (especially in stratum-specific denominators) or whether the bias lies in the SMR, by looking more closely at the underlying strata.

There are three conditions under which the SMR and CMF give substantially different results. The first is when there are non-negligible differences in the age distributions of the study group(s) and the standard population. Indirect standardisation produces biased results in this situation, due to residual confounding by age, but direct standardisation is not affected. The second is when the ratio of mortality rates of the study group(s) compared to the standard population vary substantially with age. The third is when both of these factors occur together.[3]

2.8 Calculating confidence intervals

Confidence intervals can be calculated by statistical programs such as Stata or in spreadsheets that are constructed in programs such as Microsoft Excel. Using these resources is time-saving and reduces the chance of errors by calculations done by hand. It is beyond the purposes of this report to explain in detail how to perform analyses that produce CIs using statistical software programs, however some of the methods are briefly outlined below. In Stata, crude rates and their confidence intervals can be calculated by exponentiating coefficients using Poisson regression (or alternatively, Negative Binomial Distribution regression) as demonstrated in Sections 3.8 and 3.9.

To input a dataset into Stata, variable names must not be hyphenated or greater than one word (i.e. person-years is input as personyears, age group is input as agegroup) and categories are defined by numbering (e.g. 1=0–29 years, 2=30–59 years, 3=60+ years).

Methods for calculating CIs will be demonstrated using mortality data from Rothman referring to the total deaths and populations in Sweden and Panama in 1962.[8]

```
. input nation agegroup deaths population
 1. 1 1 3523 3145000
 2. 1 2 10928 3057000
 3. 1 3 59104 1294000
 4. 2 1 3904 741000
 5. 2 2 1421 275000
 6. 2 3 2456 59000
 7. end
. label define nation 1 “Sweden” 2 “Panama”
. label val nation nation
. label define agegroup 1 “0-29 yrs” 2 “30-59 yrs” 3 “60+ yrs”
. label val agegroup agegroup
. list, nolabel
```

[the ‘list’ command can be used to view a tabulation of the dataset]

Nation	Age group		
1=Sweden, 2=Panama	1=0–29 yrs, 2=30–59 yrs, 3=60+ yrs	Deaths	Population
1	1	3,523	3,145,000
1	2	10,928	3,057,000
1	3	59,104	1,294,000
2	1	3,904	741,000
2	2	1,421	275,000
2	3	2,456	59,000

Using this dataset, the ‘tabrate’ or the ‘ci’ command provide exact Poisson CIs for crude rates. As an aside, some commands in Stata may be updates or user-written additions that need to be downloaded in Stata. By using the ‘findit’ command (e.g. ‘findit tabrate’), Stata searches for information on a topic across all Stata-related internet sources including user-written additions. From ‘findit’, you can click to go to the source to install the additions.

. tabrate deaths nation, e(population)

```
WARNING: response deaths not coded 0/1

table of cases (D), person-years (Y), and rates per 1000 person-years

  +-----------------------------------------------------+
  | nation      _D        _Y    _rate   ci_low   ci_high |
  |-----------------------------------------------------|
  | Sweden   73555   7.5e+06   9.813    9.742     9.884 |
  | Panama    7781   1.1e+06   7.238    7.079     7.401 |
  +-----------------------------------------------------+

Chisq test for unequal rates =   656.62 (1 df, p =  0.000 )
```

For Sweden, the crude rate is 9.8 deaths [95% CI: 9.7 to 9.9] per 1,000, whereas for Panama it is 7.2 deaths [95% CI: 7.1 to 7.4] per 1,000.

The ‘ci’ command (see ‘help ci’ in Stata) can also be used to calculate CIs for crude rates based on the Poisson distribution.

. bysort nation: ci deaths, exposure(population)

```
-------------------------------------------------------------------------------
-> nation = Sweden

                                                         -- Poisson  Exact --
    Variable |   Exposure        Mean    Std. Err.       [95% Conf. Interval]
-------------+---------------------------------------------------------------
      deaths |    7496000    .0098126    .0000362        .0097418    .0098837

-------------------------------------------------------------------------------
-> nation = Panama

                                                         -- Poisson  Exact --
    Variable |   Exposure        Mean    Std. Err.       [95% Conf. Interval]
-------------+---------------------------------------------------------------
      deaths |    1075000    .0072381    .0000821        .0070782    .0074008
```

An immediate form of the command ‘cii’ is available which can be used in calculator mode, with numbers for ‘exposure’ and ‘events’ instead of requiring a datafile of variables. For example, for Panama there are 7,781 deaths among 1,075,000 persons and the command is:

. cii 1075000 7781

```
                                                         -- Binomial Exact --
    Variable |        Obs        Mean    Std. Err.       [95% Conf. Interval]
-------------+---------------------------------------------------------------
             |    1075000    .0072381    .0000818        .0070788    .0074002
```

The ‘ci’ and ‘tabmore’ command have the added function of being able to provide age-specific rates and their CIs.

. quietly bysort nation agegroup: ci deaths, exposure(population)
[quietly suppresses screen output so results are not shown]
. db tabmore
. tabmore, res(deaths) typ(count) row(nation) rate col(agegroup) fup(population) ci

```
Response variable is: deaths which is count
Follow-up time variable is: pop
Row variable is: nation
Column variable is: agegroup
Number of records used:    6
95% confidence intervals

Summary using rates per 1000

----------------------------------------
          |           nation
 agegroup |        Sweden         Panama
----------+-----------------------------
 0-29 yrs |          1.12           5.27
          |   1.08 - 1.16    5.11 - 5.44
          |
30-59 yrs |          3.57           5.17
          |   3.51 - 3.64    4.91 - 5.44
          |
  60+ yrs |         45.68          41.63
          | 45.31 - 46.05  40.01 - 43.31
----------------------------------------
```

Age-standardised rates and their CIs can be calculated using the ‘dstdize’ and ‘istdize’ command for direct and indirect standardisation, respectively.

To demonstrate direct standardisation, using the Rothman data for Sweden and Panama, first choose a standard population to use, a file called ‘1962.dta’.

Age group **1=0–29 yrs, 2=30–59 yrs, 3=60+ yrs**	**Population**
1	0.35
2	0.35
3	0.30

The directly age-standardised rates and their CIs are calculated for the mortality data for Sweden (1) and Panama (2) using the ‘dstdize’ command:

. dstdize deaths population agegroup, by(nation) using(1962)

```
----------------------------------------------------------
-> nation= 1
                         -----Unadjusted-----  Std.
                                Pop.  Stratum  Pop.
  Stratum       Pop.     Cases  Dist. Rate[s] Dst[P]  s*P
----------------------------------------------------------
 0-29 yrs    3145000      3523  0.420 0.0011  0.350 0.0004
 30-59 yr    3057000     10928  0.408 0.0036  0.350 0.0013
  60+ yrs    1294000     59104  0.173 0.0457  0.300 0.0137
----------------------------------------------------------
Totals:      7496000     73555    Adjusted Cases: 115032.5
                                      Crude Rate:   0.0098
                                   Adjusted Rate:   0.0153
                       95% Conf. Interval: [0.0152, 0.0155]

----------------------------------------------------------
-> nation= 2
                         -----Unadjusted-----  Std.
                                Pop.  Stratum  Pop.
  Stratum       Pop.     Cases  Dist. Rate[s] Dst[P]  s*P
----------------------------------------------------------
 0-29 yrs     741000      3904  0.689 0.0053  0.350 0.0018
 30-59 yr     275000      1421  0.256 0.0052  0.350 0.0018
  60+ yrs      59000      2456  0.055 0.0416  0.300 0.0125
----------------------------------------------------------
Totals:      1075000      7781    Adjusted Cases:  17351.2
                                      Crude Rate:   0.0072
                                   Adjusted Rate:   0.0161
                       95% Conf. Interval: [0.0156, 0.0166]

Summary of Study Populations:
   nation             N      Crude     Adj_Rate       Confidence Interval
 --------------------------------------------------------------------------
        1       7496000   0.009813     0.015346    [  0.015235,    0.015457]
        2       1075000   0.007238     0.016141    [  0.015645,    0.016637]
----------------------------------------------------------
```

For Sweden, the directly age-adjusted rate is 15.3 deaths [95% CI: 15.2 to 15.5] per 1,000, whereas for Panama it is 16.1 deaths [95% CI: 15.6 to 16.6] per 1,000.

In the following section, equations are presented to aid the reader in constructing or using spreadsheets in programs such as Microsoft Excel or performing calculations by hand.

The Poisson distribution is asymmetrical with zero as its lower bound. If the numbers of deaths are large then normal approximations can be used to calculate confidence intervals, but care has to be exercised when both the rates are low and the numbers of deaths are small. If a normal approximation is assumed, the resulting CIs can result in the lower limit being less than zero, and death rates cannot be negative. There are a number of known methods of confidence interval estimation; some perform better than others in the case of small numbers.[9] A method in which the approximation of confidence intervals is based on the gamma distribution has been shown to outperform existing methods (including a method proposed by Dobson et al [9]) when case numbers are small and when there is large variability in the weights applied to strata in age-standardisation.[10] Anderson et al[5] demonstrates how the gamma distribution method[10] can be used to generate a set of confidence factors (see Appendix I) to apply to crude and age-specific rates and age-adjusted rates (direct and indirect) to calculate 95% CIs for Poisson distributed observations when case numbers are small (1–99 deaths). When the number of observations is greater (around 100 cases or above) a normal distribution is approximated and near symmetry is achieved. [5] Therefore, when constructing CIs, a normal approximation can be applied above this threshold to simplify calculations. If the standard error of the SMR or the CMF is to be used to construct CIs, it is necessary to make a transformation to the log scale.[3]

2.8.1 Crude rates and age-specific rates

For crude rates and age-specific rates (1–99 deaths)—see Appendix 1:

$$\begin{aligned}&\text{Lower limit : Rate x Lower confidence factor for observed deaths}\\&\text{Upper Limit : Rate x Upper confidence factor for observed deaths}\end{aligned} \qquad (13)$$

When the number of cases is 100 or more, the normal approximation may be used to calculate the CIs.

For crude rates and age-specific rates (100 or more deaths) — see Appendix 1:

$$\begin{aligned}&\text{Lower limit : Rate} - 1.96\frac{\text{Rate}}{\sqrt{\text{Deaths}}}\\&\text{Upper limit : DSR} + 1.96\frac{\text{Rate}}{\sqrt{\text{Deaths}}}\end{aligned} \qquad (14)$$

2.9 Confidence intervals for direct standardisation

2.9.1 Directly age-standardised rate

For the directly age-standardised rate (1–99 deaths)—see Appendix 1:

$$\begin{aligned}&\text{Lower limit : DSR x Lower confidence factor for observed deaths}\\&\text{Upper Limit : DSR x Upper confidence factor for observed deaths}\end{aligned} \qquad (15)$$

For a normal approximation (100 or more deaths), 95% CIs can be formed for age-standardised rates using the variances.

For the directly standardised rate (100 or more deaths):

$$\begin{aligned}&\text{Lower limit : DSR} - 1.96\sqrt{\text{var}(\text{r}_\text{i})}\\&\text{Upper limit : DSR} + 1.96\sqrt{\text{var}(\text{r}_\text{i})}\end{aligned} \qquad (16)$$

Using the example in Section 2.6 the CIs are:

Alton $= \text{DSR} \pm 1.96\sqrt{0.134} = 0.717$

$= 9.4$ per 1,000 (95% CI: 8.7 to 10.1)

Newton $= \text{DSR} \pm 1.96\sqrt{0.299} = 1.072$

$= 10.7$ per 1,000 (95% CI: 9.6 to 11.8)

2.9.2 Comparative Mortality Figure

From equations (6) and (7), the log transformed 95% CIs are given by:

$$\text{Lower limit} : \frac{\text{CMF}}{\exp\left[\dfrac{1.96 \cdot \text{SE(CMF)}}{\text{CMF}}\right]} \qquad (17)$$

$$\text{Upper limit} : \text{CMF} \cdot \exp\left[\frac{1.96 \cdot \text{SE(CMF)}}{\text{CMF}}\right]$$

Using the example in Section 2.4:

For Alton, the CMF is 0.94 (95% CI: 0.84 to 1.04)

For Newton, the CMF is 1.07 (95% CI: 0.92 to 1.24)

2.10 Confidence intervals for indirect standardisation

For the SMR (1–99 expected deaths)—see Appendix 1:

$$\begin{aligned}&\text{Lower limit : SMR x Lower confidence factor for observed deaths}\\&\text{Upper Limit : SMR x Upper confidence factor for observed deaths}\end{aligned} \qquad (18)$$

Indirectly standardised rates and CIs can be generated by multiplying the SMR and its CIs by the crude rate for the standard population (here it is 10.0).

For the SMR (100 or more deaths):

To account for the skewed distribution of the SMR, the log transformed 95% CIs from equation (12) are given by:

$$\text{Lower limit : } \frac{\text{SMR}}{\exp\left[\frac{1.96}{\sqrt{\text{Observed}}}\right]} \qquad (19)$$

$$\text{Upper limit : SMR} \cdot \exp\left[\frac{1.96}{\sqrt{\text{Observed}}}\right]$$

For Alton, the SMR is 0.96 (95% CI: 0.90 to 1.02)

For Newton, the SMR is 1.09 (95% CI: 1.00 to 1.19)

Indirectly standardised rates and CIs can be generated by multiplying the SMR and its CIs by the crude rate for the standard population (here it is 10.0).

Alton = 9.6 per 1,000 (9.0, 10.2)

Newton = 10.9 per 1,000 (10.0, 11.9)

2.11 Conclusion

Directly and indirectly standardised mortality rates are two basic methods commonly used to permit comparison of mortality risk for various groups free from the distortion introduced by one group having a different age distribution than another. The corresponding comparative measures are known as the age-standardised rate ratio or comparative mortality figure (CMF) and the standardised mortality ratio (SMR). This section has discussed the rationale for choosing an appropriate method of age-standardisation when case numbers are small. It has been shown that the size of the denominator of an age-stratum is an important factor for the stability of the CMF when using direct standardisation, and has more influence than the size of the numerator. In many instances, when the denominator is around 30 or above, direct standardisation is an acceptable method to use. A number of strategies are suggested for determining in which instances, direct standardisation can be used and when it is more appropriate to use indirect standardisation.

2.12 References

1. Kahn HA, Sempos CT. Statistical methods in Epidemiology. New York NY; Oxford University Press, 1989.
2. Fleiss JL. Statistical methods for rates and proportions. 2nd edition. John Wiley and Sons, 1981.
3. Breslow NE, Day NE. 'Rates and Rate Standardization.' In Statistical methods in cancer research. Volume II—the design and analysis of cohort studies. Lyon; International Agency for Research on Cancer, 1987.
4. Population Division of the Department of Economic and Social Affairs of the United Nations Secretariat, World Population Prospects: The 2002 Revision. Available at: http://www.un.org/esa/population/unpop.htm (accessed Aug 2004).
5. Anderson RN, Rosenburg HM. Age standardisation of death rates: implementation of the year 2000 standard. National Vital Statistics Report 1998; 47(3). Hyattsville, Maryland: National Center for Health Statistics. 1998.
6. Julious AS, Nicholl J, George S. Why do we continue to use standardized mortality ratios for small area comparisons? J Pub Health Med 2001; 23(1): 40–6.
7. Armitage P, Berry G. Statistical methods in medical research (4th edition). Oxford: Blackwell Scientific, 2002.
8. Rothman KJ. Modern Epidemiology. Boston: Little, Brown and Company. 1986.
9. Dobson AJ, Kuulasmaa K, Eberle E, Scherer, J. Confidence intervals for weighted sums of poisson parameters. Stat Med 1991; 10(3): 457–462.
10. Fay MP, Feuer EJ. Confidence intervals for directly standardised rates: a method based on the gamma distribution. Stat Med 1997; 16(7): 791–801.

3 Measuring trend

3.1 Tests for trend

There are two situations commonly encountered by a researcher where they may want to measure trend. These are:

- Trends over time e.g. mortality rates over a number of years, and
- Trends or patterns over other naturally ordered characteristics such as age.

When looking at graphs in many agency reports, trends over time in mortality rates are often determined by simply eyeballing the graph and making a statement such as 'there was an increase in mortality between 1990 and 2000 for males but not for females'. Simple graphs of annual rates can be informative and are a valid way of presenting data clearly. When there are marked trends over time or large differences between groups, this method is sufficient, but in some instances, it may not be clear-cut as to whether there is a trend over time or differences between groups.

Some journal papers and agency reports may use a formal test for trend, but often there is no discussion as to why a particular test for trend was chosen. There are many methods of test for trend in use—chi-squared statistics for trend, Pearson's correlation, ordinary least squares regression, logistic regression, Poisson regression models and others. A researcher may find it difficult or confusing to decide on which test for trend is appropriate for their data, and whether any of the tests for trend used in the literature are based on the wrong assumptions and should be avoided. The researcher may also want to choose a test that measures the magnitude of difference between groups, or per cent change over time.

One of the first considerations before embarking on any data analysis is deciding what type of data you are dealing with. A categorical (sometimes called a nominal variable) is one that has two or more categories, in which no ordering is implied (e.g. gender). If the variable has a clear ordering but the differences between categories may not be necessarily equal, then the variable is an ordinal variable (e.g. low, medium and high socioeconomic status). An interval variable is similar to an ordinal variable, except the values are evenly spaced (e.g. aged 0–4 years, 5–9 years, 10–14 years etc).[1]

It is often useful to consider both ordinal and interval variables as ordered, and to distinguish between nominal and ordered data.[1] In the following section, we will differentiate between chi-square tests that are sensitive to departures from the null hypothesis that could occur in various ways (e.g. unequal proportions between groups can occur in any combination) and tests that are suitable for testing for a certain form of departure from the null hypothesis (i.e. a trend).[1] Trend is a simultaneous test for increasing or decreasing relationship between risk and exposure.

The following section will use Stata statistical software for data manipulations.[2]

3.2 Chi-square statistics

Chi-square is the name of a class of continuous probability distributions that can be used to test the assumption under the null hypothesis of the independence of row and column classifications in a 2 X k contingency table. Chi-square statistics can be used to determine if there is an association between variables for counts, rates and proportions. The tests described in the following section include:

- Overall Pearson's chi-square
- Cochran-Armitage test for trend (chi-square test for trend and chi-square test for departure from linearity)
- Mantel-Haenszel test for trend

3.2.1 Trends for ordered data

In this section, we will consider chi-square statistics suitable for 2 X k or k X 2 contingency tables. That is, 2 rows by any number of columns, or 2 columns by any number of rows.

Consider this fictitious data for the prevalence of a disease (a k X 2 contingency table). We will examine the relationship between the prevalence of the disease and age. Are the proportions of cases in each age group homogenous, or is there a trend?

Age (years)	Age group	Cases	Non-cases
25–29	1	12	108
30–34	2	24	156
35–39	3	36	108
40–44	4	60	120
45–49	5	72	84
50–54	6	60	36
55–59	7	96	24
60–64	8	108	12

A graph of the data shows what appears to be an increasing prevalence of cases by age.

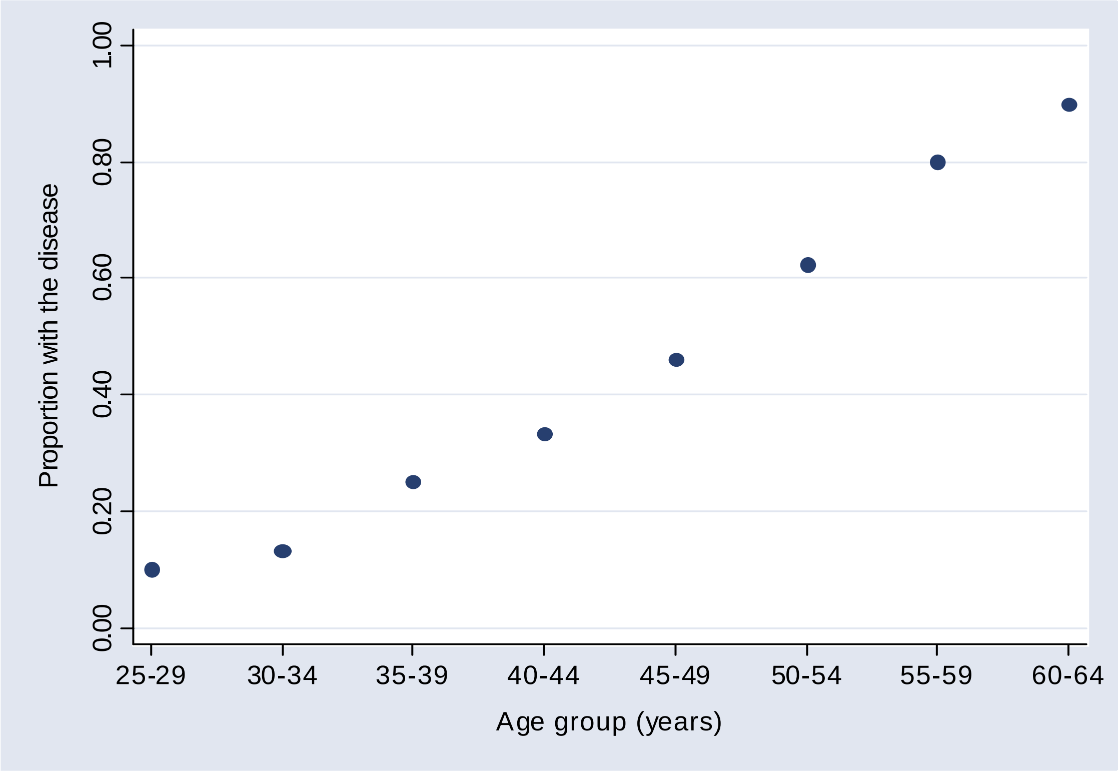

Figure 3.1: Prevalence of a disease, in chronological order

In Stata,[2] use the 'ptrend' command to give output of chi-square statistics. 'ptrend' calculates an overall chi-square (Pearson's chi-square) and calculates a chi-square statistic for the trend (linear regression) of the proportion of cases on age (a variable called _prop is generated), and also gives a chi-square test for departure from the trend line. Both the chi-square test for trend and chi-square for departure are usually performed at the same time, and are sometimes loosely lumped together under the same name 'Cochran-Armitage test for trend'.[2]

```
. ptrend cases noncases agegroup
```

```
Trend analysis for proportions
Regression of p =  cases/(cases+non-cases) on agegroup:
Slope =  .12173, std. error =  .00672, Z = 18.111


Overall chi2(7) =       336.238,  pr>chi2 = 0.0000
Chi2(1) for trend =     328.022,  pr>chi2 = 0.0000
Chi2(6) for departure =   8.216,  pr>chi2 = 0.2227
```

The overall chi-square (Pearson's chi-square) compares the observed and expected frequency counts. The null hypothesis of the Pearson's chi-square is that the rows and columns in a two-way table are independent. A rough rule is that for the valid use of the Pearson's chi-square test, relatively few expected frequency counts should be less than 5 (say 1 cell out of 5) and no expected frequency should be less than 1.[1] The number in brackets is the degrees of freedom, here it is 7 or (row–1) x (column–1).

The chi-square test for trend tests the null hypothesis that there is no association between the two variables being studied (i.e. the slope=0). If the test is significant, we reject the null hypothesis (i.e. we conclude that the slope does not equal 0). The sign of the slope coefficient indicates the direction of the trend: positive for an increasing trend and negative for a decreasing trend.

The chi-square test for departure tests for 'departures' from linearity—it is simply a goodness-of-fit test for the linear model.[3] A goodness-of-fit test compares the observed counts and those predicted by the model.[4] If the model fits the data well, then the observed and expected counts should be close, the chi-square statistics will be small and the corresponding p-value will be large and non-significant. If the model is not a good fit, the chi-square statistic will be large and the p-value will be small (<0.05) and significant.[4]

The chi-square test for trend has a reduced number of degrees of freedom (1 df) and is likely to be satisfactory, provided that only a small proportion of expected frequencies are less than about 2 and that these do not occur in adjacent rows.[1] The chi-square test for departure is likely to be adequate if only a small proportion of the expected frequencies are less than about 5.[1] Its degrees of freedom are (k – 2) or the number of rows – 2.

In the example above, the conclusion is that age and the disease are associated. The proportion of cases present in each age group is not the same (chi-square test for difference in distributions, p<0.001) and the test for trend is significant (p<0.001). The chi-square for departure is not significant (chi-square test for trend, p=0.2227) so a linear model is a good fit for the data. The overall chi-square and the chi-square test for trend are both significant—but what are each of these are actually testing? What would we get if we used the same case numbers and mixed up the age groups, so we disguised the trend by age?

Age (years)	Age group	Cases	Non-cases
25–29	6	12	108
30–34	5	24	156
35–39	2	36	108
40–44	1	60	120
45–49	8	72	84
50–54	7	60	36
55–59	4	96	24
60–64	3	108	12

Looking at the graph, when the age groups are mixed up there is no longer a linear trend.

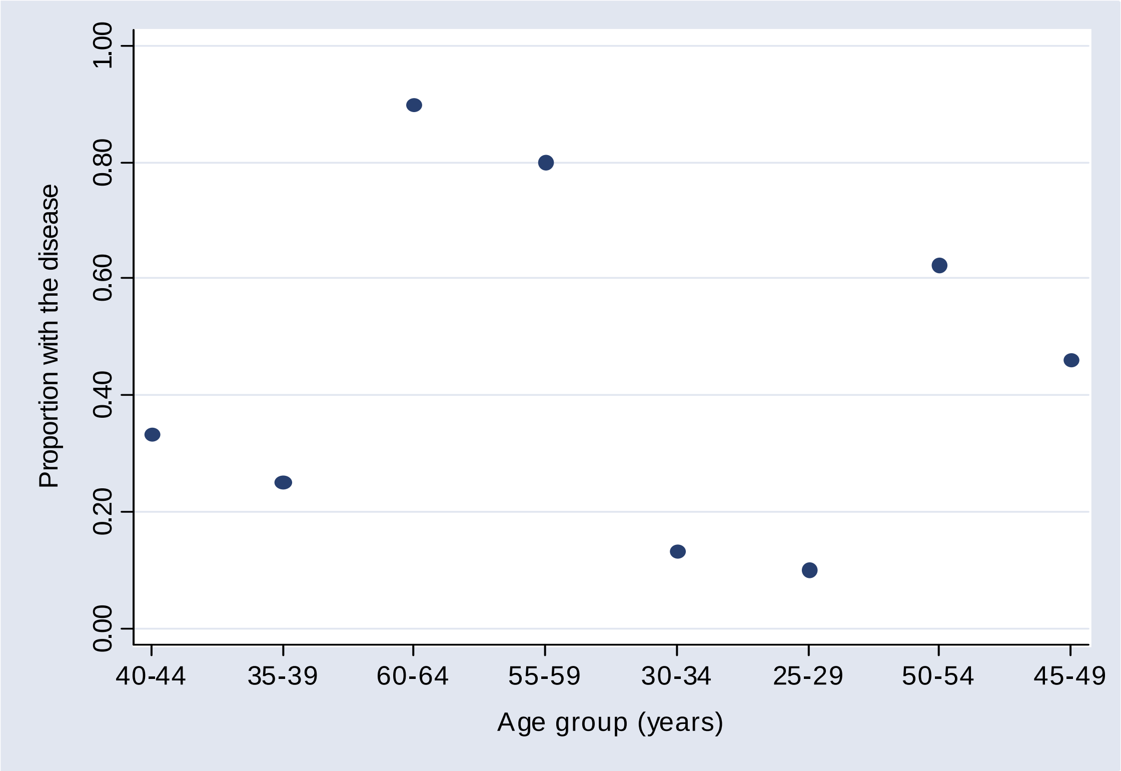

Figure 3.2: Prevalence of a disease, not in chronological order

In Stata, use the 'ptrend' command to give output of chi-square statistics.

. ptrend cases noncases agegroup

```
Trend analysis for proportions
Regression of p = cases/(cases+non-cases) on agegroup:
Slope = -.00081, std. error =  .00626, Z =   0.129


Overall chi2(7) =       336.238,  pr>chi2 = 0.0000
Chi2(1) for trend =       0.017,  pr>chi2 = 0.8970
Chi2(6) for departure = 336.221,  pr>chi2 = 0.0000
```

The overall chi-square analysis shows identical results to the previous example. So the overall chi-square test is simply testing whether the 11 proportions are the same (the null hypothesis) or different (the alternative hypothesis) and is not a test for trend (an increasing or decreasing proportion). The chi-square test for trend uses linear regression to correctly identify that there is no trend over time in incident cases of the disease (p=0.8970). The chi-square for departure is significant (p<0.001) so a linear trend is not a good model for the data.

This example highlights that it is inappropriate to use an overall chi-square (Pearson's chi-square) to test for trend. The chi-square statistics tests for overall association between the rows and columns and assumes no ordering of either the rows or the columns. For dose-response or trend data, we need to use different statistics. The data is presented as a binary outcome (disease or no disease, case or no case) for which we have an ordered margin (which is the row or column depending how the data is set up). The chi-square test for trend can be used to identify trends over time when the data is ordered in this manner.

Mantel-Haenszel test for trend

The Mantel-Haenszel chi-square test for trend is very similar to the Cochran-Armitage test,[5] and gives only slightly different test statistics.[3, 5] Using the same fictitious data for the prevalence of a disease, the example below demonstrates that the Cochran-Armitage test for trend gives comparable results to the Mantel-Haenszel chi-square test for trend.

The Mantel-Haenszel chi-square test for trend can be calculated using the 'tabodds' command in Stata. The data needs to be entered in a different format to ptrend.

Age (years)	Age group	Disease	Observations
25–29	1	0	108
30–34	2	0	156
35–39	3	0	108
40–44	4	0	120
45–49	5	0	84
50–54	6	0	36
55–59	7	0	24
60–64	8	0	12
25–29	1	1	12
30–34	2	1	24
35–39	3	1	36
40–44	4	1	60
45–49	5	1	72
50–54	6	1	60
55–59	7	1	96
60–64	8	1	108

```
. expand observations
(1100 observations created)
```

. tabodds disease agegroup

```
--------------------------------------------------------------------------
  agegroup  |      cases     controls       odds      [95% Conf. Interval]
------------+-------------------------------------------------------------
          1 |         12          108    0.11111        0.06120   0.20173
          2 |         24          156    0.15385        0.10010   0.23644
          3 |         36          108    0.33333        0.22859   0.48606
          4 |         60          120    0.50000        0.36676   0.68164
          5 |         72           84    0.85714        0.62567   1.17425
          6 |         60           36    1.66667        1.10255   2.51940
          7 |         96           24    4.00000        2.55741   6.25632
          8 |        108           12    9.00000        4.95713  16.34011
--------------------------------------------------------------------------
Test of homogeneity (equal odds): chi2(7)  =   335.94
                                  Pr>chi2  =   0.0000

Score test for trend of odds:     chi2(1)  =   327.73
                                  Pr>chi2  =   0.0000
```

The ‘or’ option specifies that odds ratios rather than odds be displayed.

. tabodds disease agegroup, or

```
---------------------------------------------------------------------------
    agegroup |  Odds Ratio       chi2       P>chi2     [95% Conf. Interval]
-------------+-------------------------------------------------------------
           1 |    1.000000          .            .             .          .
           2 |    1.384615       0.76       0.3849      0.662430   2.894131
           3 |    3.000000       9.86       0.0017      1.459309   6.167303
           4 |    4.500000      21.42       0.0000      2.237492   9.050312
           5 |    7.714286      41.72       0.0000      3.704492  16.064335
           6 |   15.000000      65.84       0.0000      6.264706  35.915493
           7 |   36.000000     118.29       0.0000     12.582646 102.999000
           8 |   81.000000     152.96       0.0000     19.802094 331.328592
---------------------------------------------------------------------------
Test of homogeneity (equal odds): chi2(7)  =   335.94
                                  Pr>chi2  =   0.0000

Score test for trend of odds:     chi2(1)  =   327.73
                                  Pr>chi2  =   0.0000
```

The test is on the log odds scale which means ‘tabodds’ reports whether there is a trend in the odds (or probability) of disease across exposure levels. Because the table is a 8 by 2, the test for equality of the odds ratio is a chi-square with (row–1) x (column–1) degrees of freedom, with (8–1) x (2–1) or 7 degrees of freedom. Because the test for homogeneity of the odds is significant at p<0.001, there is indication that there is a difference among the exposure levels. The score test for trend indicates that there is a dose-response relationship. It has 1 degree of freedom because it is testing the slope of a regression line. Both the test of homogeneity chi-square (335.94) and the score test for trend of the odds (327.73) give very similar test statistics to the Pearson’s chi-square (336.238) and chi-square test for trend (328.022) using the Cochran-Armitage test for trend (‘ptrend’ command) at the beginning of Section 3.1.1.

Another command which uses average ranks instead of different scores is ‘nptrend’. ‘nptrend’ performs a nonparametric test for trend across ordered groups and is an extension of the Wilcoxon rank-sum test. The ‘nptrend’ command has extra flexibility compared to ‘ptrend’ as it allows the inclusion of more groups in the table (e.g. so that the Pearson's correlation can be extended to 3 X 3 tables). A modification of the 'nptrend' command (which can be located in Stata using 'findit snp12') is available which allows for stratified tests for trend using exact and large sample methods.[6] This modified command is well-suited for matched stratified data that are sparse and remains valid even if any cell counts in the stratums are 0s or 1s.

3.2.2 Trends for ordered person-time data

A test for trend is available in Stata for person-time data. It is called 'tabrate' and it calculates prevalence rates with 95% confidence intervals using the Poisson distribution. Consider the following fictitious data:

Year	Cases	Population
1980	12	1,000
1981	24	1,000
1982	36	1,000
1983	60	1,000
1984	72	1,000
1985	60	1,000
1986	96	1,000
1987	108	1,000

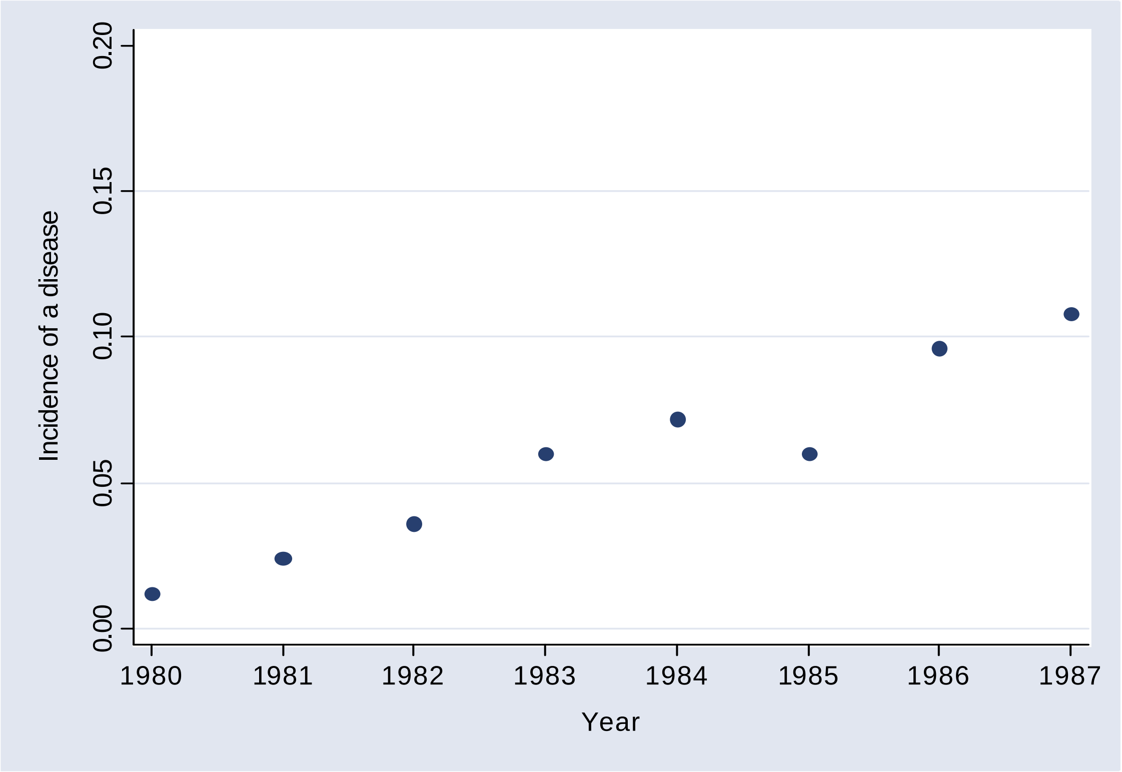

Figure 3.3: Incidence of a disease, 1980–1987, in chronological order

. tabrate cases year, e(population) graph trend

```
table of cases (D), person-years (Y), and rates per 1000 person-years

  +---------------------------------------------------+
  | year    _D       _Y     _rate   ci_low    ci_high |
  |---------------------------------------------------|
  | 1980    12   1000.0    12.000    6.815     21.130 |
  | 1981    24   1000.0    24.000   16.086     35.807 |
  | 1982    36   1000.0    36.000   25.968     49.908 |
  | 1983    60   1000.0    60.000   46.587     77.275 |
  | 1984    72   1000.0    72.000   57.150     90.708 |
  | 1985    60   1000.0    60.000   46.587     77.275 |
  | 1986    96   1000.0    96.000   78.595    117.259 |
  | 1987   108   1000.0   108.000   89.437    130.416 |
  +---------------------------------------------------+

chi-squared for trend    126.71 ( 1 df, p =  0.000 )
```

The chi-square test for trend in significant, so there is trend in incidence rate of the disease over time.

If the data were reordered, so there was no linear trend.

Year	New order	Cases	Population
1980	6	12	1,000
1981	5	24	1,000
1982	2	36	1,000
1983	1	60	1,000
1984	8	72	1,000
1985	7	60	1,000
1986	4	96	1,000
1987	3	108	1,000

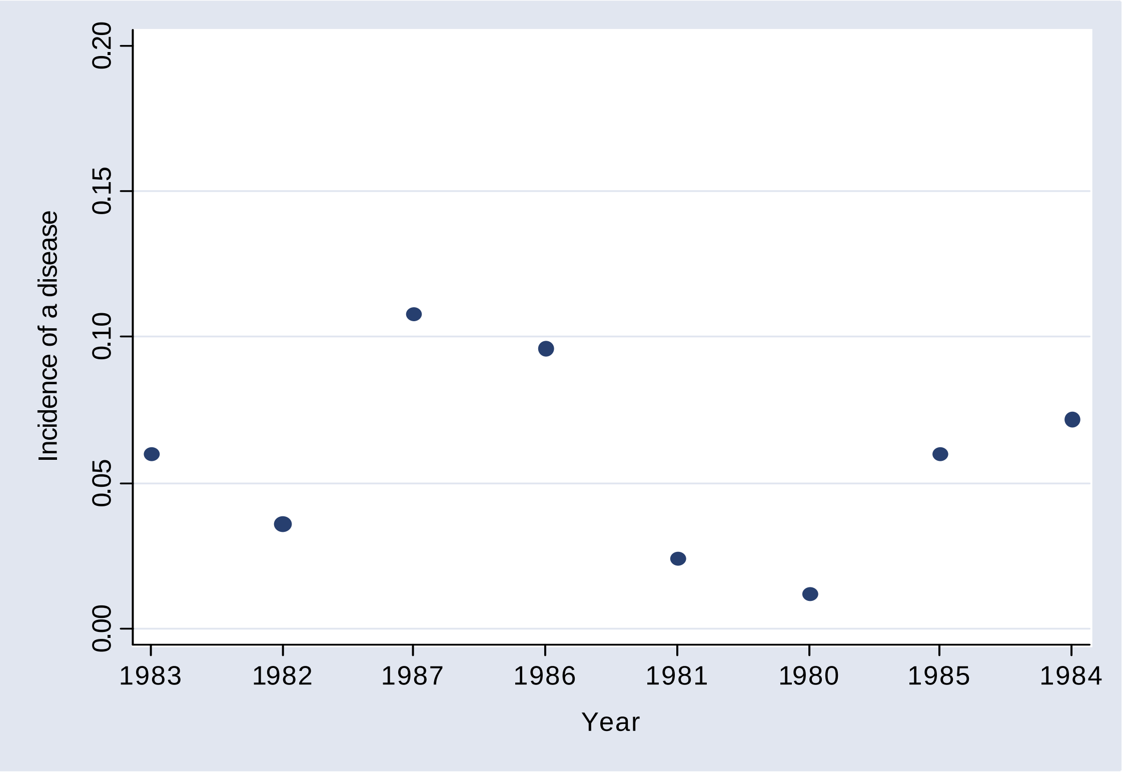

Figure 3.4: Incidence of a disease, 1980–1987, not in chronological order

. tabrate cases neworder, e(population) graph trend

```
table of cases (D), person-years (Y), and rates per 1000 person-years

  +-------------------------------------------------------+
  | neworder    _D       _Y     _rate    ci_low   ci_high |
  |-------------------------------------------------------|
  |        1    60   1000.0    60.000    46.587    77.275 |
  |        2    36   1000.0    36.000    25.968    49.908 |
  |        3   108   1000.0   108.000    89.437   130.416 |
  |        4    96   1000.0    96.000    78.595   117.259 |
  |        5    24   1000.0    24.000    16.086    35.807 |
  |        6    12   1000.0    12.000     6.815    21.130 |
  |        7    60   1000.0    60.000    46.587    77.275 |
  |        8    72   1000.0    72.000    57.150    90.708 |
  +-------------------------------------------------------+

chi-squared for trend      2.48 ( 1 df, p =  0.116 )
```

The chi-square test for trend is not significant, so there is no trend in the incidence rate of the disease.

3.2.3 Limitations of chi-square tests

Sribney[3] evaluated various tests for trend (including Mantel-Haenszel chi-square test for trend, Pearson's correlation, 'ptrend' or Cochran-Armitage test, and 'nptrend' command) and demonstrated that these tests are simply a Pearson's correlation coefficient (a test suitable to determine the correlation between two variables and its significance). A Pearson's correlation coefficient uses linear regression to compute the slope, and the null hypothesis is that the slope is equal to zero. As Pearson's correlation coefficient is a measure of the strength of the linear relationship between two variables, it assumes the relationship would be a straight line on a scatterplot. This test is powerful against alternative hypotheses of consistently increasing or decreasing trend, but not at all powerful against curvilinear (or other) associations, with no linear component.[3] If the relationship is curvilinear or non-linear, the graph of the relationship might bend or even resemble a 'U' and Pearson's correlation coefficient will understate the true correlation, sometimes to the point of being useless or misleading. In these instances, the usual overall Pearson's chi-square can detect the association with more power.

False 'linear' trends can be obtained if there is an incremental (series of steps with levelling off) rather than smooth dose-response or trend relationship. For this reason, it is important to graph or tabulate the data to see if a trend makes sense in the range of exposures of interest, and check for any incremental increases in odds ratios, relative risks, or incidence rate ratios.

To demonstrate that the chi-square test for trend (using the 'ptrend' command) is almost identical to linear regression using the example in 3.2.1:

. regress _prop agegroup

```
Source |       SS       df       MS              Number of obs =       8
-------------+------------------------------           F(  1,     6) =  255.69
       Model |  .617652158     1  .617652158           Prob > F      =  0.0000
    Residual |  .014493562     6  .002415594           R-squared     =  0.9771
-------------+------------------------------           Adj R-squared =  0.9733
       Total |   .63214572     7  .090306531           Root MSE      =  .04915

------------------------------------------------------------------------------
       _prop |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    agegroup |   .1212683   .0075838    15.99   0.000     .1027114    .1398252
       _cons |  -.0953068   .0382964    -2.49   0.047    -.1890146    -.001599
------------------------------------------------------------------------------
```

The slope determined by linear regression (0.1212683) is similar to that obtained using the 'ptrend' command (0.12173) for the variable 'age group', and both methods show a trend over time with p<0.001.

Similarly, for the example in 2.3.1 in which the case numbers were the same, but the age groups were mixed up to disguise the trend by age:

. regress _prop agegroup

```
Source |       SS       df       MS              Number of obs =       8
-------------+------------------------------           F(  1,     6) =    0.00
       Model |  .000515308     1  .000515308           Prob > F      =  0.9465
    Residual |  .631630412     6  .105271735           R-squared     =  0.0008
-------------+------------------------------           Adj R-squared = -0.1657
       Total |   .63214572     7  .090306531           Root MSE      =  .32446

------------------------------------------------------------------------------
       _prop |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    agegroup |  -.0035027   .0500647    -0.07   0.946    -.1260065    .1190011
       _cons |    .466163   .2528141     1.84   0.115    -.1524509    1.084777
------------------------------------------------------------------------------
```

The slope determined by linear regression (-0.0035027) is similar to that obtained using 'ptrend' (-.00081) for the variable 'age group', and both methods show no trend over time (p=0.946 and 0.897, respectively).

3.3 Regression procedures

Regression procedures can be simply thought of as methods which allow investigation of an input/output relationship.[4] A mathematical model is constructed which relates the input (variables that are thought to be related to the outcome called 'independent variables' and denoted by X) with the output (the 'dependent variable' and denoted by Y).[4] The most commonly used models are 'linear models' which assume that the X variables combine in a linear fashion to predict Y, a process which gives useful regression parameters such as the slope and intercept of the line. No model can predict the Y variable perfectly, and the model provides an error term (otherwise known as the 'residuals') to account for this. Linear models are appropriate when the outcome variable is normally distributed, but the models can be generalised so that the modelling procedure is similar for many different situations such as when the distribution is non-normal or discrete (can only be integers) (see Section 3.4).[4]

Common regression procedures include Ordinary Least Squares regression, logistic regression, and Poisson regression (see Table 3.1). The choice of an appropriate model should be guided by the characteristics of the outcome variable (Y). For example, when the outcome variable is continuous, then an appropriate analysis is linear regression (see Table 3.1). Regression procedures are more powerful tests of association than chi-square statistics.[1] That is, for a given sample size, one can demonstrate an effect with a narrower confidence interval (smaller p-value). Another advantage is that regression procedures are able to generate estimates of future rates as well as average annual per cent change, neither of which are automatically obtained when conducting a chi-square test for linear trend.[7] Graphs of the predicted values, projected values and their confidence bands can be plotted.[7] Another advantage of using regression models is that other variables can be included in the model and can be simultaneously adjusted for.[1] For example, a regression that models the association between time and the outcome of interest, can also include variables that are confounders and effect modifiers of this association, thereby adjusting the predicted rates, projected rates and their confidence bands appropriately.[7]

Table 3.1: Common regression models

Outcome (Y)	Example	Regression model	Estimate
Continuous	Lab values (e.g. blood alcohol levels)	Linear	$\Delta Y/\Delta X$
Interval	1,2,3,4	Linear	$\Delta Y/\Delta X$
Ordered categorical	Scales (good, better, best)	Logit	OR
Number of events	Deaths/year	Poisson	IRR
	Strokes/1000 persons	Poisson	RR
Binary (0/1)	Disease, death	Logistic	OR
Time to event	Death after operation	Cox regression	HR

OR = odd ratio; RR = relative risk; IRR = incidence rate ratio; HR = hazard ratio

3.4 Generalised linear models

The concept of generalised linear models places all the commonly used regression models into a unified framework.[1] In its simplest form, a linear model specifies the linear relationship between a dependent variable (Y) and a set of predictor variables (Xs) (also referred to as covariates).

$$Y = \text{Intercept} + (\text{Slope}_1 \cdot X_1) + (\text{Slope}_2 \cdot X_2) + ... + (\text{Slope}_k \cdot X_k)$$

For example, from a sample of data measuring height and weight and recording gender, you could use linear regression to estimate (i.e. predict) a person's weight as a function of the person's height and gender. For many data analysis problems, a linear relationship between variables is adequate to describe the observed data, and to make reasonable predictions for new observations. However, there are many relationships that cannot adequately be summarised by a simple linear equation. For example, the distribution of the data may not be normal and the outcomes may not be continuous; they may be binary, multinomial (can take only a distinct number of values), skewed, or discrete.

A generalised linear model is an extension of the linear modelling process that allow models to be fit that follow probability distributions other than normal and have residuals (errors) that are not normally distributed.[1] The dependent variable values are predicted from a linear combination of predictor variables, which are 'connected' to the dependent variable via a link function.

$$g(E[y]) = g(\mu) = \text{Intercept} + (\text{Slope}_1 \cdot X_1) + (\text{Slope}_2 \cdot X_2) + ...(\text{Slope}_k \cdot X_k)$$

Where:

$E(y) = \mu$ is the mean outcome

$g(\,)$ is the function of the mean outcome. It is called the 'link' function since it is the link between the mean and the linear predictor

These generalised linear models are all characterised by:

1) a link
2) a family or error term
3) a linear predictor

Linear regression, logistic regression and Poisson regression are all members of the generalised linear model family. There are 'canonical forms' (i.e. pairs) of link functions and error structures that commonly go together. For example, the canonical form for the logit function is the binomial error structure, which results in logistic regression.

Table 3.2: Canonical forms

Regression Model	Error	Link
Linear	Normal	Identity
Logistic	Binomial	Logit
Poisson	Poisson	Log

In logistic regression, in order to relate the expected outcome E[y], where y is binary, to the linear combination of predictors, the expected outcome is transformed using the logit function. This enables the transformed mean to follow a linear model, so that the right hand side of the regression equations for both linear and logistic regressions is of similar form. The regression parameters (the intercept and slope) are derived by a general method of estimation called maximum likelihood estimation (MLE). A detailed explanation of the theory behind generalised linear models is beyond the scope of this report.

When using Stata for data manipulations,[2] there are usually two options for the syntax. A model can be specified either using 1) the 'glm' command where the link function and family distribution must be specified, or 2) the model-specific command (e.g. 'regress' 'logistic' or 'poisson'). Both command options will converge to the same result, but sometimes the 'glm' option provides extra flexibility or functions and other times the model-specific command is preferable. The help option in Stata can aid decision-making (e.g. see 'help glm', 'help regress', 'help logistic' and 'help poisson').

3.5 Ordinary Least Squares regression

For linear regression, the algorithm used to fit the data is called ordinary least squares (OLS). It determines what estimate minimises the squared distance between the observed data and the fitted values from the model. [8] The OLS model is appropriate when the dependent variable (Y) is continuous or an interval variable.[9] Each observation has a corresponding error term or 'residual', which is the difference between the actual value and the predicted value at that level of the independent predictor. When using linear regression, there are five assumptions; 1) that a linear model is appropriate for the data, 2) the observations are independent, 3) the residuals are therefore uncorrelated, 4) there is homogeneity of variance — the spread of residuals at each level of the predictor variable (X) is symmetrical around the regression line, and the spread is similar for each distribution, and 5) the residuals have a normal distribution with mean 0 and a common variance of σ^2 (Figure 3.5).[8]

For modelling rates:

$$\text{rate}_i = \text{Intercept} + (\text{Slope} \cdot \text{Year}_i) + \varepsilon_i$$

$$\text{where } i = 1 \text{ to the number of years being analysed, and } \varepsilon \text{ is an error term}$$

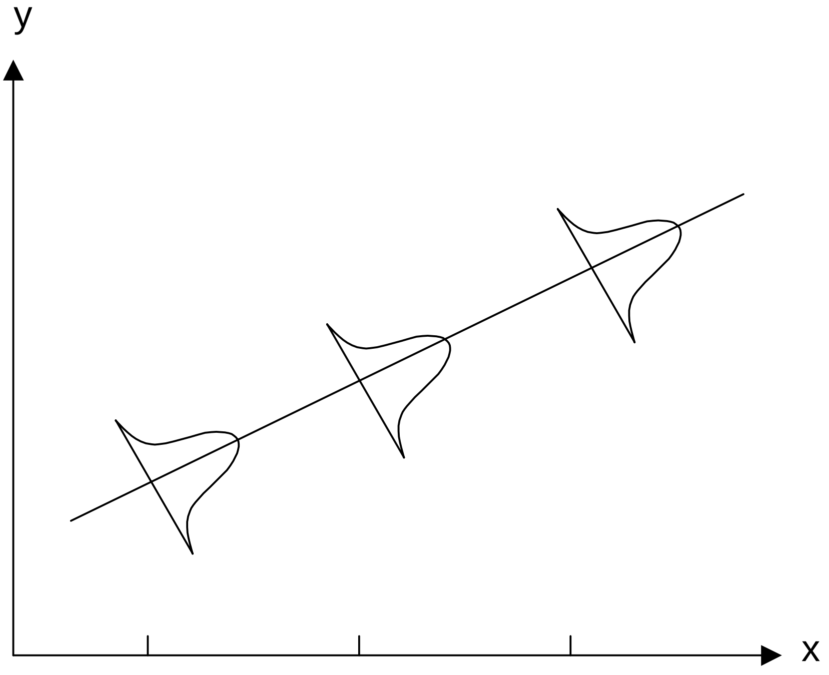

Figure 3.5: The distribution of the residuals at each level of the predictor variable

As an example of linear regression, we will look at infant mortality rates in Australia from 1900–2000 using data from the Australian Bureau of Statistics.[10] Count data often follow a Poisson distribution,[9] so some type of Poisson analysis might be more appropriate, provided the numerator (counts) and denominator (population) data were available for the infant mortality rates. However, for the purposes of this example, we will use linear regression.

Year	Infant mortality rate (per 1,000 live births)
1900	103.6
1910	74.8
1920	69.1
1930	47.2
1940	38.4
1950	24.5
1960	20.2
1970	17.9
1980	10.7
1990	8.2
2000	5.2

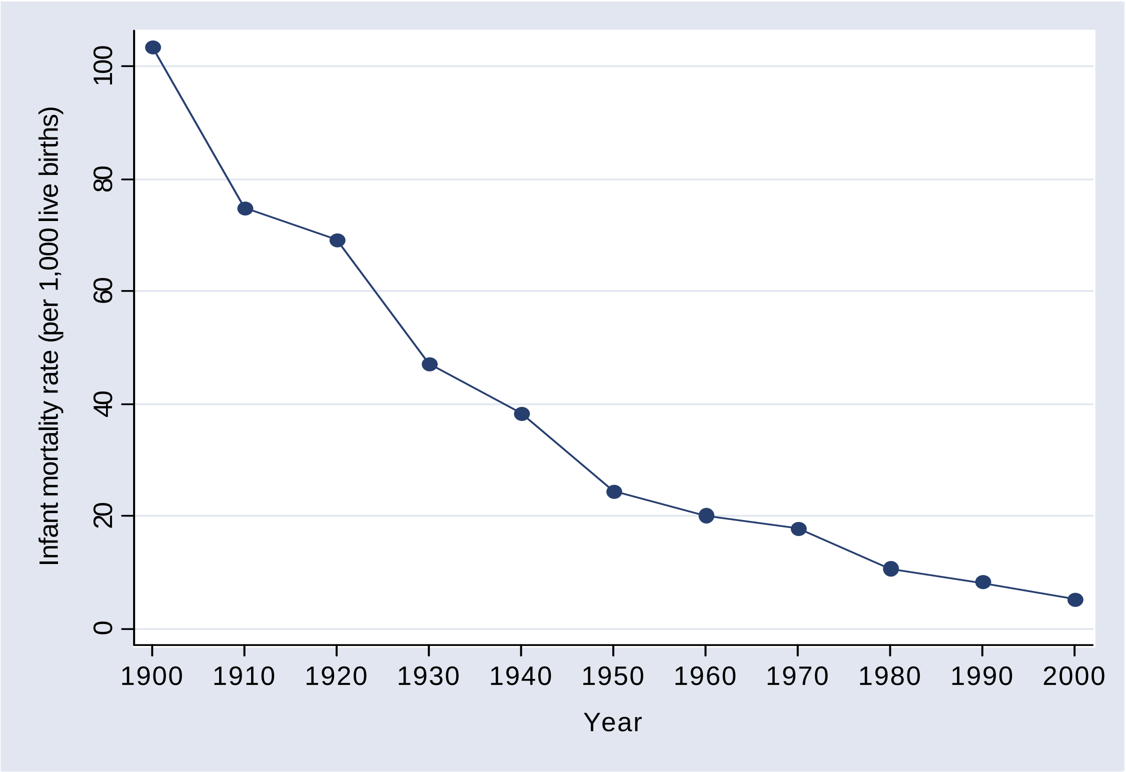

Figure 3.6: Infant mortality rate in Australia, 1900–2000

The graph shows there has been a decline in infant mortality rates over the 100 years. Infant mortality rates appear to follow a slight curvilinear trend which suggests a polynomial model may be more appropriate than a linear model. We will firstly test the linear model.

. regress infant year

```
      Source |       SS       df       MS              Number of obs =      11
-------------+------------------------------           F(  1,     9) =   80.27
       Model |  9280.98337     1  9280.98337           Prob > F      =  0.0000
    Residual |  1040.60205     9   115.62245           R-squared     =  0.8992
-------------+------------------------------           Adj R-squared =  0.8880
       Total |  10321.5854    10  1032.15854           Root MSE      =  10.753

------------------------------------------------------------------------------
      infant |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        year |  -.9185455   .1025238    -8.96   0.000     -1.15047   -.6866205
       _cons |   1829.327   199.9477     9.15   0.000     1377.014     2281.64
------------------------------------------------------------------------------
```

The estimate of the slope for the variable ‘year’ is -0.9185455. The intercept at year 1900 is denoted by _cons and is 1829.327, which is not meaningful. If we centre the intercept by subtracting 1900, the intercept (β_0) has a natural interpretation of the baseline mortality rate in the first year of the observations.

. gen c_year = year-1900

. regress infant c_year

```
      Source |       SS       df       MS              Number of obs =      11
-------------+------------------------------           F(  1,     9) =   80.27
       Model |  9280.98337     1  9280.98337           Prob > F      =  0.0000
    Residual |  1040.60205     9   115.62245           R-squared     =  0.8992
-------------+------------------------------           Adj R-squared =  0.8880
       Total |  10321.5854    10  1032.15854           Root MSE      =  10.753

------------------------------------------------------------------------------
      infant |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      c_year |  -.9185455   .1025238    -8.96   0.000     -1.15047   -.6866205
       _cons |   84.09091    6.06539    13.86   0.000     70.37004    97.81178
------------------------------------------------------------------------------
```

Test for linear trend

Test $H_0 : \beta_1 = 0$ vs. $H_1 : \beta_1 \neq 0$

The ‘test for linear trend’ is the test of the null hypothesis that the coefficient year=0. The probability is <0.001, so we reject the null hypothesis and state that there is a linear trend.

Are the coefficients meaningful?

-0.9185455 is the ‘slope’—the decline in infant mortality per one year increase

84.09091 is the ‘intercept’—infant mortality rate when the year is zero

The interpretation would be that the infant mortality rate dropped by approximately 0.92 deaths per year [95% CI: -1.15 to -0.687; p<0.001].

This means that in the year 2010 (110 years from the first observation), the infant mortality rate would be predicted to be -16.9 deaths, which is impossible.

$$\text{rate}_i = \text{Intercept} + (\text{Slope} \cdot \text{Year}_i) + \varepsilon_i$$

where $i = 1$ to the number of years being analysed, and ε_i is an error term

```
. display 84.09091 + (-.9185455*110)
-16.949455
```

To display this graphically, we generate the mortality rate predicted by the linear regression model, and graph both the observed and predicted infant mortality rates.

```
. regress infant c_year (output omitted)
. predict exp_infant
```

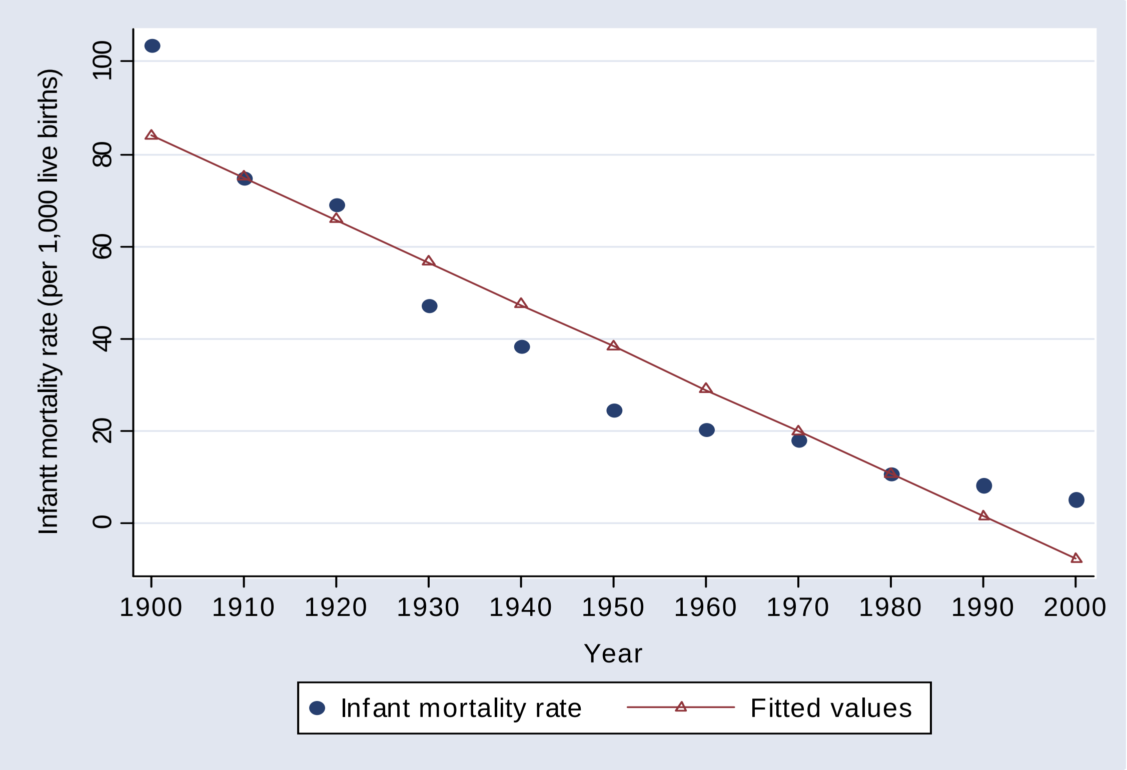

Figure 3.7: Infant mortality rate in Australia, 1900–2000, predicted by linear regression

It is often preferable to model the natural logarithm of the rates.

.gen loginfant = log(infant)

.regress loginfant c_year

```
      Source |       SS       df       MS              Number of obs =      11
-------------+------------------------------           F(  1,     9) =  760.53
       Model |  9.29723867     1  9.29723867           Prob > F      =  0.0000
    Residual |  .110022309     9  .012224701           R-squared     =  0.9883
-------------+------------------------------           Adj R-squared =  0.9870
       Total |  9.40726098    10  .940726098           Root MSE      =  .11057

------------------------------------------------------------------------------
   loginfant |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      c_year |  -.0290724   .0010542   -27.58   0.000    -.0314571   -.0266876
       _cons |    4.71776   .0623673    75.64   0.000     4.576676    4.858845
------------------------------------------------------------------------------
```

The average annual per cent change in infant mortality rates can be calculated from the log linear regression estimates. The estimates must be exponentiated in order to report the results in the usual units. The formula is $100\left[\exp(\beta_1) - 1\right]$

. di 100*(exp(-.0290724)-1)

-2.8653864

. di 100*(exp(-.0266876)-1), 100*(exp(-.0314571)-1)

-2.6334633 -3.0967473

The average annual per cent change is -2.9% [95% CI: -2.6% to -3.1%]

To predict the infant mortality rate in 2010:

$\ln(\text{rate}_i) = \text{Intercept} + (\text{Slope} \cdot \text{Year}_i) + \varepsilon_i$

where $i = 1$ to the number of years being analysed, and ε_i is an error term

. display 4.71776 + (-.0290724*110)

1.519796

Then exponentiate the regression estimate:

. display exp(1.519796)

4.5712926

The infant mortality rate is predicted to be 4.6 per 1,000 live births in the year 2010. When modelling a natural logarithm of the infant mortality rates, an infant mortality rate of zero will never be predicted, which is a more realistic result.

Generate the mortality rate predicted by the log linear regression model.

```
. regress loginfant c_year (output omitted)
. predict logexp_infant
```

Then exponentiate the regression estimates in order to report the results in the usual units.

```
. gen exp_infant = exp(logexp_infant)
```

The observed infant mortality rates and the rates predicted by the log linear regression model are graphed below.

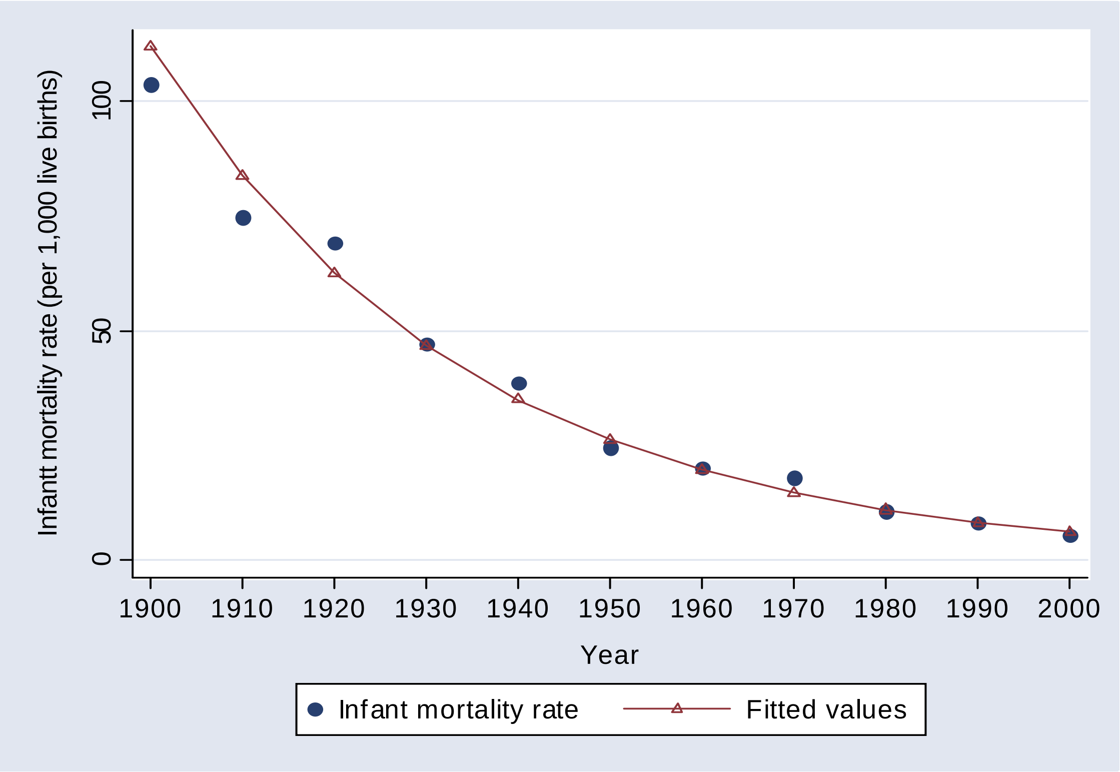

Figure 3.8: Infant mortality rate in Australia, 1900–2000, predicted by log linear regression

There may be different non-linear models (e.g. quadratic) that better fit the data.

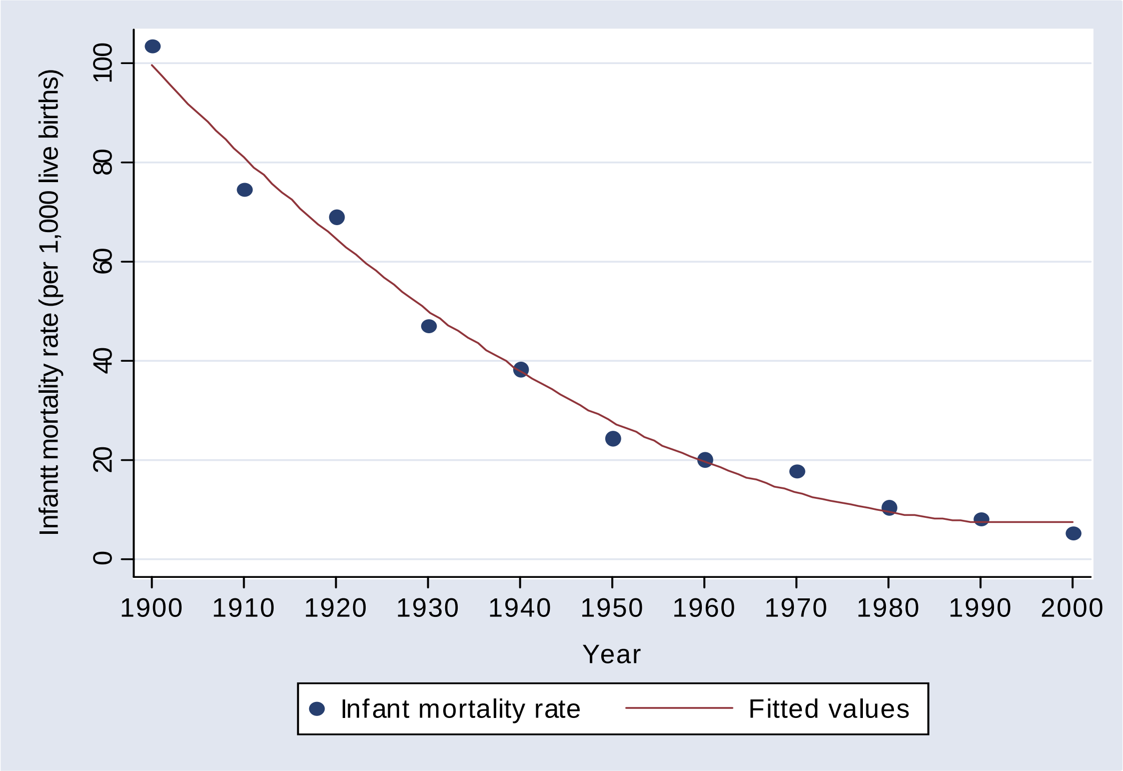

Figure 3.9: Infant mortality rate in Australia, 1900–2000, quadratic prediction

To test whether it is quadratic, first generate the parameters for the quadratic model below:

$$\text{rate}_1 = \text{Intercept} + (\text{Slope}_{\text{year}} \cdot \text{Year}_i) + (\text{Slope}_{\text{year2}} \cdot \text{Year}_1) + \varepsilon_i$$

where $i = 1$ to the number of years being analysed, and ε_i is an error term

. gen c_year2 = c_year^2 (generates the variable year_i^2)

. regress infant c_year c_year2

```
      Source |       SS       df       MS              Number of obs =      11
-------------+------------------------------           F(  2,     8) =  331.09
       Model |  10198.3771     2  5099.18853           Prob > F      =  0.0000
    Residual |  123.208359     8  15.4010449           R-squared     =  0.9881
-------------+------------------------------           Adj R-squared =  0.9851
       Total |  10321.5854    10  1032.15854           Root MSE      =  3.9244

------------------------------------------------------------------------------
      infant |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      c_year |  -1.952578   .1391044   -14.04   0.000    -2.273353   -1.631803
     c_year2 |   .0103403   .0013398     7.72   0.000     .0072508    .0134298
       _cons |    99.6014   2.989827    33.31   0.000     92.70684     106.496
------------------------------------------------------------------------------
```

To predict the infant mortality rate for 2010:

```
. display 99.6014 + (-1.952578*110) + (.0103403*110^2)
9.93545
```

The infant mortality rate is predicted to be 9.9 per 1,000 live births in the year 2010.

Test for non-linear trend

$$\text{Test } H_0 : \beta_2 = 0 \text{ vs. } H_1 : \beta_2 \neq 0$$

The test for non-linear trend is the test of the null hypothesis that the coefficient 'c_year2' is=0. Here the probability is <0.001, so we reject the null hypothesis and state that there is a non-linear trend between 'c_year2' and infant mortality.

3.5.1 Limitations of ordinary least squares regression

One of the main assumptions for OLS regression is homogeneity of variance of the residuals.[8] When plotting rates for a defined time period, the regression procedure does not have access to information on the population sizes that gave rise to the rates at each time point.[7] When modelling 10 years of mortality rate data, the sample size equals 10 observations regardless of whether the underlying population denominator for each rate is in the 1,000s or 100,000s and varies in size from year to year. This means that OLS regression accounts for the variance across the time points plotted, but cannot account for the variability or random error in each individual rate.[7] Therefore the assumption of homogeneity of variance does not hold, and you should not do a simple linear regression through the plotted rates. A better alternative as a test for trend is variance-weighted least squares regression, as unlike OLS regression, homogeneity of variance is not assumed. [3, 11] In Stata, the command to use is 'vwls'.

3.6 Variance-weighted Least Squares regression

Variance-weighted least squares regression is also a linear model, but differs from OLS regression by weighting the observations to account for non-homogeneity of variance.[11] Each observation is weighted proportional to the reciprocal of the variance of the dependent variable at each value of the independent variable. In other words, the weights are $1/\mathrm{Var}(y \mid x)$.[3] Observations at values of the independent variable with large variances are down-weighted and observations with small variances are up-weighted prior to the regression being performed.[3, 11] This weighting decreases the heterogeneity of the spread of residuals and results in a better fit for the linear model.

Some data sets have observations for individuals (e.g. cohort studies) and provide a large enough sample size for each value of the independent variable to allow standard deviations for each value to be calculated internal to the data in Stata. The conditional standard deviations are calculated prior to the regression using the groups defined by the independent variable.[11]

In contrast, mortality rates are calculated from grouped observations (number of deaths in an age- and sex-specific strata) and population sizes that gave rise to each observation are not included in the data set. This prevents calculation of standard deviations for each value of the independent variable (e.g. year) internal to the data. This means that for mortality rates (where there is only one observation per population group) an estimate of the conditional standard deviation (which will vary observation by observation) must be provided as an extra variable in the data set. The process of age-standardisation generates the variance for each mortality rate (see Section 2). The standard deviation is simply the square root of the variance, so it can easily be calculated.[8]

As an example of variance-weighted least squares regression, we will look at injury mortality rates in Australia (1997–2000) by Indigenous status and remoteness.[12] The injury rates are directly age-adjusted using the total population of Australia in 2001 and the remoteness zones are determined according to the Australian Standard Geographical Classification (ASGC). The data is for South Australia, Western Australia, the Northern Territory and Queensland only, as these jurisdictions are considered to have the most reliable identification of Indigenous status. Some type of Poisson analysis might be more appropriate if this data were in numerator (count) and denominator (population) format. If the data is only available as directly adjusted rates (as presented here) then variance-weighted least squares regression is appropriate.

ASGC zone 1=cites, 2=inner regional, 3=outer regional, 4=remote, 5=very remote	Indigenous 0=Other, 1=Indigenous	Age-adjusted rate (per 100,000)	Standard deviation
1	0	34.6	0.5
2	0	43.0	1.0
3	0	46.2	1.1
4	0	49.9	2.4
5	0	47.0	3.3
1	1	70.0	6.5
2	1	52.3	7.6
3	1	102.6	7.7
4	1	150.7	12.7
5	1	141.3	8.0

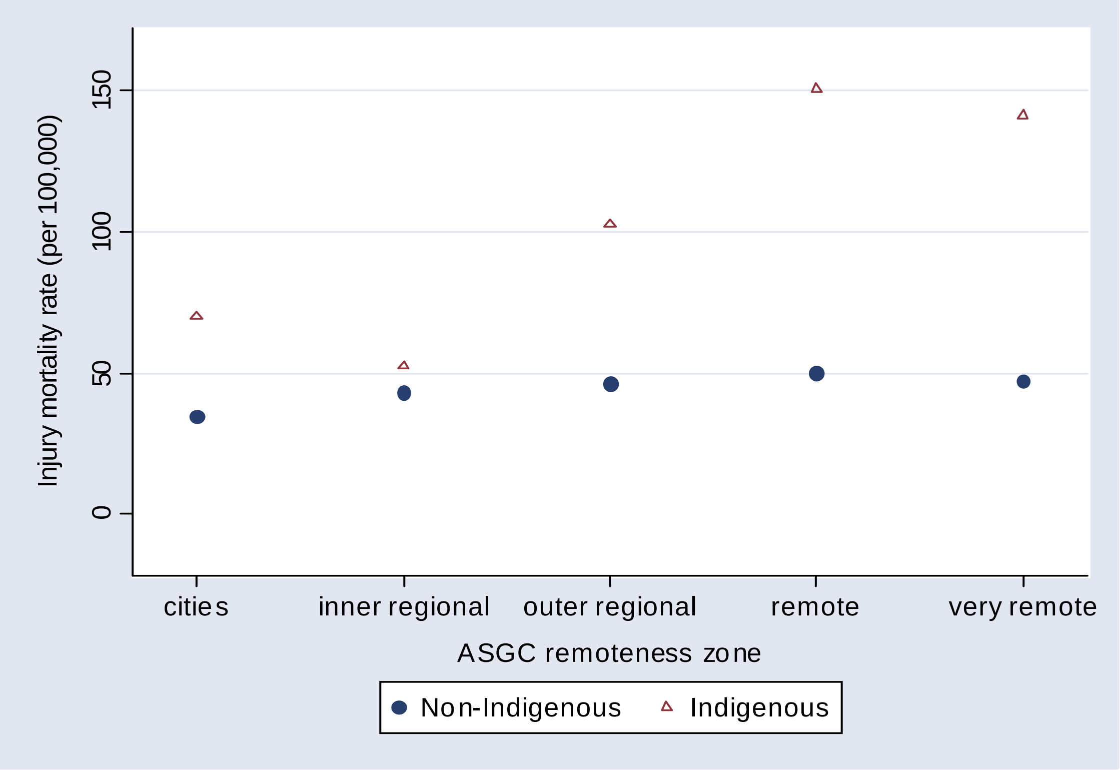

Figure 3.10: Age-adjusted injury mortality rate in Australia by remoteness zone

The graph shows a much greater increase in injury mortality for Indigenous Australians than for Non-Indigenous Australians as the ASGC zones become increasingly more remote. The tabulated data shows that there is a much greater variance for Indigenous Australians compared to Non-Indigenous Australians for age-adjusted injury mortality rates. Indigenous Australians comprise only about 2% of the Australian population, and therefore their small numbers would mean greater variance is to be expected.

To illustrate why simple linear regression is inappropriate for looking at trends over time, compare the results from simple linear regression with variance-weighted least squares regression.

Non-Indigenous age-adjusted mortality by remoteness zone

Simple linear regression

. regress adjrate ASGC if ind==0

```
      Source |       SS       df       MS              Number of obs =       5
-------------+------------------------------           F(  1,     3) =    8.22
       Model |   100.48678     1   100.48678           Prob > F      =  0.0642
    Residual |  36.6613667     3  12.2204556           R-squared     =  0.7327
-------------+------------------------------           Adj R-squared =  0.6436
       Total |  137.148147     4  34.2870366           Root MSE      =  3.4958

------------------------------------------------------------------------------
     adjrate |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        ASGC |   3.169965   1.105462     2.87   0.064    -.3481075    6.688037
       _cons |   34.62889   3.666402     9.44   0.003     22.96076    46.29701
------------------------------------------------------------------------------
```

The slope is positive (3.169965) but the confidence intervals range from -.3481075 to 6.688037 and death rates cannot be negative in the real world, but the model would generate negative death rates with its CI limits.

Variance-weighted least squares regression

. vwls adjrate ASGC if ind==0, sd(stdev)

```
Variance-weighted least-squares regression           Number of obs   =       5
Goodness-of-fit chi2(3)    =   16.22                 Model chi2(1)   =  150.35
Prob > chi2                =  0.0010                 Prob > chi2     =  0.0000
------------------------------------------------------------------------------
     adjrate |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        ASGC |   5.225352    .426147    12.26   0.000     4.390119    6.060585
       _cons |   29.89375   .7594632    39.36   0.000     28.40523    31.38227
------------------------------------------------------------------------------
```

The slope is positive (5.225352) but the confidence intervals do not cross zero (4.390119 to 6.060585). There is a significant increase in non-Indigenous mortality by remoteness. The ‘Goodness-of-fit’ chi-squared test that ‘vwls’ produces is a test for the adequacy of the linear model. The large value of this statistic (and the small p-value corresponding to it) strongly suggest that the linear model doesn't fit the data well.

Indigenous age-adjusted mortality by remoteness zone

Simple linear regression

. regress adjrate ASGC if ind==1

```
      Source |       SS       df       MS              Number of obs =       5
-------------+------------------------------           F(  1,     3) =   10.97
       Model |  5814.58059     1  5814.58059           Prob > F      =  0.0453
    Residual |   1590.8369     3  530.278967           R-squared     =  0.7852
-------------+------------------------------           Adj R-squared =  0.7136
       Total |  7405.41749     4  1851.35437           Root MSE      =  23.028

------------------------------------------------------------------------------
     adjrate |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        ASGC |   24.11344   7.282026     3.31   0.045      .938786     47.2881
       _cons |   31.04069   24.15175     1.29   0.289    -45.82094    107.9023
------------------------------------------------------------------------------
```

The slope is positive (24.11344) and just reaches significance.

Variance-weighted least squares regression

. vwls adjrate ASGC if ind==1, sd(stdev)

```
Variance-weighted least-squares regression           Number of obs   =       5
Goodness-of-fit chi2(3)    =   21.30                 Model chi2(1)   =   78.91
Prob > chi2                =  0.0001                 Prob > chi2     =  0.0000
------------------------------------------------------------------------------
     adjrate |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        ASGC |   21.42015   2.411285     8.88   0.000     16.69412    26.14618
       _cons |   36.32533   7.328247     4.96   0.000     21.96223    50.68843
------------------------------------------------------------------------------
```

The slope is positive (21.42015) and the confidence intervals are much narrower (16.69412 to 26.14618). There is a significant positive increase in Indigenous mortality by remoteness, and it is greater than for non-Indigenous Australians. The 'goodness-of-fit' chi-squared test is significant, suggesting the linear model is not a good fit for the data.

3.7 Logistic regression

Logistic regression is used when the dependent variable is binary or dichotomous so that y is a 0/1 variable (0 is 'unexposed', 1 is 'exposed').[13] The independent predictor variables can be binary, categorical (more than two categories) or continuous.[4] Some examples where logistic regression is appropriate are case-control studies and analysis of data where there is a yes/no outcome (e.g. death, coronary heart disease etc). If we use linear regression with binary outcomes data we encounter several problems: 1) the model can give predicted values that exceed the bounds of 0 and 1, and 2) the model assumes a normal distribution of residuals, when the errors actually follow a binomial distribution. Logistic regression solves these problems by transforming the dependent variable so that the assumptions of linearity, normality and homogeneity of variance are better met. Logistic regression uses a maximum likelihood estimation procedure rather than the least square estimation procedure used in simple linear regression. [13] As an aside, in matched case-control studies each case is matched directly with one or more controls on some factor (e.g. age, socioeconomic status) and this requires a particular analysis known as conditional logistic regression, not described in this report.

Consider the following fictitious data of a case-control study into drunk driving (blood alcohol concentration greater than 0.05) and the probability of death when injured in a car accident. A case (1) is someone who died in a car accident and a control (0) is someone in a car accident who did not die. For driving while under the influence of alcohol, the number of exposed cases was 50, exposed controls 260, non-exposed cases 140 and non-exposed controls 9,570.

Case (Dead)	Exposure (BAC > 0.05)	Population
1	1	50
1	0	140
0	1	260
0	0	9,570

With this grouped data, the analysis can be carried out in Stata using the following command:

```
. logistic case exposed [fw=pop]
```

Otherwise, to convert the data to individual observations, use the following:

```
. expand pop
(10016 observations created)
```

.logistic case exposed

```
Logistic regression                               Number of obs   =      10020
                                                  LR chi2(1)      =     144.31
                                                  Prob > chi2     =     0.0000
Log likelihood = -869.44256                       Pseudo R2       =     0.0766

------------------------------------------------------------------------------
        case | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     exposed |    13.1456    2.31789    14.61   0.000       9.3045     18.5724
------------------------------------------------------------------------------
```

This output is interpreted as the odds of exposure to alcohol (BAC > 0.05) among cases is 13 times higher than among controls, and is statistically significant at p<0.001.

To obtain the coefficients of the logistic regression (intercept and slope), the logit command can be used. Logit is defined as the log base e (log) of the odds.

. logit case exposed

```
Iteration 0:   log likelihood = -941.59682
Iteration 1:   log likelihood = -937.03099
Iteration 2:   log likelihood =  -871.3212
Iteration 3:   log likelihood = -869.45519
Iteration 4:   log likelihood = -869.44256

Logit estimates                                   Number of obs   =      10020
                                                  LR chi2(1)      =     144.31
                                                  Prob > chi2     =     0.0000
Log likelihood = -869.44256                       Pseudo R2       =     0.0766

------------------------------------------------------------------------------
        case |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     exposed |   2.576087   .1763243    14.61   0.000     2.230498    2.921677
       _cons |  -4.224746   .0851314   -49.63   0.000      -4.3916   -4.057892
------------------------------------------------------------------------------
```

The results are interpreted just like linear regression. If there is no relationship between blood alcohol concentration and the probability of death, the line will be flat (slope=0), otherwise there will be a non-zero slope if there is a relationship. There is an increase in the log of the odds of death when the driver is exposed to alcohol (p<0.001). The slope estimate is interpreted as the log odds ratio—the odds ratio can be computed by raising e to the power of the logistic coefficient [exp(slope) = OR].

. display exp(2.576087) = 13.1456

[same result as logistic]

Logistic regression can work for continuous predictors as well. For example, instead of coding exposure to alcohol (blood alcohol concentration >0.05) as 0/1, the actual blood alcohol levels could be included (e.g. 0.00, 0.05, 0.10) to provide a more informative model.

Logistic regression can be used to monitor trends over time. Drivers in car accidents who are admitted to hospital or die at the scene are routinely tested for blood alcohol concentration in Australia. The fictitious data below is a case-control study of alcohol involvement in car accidents for three consecutive years.

Year	Case (Dead)	Exposure (BAC > 0.05)	Population
2001	1	1	50
	1	0	140
	0	1	260
	0	0	9,570
2002	1	1	79
	1	0	173
	0	1	275
	0	0	9,563
2003	1	1	112
	1	0	166
	0	1	314
	0	0	9,508

If year is categorised as 1=2001, 2=2002 and 3=2003, then Stata will assign the lowest category as the reference category (in this case 2001). The command in Stata that permits categories is the 'xi:' command and an 'i.' is placed preceding the variable to be categorised, in this case, year.

. xi: logistic case exposed i.year [fw=pop]

```
i.year            _Iyear_1-3          (naturally coded; _Iyear_1 omitted)

Logistic regression                                Number of obs   =      30210
                                                   LR chi2(3)      =     777.56
                                                   Prob > chi2     =     0.0000
Log likelihood = -3012.9788                        Pseudo R2       =     0.1143

------------------------------------------------------------------------------
        case | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     exposed |   16.77956   1.450959    32.61   0.000     14.16367    19.87859
    _Iyear_2 |   1.294977   .1293915     2.59   0.010     1.064662    1.575116
    _Iyear_3 |   1.351432   .1327056     3.07   0.002     1.114832    1.638246
------------------------------------------------------------------------------
```

This output is interpreted as the odds of exposure to alcohol (BAC > 0.05) among cases is over 16 times higher than among controls, and is statistically significant at p<0.001. There is an increase in the odds of death over time (statistically significant at p<0.05). Compared to 2001, the odds of death increased by 29% in 2002 and by 35% in 2003.

If we look at those whose blood alcohol concentration was less than 0.05, there is no statistically significant increase in the odds of death from a car accident over the three years.

. xi: logistic case i.year if exposed==0 [fw=pop]

```
i.year            _Iyear_1-3         (naturally coded; _Iyear_1 omitted)

Logistic regression                                 Number of obs   =      29120
                                                    LR chi2(2)      =       3.90
                                                    Prob > chi2     =     0.1421
Log likelihood = -2440.5705                         Pseudo R2       =     0.0008

------------------------------------------------------------------------------
        case | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    _Iyear_2 |   1.236619   .1417117     1.85   0.064     .9878511    1.548033
    _Iyear_3 |   1.193446   .1380308     1.53   0.126     .9513824    1.497099
------------------------------------------------------------------------------
```

Whereas for those with a blood alcohol concentration greater than 0.05, the odds of death significantly increased over time.

. xi: logistic case i.year if exposed==1 [fw=pop]

```
i.year            _Iyear_1-3         (naturally coded; _Iyear_1 omitted)

Logistic regression                                 Number of obs   =       1090
                                                    LR chi2(2)      =      11.09
                                                    Prob > chi2     =     0.0039
Log likelihood =  -570.3016                         Pseudo R2       =     0.0096

------------------------------------------------------------------------------
        case | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    _Iyear_2 |   1.493818   .2992888     2.00   0.045     1.008691    2.212266
    _Iyear_3 |   1.854777   .3517209     3.26   0.001     1.279024    2.689706
------------------------------------------------------------------------------
```

Drivers who drink and drive when over the limit are clearly at greater risk of death if injured in a car accident than those who abstain. The three years of data (if not fictitious) would show an increasing trend in deaths from drink-driving and would provide a warning that prevention strategies had not been effective.

3.8 Poisson regression

Count variables indicate how many times something has happened and examples include number of deaths, hospitalisations, injuries, number of bombs dropped during a war, to name just a few. Poisson regression is an extension of logistic regression and is useful when the risk of an event to an individual is small, but there are a large number of individuals.[4] Count data often follow a Poisson distribution—when the population size n is large, the probability of an individual event π is small, but the expected number of events, $n\pi$, is moderate (say five or more).[14]

If we let $\lambda = n\pi$ (λ is the rate of occurrence or the expected number of times an event will occur over a given period of time) and y be a random variable indicating the number of times an event did occur, and let $n \to \infty$, we would find the relationship between the expected count (λ) and the probability of observing any observed count (y) is specified by the Poisson distribution.

$$\Pr(y) = \frac{\lambda^y}{y!} \cdot e^{-\lambda} \qquad \text{for } y = 0, 1, 2, 3......$$

The Poisson distribution can be visualised in Figure 3.2 (recreated from an example in Long et al[9]).

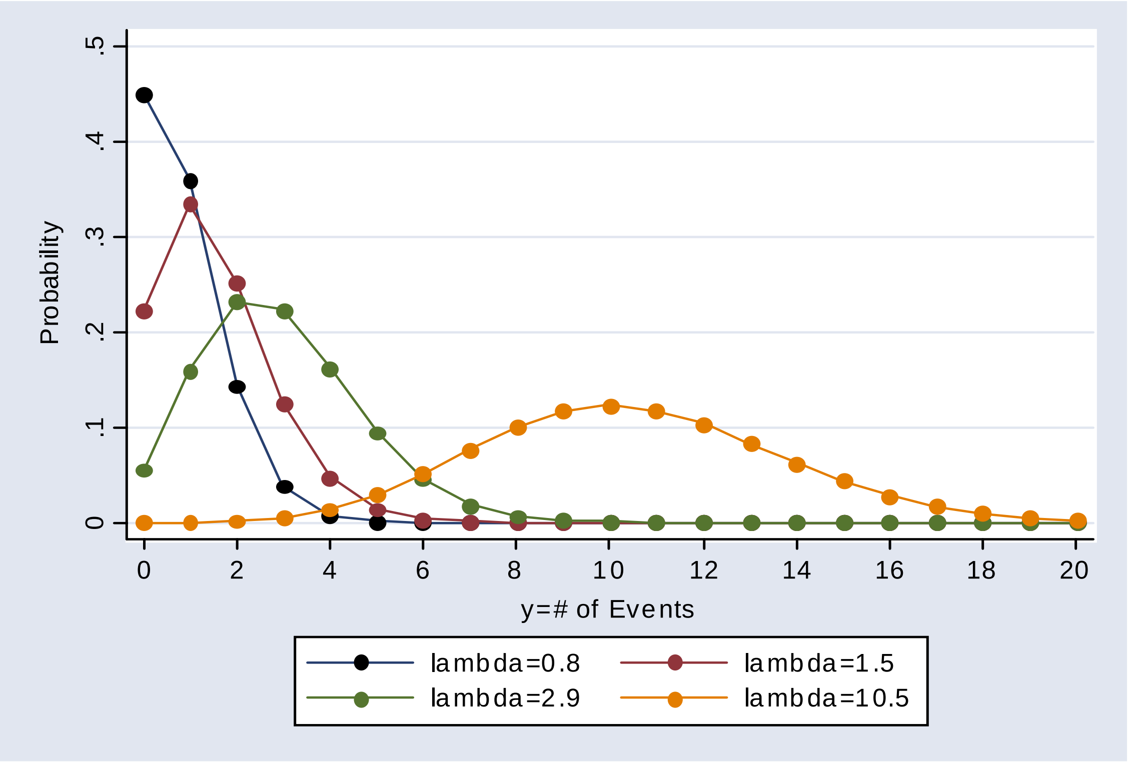

Figure 3.11: A plot of the predicted probabilities for the rate parameter λ

The mean of the distribution is equal to its variance and both are given by λ (as λ increases, the distribution shifts to the right).[9] This means the Poisson distribution has a variance which is equal to the mean and a standard deviation equal to the square root of the mean. As λ increases, the probability of a zero count decreases and the Poisson distribution approximates a normal distribution (shown by the distribution for λ =10.5).[9]

The Poisson regression model extends the Poisson distribution by allowing each observation to have a different value of λ .[9] Poisson regression involves the regression of a count as the outcome on one or more predictor variables. A log transformation is used to adjust for skewness and prevents the Poisson regression model from producing negative predicted values. Poisson regression also models the variance as a function of the mean.[9]

OLS regression has often been applied to count outcomes but this can result in inefficient and biased estimates.[9] Poisson distributions have three characteristics that make OLS regression problematic:

1. A skewed distribution (OLS assumes a symmetric distribution of errors)
2. A non-negative distribution (OLS can produce predicted negative values)
3. The variance of the distribution increases as the mean increases (OLS assumes a constant variance)

It is much better to use models which are designed to deal with count outcomes, such as Poisson regression.[9] Poisson regression has the advantage over OLS regression of accounting for variance across the time points plotted and the variability at each time point.[7] The Poisson regression models the counts in the numerator and denominator for each time period rather than the pre-calculated rates. This means the confidence intervals will differ for sets of the same rates when the populations they arise from are of different sizes.[7]

Poisson regression is useful for analysis of studies when the objective is to relate rates of injury or disease (counts of events divided by person-years or persons-at-risk) to predictor variables such as age, gender, socioeconomic status, exposure to substances (e.g. alcohol , tobacco) or other covariates and confounders.

The Poisson distribution assumes that events are independent and individuals have the same risk of experiencing an event over time. In real data, many count variables have a variance greater than the mean ('overdispersion'). This means that the observed data vary by more than would be expected by a Poisson distribution. If over-dispersion occurs, ignoring it will result in underestimating the standard errors of the regression parameter estimates, which may lead to incorrect conclusions. It may arise because an important covariate is omitted from the model.[4] Another common explanation is when the counts display autocorrelation (see Section 1.4). For example, events can recur in some individuals and some individuals may be at greater risk of an event than others. Data may be overdispersed if there are counts within an individual, such as the number of falls or asthma attacks a year, rather than counts within groups of separate individuals. [4] In the case of overdispersion, the Negative Binomial Distribution model is a better alternative to the Poisson model because it has a variance which is larger than the mean.[9] For many count variables, there are more observed zeros than predicted by the Poisson distribution. There are models that incorporate Poisson probabilities but allow the probability of a zero to be a bit larger or a lot larger than is expected according to a Poisson distribution. These are called Zero Inflated Poisson (ZIP) models.[9]

As an example of Poisson regression, we will look at deaths due to unintentional motor vehicle traffic injury in Australia (1990–2000).[15] The data is available as counts (numerator) and population numbers (denominator) so it is suitable for Poisson regression analysis.

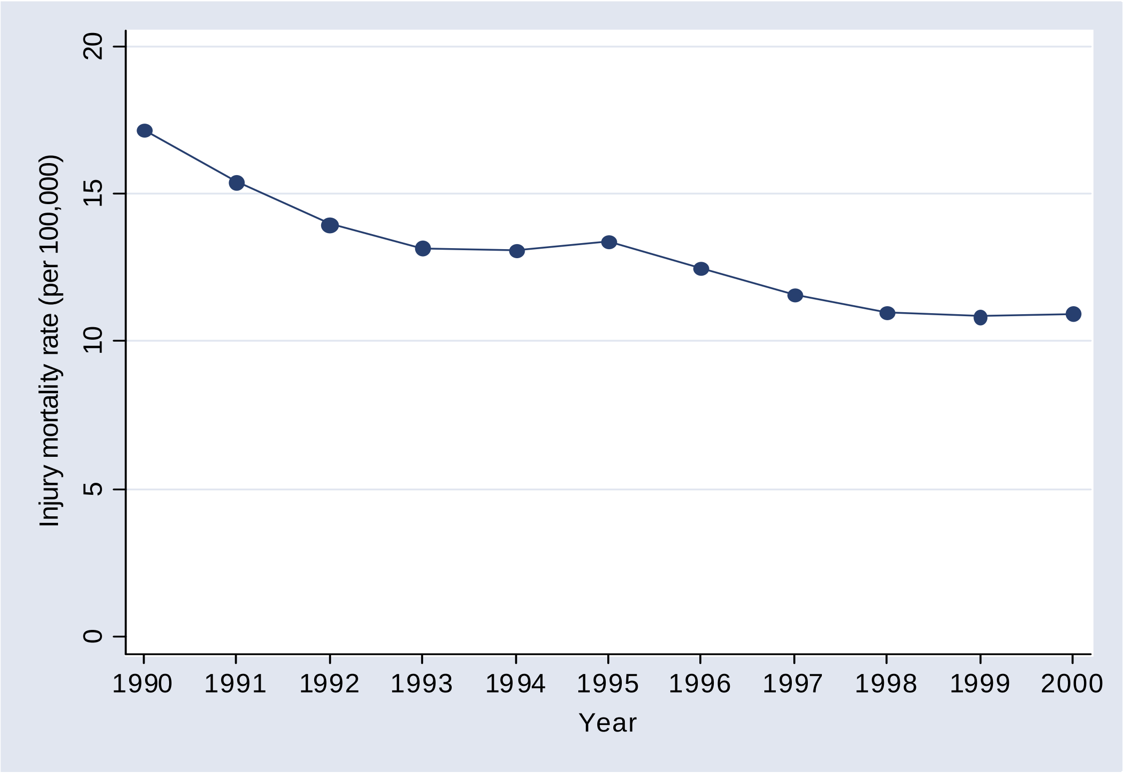

Figure 3.12: Motor vehicle injury mortality rate in Australia, 1990–2000

The graph shows there has been a decline in fatalities from motor vehicle traffic accidents over the decade.

We might be interested to see what effect predictors such as age and gender have on the mortality rates from unintentional motor vehicle accidents.

To do these analyses, the data is entered into Stata using the variables gender: 0 = males and 1=females, age-group: 1 = 15–39 years, 2 = 40–59 years, and 3 = 60+ years. There are 66 rows in total for males and females. Mid-year age- and sex-specific Australian populations are taken to be reasonable estimates of years lived by persons who could have become cases.

Year	Gender	Age-group	Deaths	Person-years
1990	0	1	1,054	3,477,224
1991	0	1	958	3,486,637
1992	0	1	861	3,497,805
1993	0	1	858	3,495,650
1994	0	1	800	3,497,722
1995	0	1	842	3,508,778
1996	0	1	811	3,522,090
1997	0	1	741	3,517,531
1998	0	1	705	3,510,959
1999	0	1	710	3,509,163
2000	0	1	741	3,512,056
1990	0	2	286	1,931,130
1991	0	2	251	1,982,700
1992	0	2	237	2,035,597
1993	0	2	237	2,088,547
1994	0	2	247	2,144,216
1995	0	2	230	2,203,838
1996	0	2	261	2,267,750
1997	0	2	215	2,329,313
1998	0	2	260	2,390,771
1999	0	2	224	2,451,431
2000	0	2	248	2,511,606
1990	0	3	286	1,176,371
1991	0	3	261	1,203,041
1992	0	3	209	1,224,682
1993	0	3	212	1,245,319
1994	0	3	253	1,266,603
1995	0	3	246	1,287,604
1996	0	3	232	1,313,126
1997	0	3	211	1,343,707
1998	0	3	194	1,375,510
1999	0	3	225	1,410,940
2000	0	3	203	1,448,498

Year	Gender	Age-group	Deaths	Person-years
1990	1	1	319	3,412,518
1991	1	1	319	3,427,986
1992	1	1	294	3,442,406
1993	1	1	267	3,441,543
1994	1	1	249	3,443,263
1995	1	1	272	3,453,589
1996	1	1	225	3,472,418
1997	1	1	257	3,477,568
1998	1	1	217	3,475,965
1999	1	1	213	3,479,449
2000	1	1	236	3,487,036
1990	1	2	120	1,857,413
1991	1	2	99	1,913,600
1992	1	2	120	1,970,900
1993	1	2	84	2,029,839
1994	1	2	108	2,091,715
1995	1	2	126	2,155,731
1996	1	2	107	2,223,602
1997	1	2	114	2,292,738
1998	1	2	96	2,362,010
1999	1	2	98	2,430,281
2000	1	2	98	2,499,645
1990	1	3	221	1,455,478
1991	1	3	191	1,484,542
1992	1	3	191	1,506,937
1993	1	3	164	1,528,669
1994	1	3	175	1,551,182
1995	1	3	180	1,574,100
1996	1	3	163	1,600,413
1997	1	3	152	1,630,212
1998	1	3	152	1,660,179
1999	1	3	154	1,693,719
2000	1	3	139	1,728,472

An an aside, the 66 rows of the dataset can be condensed to 33 rows using the ‘collapse’ command for the purpose of summarising data for all persons for each age group.

. collapse (sum) gender deaths personyears, by (year agegroup)

Likewise, the 66 rows can be condensed to 11 rows to summarise data for all persons with all age groups combined. The ‘collapse’ function is helpful for formatting data to produce summary graphs and summary statistics for trend.

. collapse (sum) gender agegroup deaths personyears, by (year)

Return to the original data.

.clear

[and reopen the original saved data file of 66 rows]

To generate a column that displays the crude (unadjusted) death rates by age, gender and year, use:

. gen rate = (deaths/personyears)*100000

What is the overall mortality rate for 1990–2000? If no independent variables are included, a model with only an intercept is fitted, which corresponds to fitting a univariate Poisson distribution (example below).

. poisson deaths, exposure(personyears)

```
Iteration 0:   log likelihood =  -3261.209
Iteration 1:   log likelihood =  -3261.209

Poisson regression                                Number of obs   =         66
                                                  LR chi2(0)      =      -0.00
                                                  Prob > chi2     =          .
Log likelihood =  -3261.209                       Pseudo R2       =    -0.0000

------------------------------------------------------------------------------
      deaths |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |  -8.953202   .0070309 -1273.40   0.000    -8.966983   -8.939422
 personyears | (exposure)
------------------------------------------------------------------------------
```

The regression coefficient is -8.953202 and it is the log of the crude rate. By exponentiating this coefficient and the CI coefficients, we can obtain the crude rate and its 95% CIs.

. display exp(-8.953202)*100000

12.932241

. display exp(-8.966983)*100000, exp(-8.939422)*10000012.755244

12.755244 13.11168

[lower CI, upper CI]

There is a crude death rate of 12.9 deaths [95% CI: 12.8 to 13.1] per 100,000 persons per year.

To estimate trends over time, we can model the rate of vehicle injury mortality as a function of calendar year using Poisson regression; the basic model being $\log(\text{rate}) = \beta_0 + \beta_1(\text{year} - 1990)$. Calendar year can be entered as a continuous variable, scaled by subtracting the initial year (1990) so that the intercept can be interpreted as the log of the baseline mortality rate in the first year (1990).

. gen c_year = year–1990

. poisson deaths c_year, exposure(personyears)

```
Iteration 0:   log likelihood = -3072.4532
Iteration 1:   log likelihood = -3072.4532

Poisson regression                                Number of obs   =         66
                                                  LR chi2(1)      =     377.51
                                                  Prob > chi2     =     0.0000
Log likelihood = -3072.4532                       Pseudo R2       =     0.0579

------------------------------------------------------------------------------
      deaths |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      c_year |  -.0432333   .0022289   -19.40   0.000    -.0476019   -.0388648
       _cons |  -8.740711    .012615  -692.88   0.000    -8.765436   -8.715987
 personyears | (exposure)
------------------------------------------------------------------------------
```

Estimated trends in the mortality rate can be reported as annual per cent change, obtained from the fit of the Poisson regression model as $100\left[\exp(\beta_1) - 1\right]$.

. di 100*(exp(-.0432333)-1)

-4.2312065

Confidence intervals (95%) for the annual per cent change can be calculated.

. di 100*(exp(-.0388648)-1), 100*(exp(-.0476019)-1)

-3.8119253 -4.6486695

[lower CI, upper CI]

The average annual per cent change in mortality rate from 1990–2000 (unadjusted for age or gender) is -4.2% [95% CI: -3.8% to -4.6%].

However, it is likely that there are differences between males and females and among age groups in the mortality rate for vehicle injury mortality.

What is the overall mortality rate for males and females for 1990–2000?

. bysort gender: poisson deaths, exposure(personyears)

```
-> gender = 0

Iteration 0:   log likelihood = -929.99854
Iteration 1:   log likelihood = -929.99854

Poisson regression                                Number of obs   =         33
                                                  LR chi2(0)      =       0.00
                                                  Prob > chi2     =          .
Log likelihood = -929.99854                       Pseudo R2       =     0.0000

------------------------------------------------------------------------------
      deaths |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |   -8.59285   .0083598 -1027.88   0.000    -8.609235   -8.576465
 personyears | (exposure)
------------------------------------------------------------------------------
```

. display exp(-8.59285)*100000

18.542687

For males, there is a crude death rate of 18.5 deaths per 100,000 persons per year.

```
-> gender = 1

Iteration 0:   log likelihood = -424.35729
Iteration 1:   log likelihood = -424.35729

Poisson regression                                Number of obs   =         33
                                                  LR chi2(0)      =       0.00
                                                  Prob > chi2     =          .
Log likelihood = -424.35729                       Pseudo R2       =     0.0000

------------------------------------------------------------------------------
      deaths |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |  -9.502091   .0129969  -731.11   0.000    -9.527564   -9.476617
 personyears | (exposure)
------------------------------------------------------------------------------
```

. display exp(-9.502091)*100000

7.4695478

For females, there is a crude death rate of 7.5 deaths per 100,000 persons per year.

The rate ratio is:

. di exp(-9.502091)/ exp(-8.59285)

.40282986

The ‘ir’ option after the ‘poisson’ command reports incident rate ratios rather than the default regression coefficients. The incident rate ratios are simply the regression coefficients exponentiated, and give a measure of the relative risk.

. xi: poisson deaths gender, exp(personyears) ir

```
Iteration 0:   log likelihood = -1354.3559
Iteration 1:   log likelihood = -1354.3558

Poisson regression                                Number of obs   =         66
                                                  LR chi2(1)      =    3813.71
                                                  Prob > chi2     =     0.0000
Log likelihood = -1354.3558                       Pseudo R2       =     0.5847

------------------------------------------------------------------------------
      deaths |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      gender |   .4028301   .0062251   -58.84   0.000     .3908121    .4152176
 personyears | (exposure)
------------------------------------------------------------------------------
```

The female crude death rate is estimated to be about 40% of the rate for males (not adjusted for age). The difference in crude rates by gender suggests that we would want to adjust for sex when looking at mortality rates from motor vehicle accidents over time.

The incidence rate ratio value of 0.40 is similar to the Standardised Mortality Ratio (SMR) (see Chapter 2) calculated internal to the data (using males as the standard population):

$$\mathrm{SMR} = \frac{\mathrm{D_i}}{\sum \mathrm{r_{si}P_i}} = \frac{319+319+294+267+\ldots+139\cdot 100000}{30.3(3412518)+27.5(3427986)+24.6(3442406)+\ldots+14.0(1728472)}$$

$$\mathrm{SMR} = 0.40326989$$

The rate ratios from the Poisson model and the SMR calculated directly from the data will not in general be identical but when the observed rate ratios are similar among all age strata, the two approaches give similar estimates.[16]

It is also important to see what effect age has on the risk of mortality from motor vehicle accidents for males and females.

The command in Stata that permits categories is the ‘xi:’ command and an ‘i.’ is placed preceding the variable to be categorised. Age is categorised as 1=15–39 years, 2=40–59 years, and 3=60+ years. By default, Stata assigns the lowest number as the reference category (in this case, the youngest age group, 15–39 years). Motor vehicle accidents in the reference category (15–39 years) are allocated a relative risk (RR) of 1.

. xi: poisson deaths i.agegroup if gender==0, exposure(personyears) ir

```
i.agegroup        _Iagegroup_1-3      (naturally coded; _Iagegroup_1 omitted)

Iteration 0:   log likelihood = -258.17584
Iteration 1:   log likelihood = -258.15361
Iteration 2:   log likelihood = -258.15361

Poisson regression                                Number of obs   =         33
                                                  LR chi2(2)      =    1343.69
                                                  Prob > chi2     =     0.0000
Log likelihood = -258.15361                       Pseudo R2       =     0.7224

------------------------------------------------------------------------------
      deaths |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
_Iagegroup_2 |   .4700924   .0103104   -34.42   0.000     .4503127     .490741
_Iagegroup_3 |   .7516159   .0168915   -12.71   0.000     .7192277    .7854627
 personyears | (exposure)
------------------------------------------------------------------------------
```

Compared to males aged 15–39 years, the death rates due to motor vehicle accidents in the 40–59 year age-group (RR=0.47) and in the 60+ year age-group (RR=0.75) are substantially and significantly lower.

. xi: poisson deaths i.agegroup if gender==1, exposure(personyears) ir

```
i.agegroup        _Iagegroup_1-3      (naturally coded; _Iagegroup_1 omitted)

Iteration 0:   log likelihood = -191.27533
Iteration 1:   log likelihood = -191.27041
Iteration 2:   log likelihood = -191.27041

Poisson regression                                Number of obs   =         33
                                                  LR chi2(2)      =     466.17
                                                  Prob > chi2     =     0.0000
Log likelihood = -191.27041                       Pseudo R2       =     0.5493

------------------------------------------------------------------------------
      deaths |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
_Iagegroup_2 |   .6508326   .0225772   -12.38   0.000     .6080529    .6966221
_Iagegroup_3 |   1.432468   .0424945    12.12   0.000     1.351556    1.518225
 personyears | (exposure)
------------------------------------------------------------------------------
```

For females, we can see there is a significant decline in the relative risk of motor vehicle accidents in the 40–59 year age-group but the risk is higher in the 60+ year age-group compared to the reference category, the 15–39 year age group.

These separate results for males and females suggest that injury mortality should be adjusted for both age and sex. This would require a multivariate Poisson regression model.

To adjust injury mortality rates by age and sex requires the generation of a mortality estimate in which the effects of age groups and gender are averaged to produce a contribution of each to the estimate of the year's mortality. In other words, the mortality estimate is the average over all sexes and age-groups assuming equal weights (numbers of subjects) in each of the age/gender cells.

The mortality estimate averaged over all sexes and ages can be calculated manually using the 'lincom' command once the Poisson model is in memory.

Using the Poisson model adjusted for age and gender:

. xi: poisson deaths c_year i.agegroup i.gender, exposure(personyears)

```
i.agegroup        _Iagegroup_1-3      (naturally coded; _Iagegroup_1 omitted)
i.gender          _Igender_0-1        (naturally coded; _Igender_0 omitted)

Iteration 0:   log likelihood = -439.51166
Iteration 1:   log likelihood = -439.37014
Iteration 2:   log likelihood = -439.37014

Poisson regression                                Number of obs   =         66
                                                  LR chi2(4)      =    5643.68
                                                  Prob > chi2     =     0.0000
Log likelihood = -439.37014                       Pseudo R2       =     0.8653

------------------------------------------------------------------------------
      deaths |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      c_year |   -.040328   .0022304   -18.08   0.000    -.0446994   -.0359566
_Iagegroup_2 |  -.6561679    .018512   -35.45   0.000    -.6924507   -.6198852
_Iagegroup_3 |  -.0614925   .0176318    -3.49   0.000    -.0960501   -.0269348
  _Igender_1 |  -.9148964   .0154665   -59.15   0.000    -.9452103   -.8845826
       _cons |  -8.217637   .0144422  -569.00   0.000    -8.245944   -8.189331
 personyears | (exposure)
------------------------------------------------------------------------------
```

There are three age groups (15–39 years, 40–59 years and 60+ years) so each contributes 1/3 to the effect. We have two sexes, so each contributes 1/2. The age- and sex-adjusted mortality rate for 1990 (i.e. c_year =0) can be calculated as follows:

. lincom (_cons + (0.3333* _Iagegroup_2) + (0.3333* _Iagegroup_3) + (0.5* _Igender_1) + (0* c_year))

```
( 1)  .3333 [deaths]_Iagegroup_2 + .3333 [deaths]_Iagegroup_3 + .5 [deaths]_Igender_1 +
[deaths]_cons = 0

------------------------------------------------------------------------------
      deaths |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |  -8.914282   .0136229  -654.36   0.000    -8.940982   -8.887581
------------------------------------------------------------------------------
```

. di exp(-8.914282)*100000

13.445486

. display exp(-8.940982)*100000, exp(-8.887581)*100000

13.091242 13.80933

[lower CI, upper CI]

The age- and sex- adjusted mortality rate in 1990 (baseline) is 13.4 [95% CI: 13.1 to -13.8] per 100,000 persons.

For 1995, it would be:

. lincom (_cons + (0.3333* _Iagegroup_2) + (0.3333* _Iagegroup_3) + (0.5* _Igender_1) + (5* c_year))

```
( 1)  5 [deaths]c_year + .3333 [deaths]_Iagegroup_2 + .3333 [deaths]_Iagegroup_3 + .5
[deaths]_Igender_1 + [deaths]_cons =
>  0

------------------------------------------------------------------------------
      deaths |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |  -9.115922   .0085614 -1064.77   0.000    -9.132702   -9.099142
------------------------------------------------------------------------------
```

. di exp(-9.115922)*100000

10.990194

. display exp(-9.132702)*100000, exp(-9.099142)*100000

10.807318 11.176166

[lower CI, upper CI]

The age- and sex- adjusted mortality rate in 1995 is 11.0 [95% CI: 10.8 to 11.2] per 100,000 persons.

The estimated age- and sex-adjusted mortality rates and 95% confidence intervals for each year can be obtained as a single table using the ‘adjust’ command once the Poisson model is in memory. The ‘exp ci’ option gives exponentiated linear predictions with 95% confidence intervals. The ‘nooffset’ option is specified because the offset or exposure (i.e. person-years) is not constant—there are different values for each year. The ‘replace’ option specifies that the data in memory are to be replaced with data containing one observation per cell corresponding to the table produced by the ‘adjust’ command (i.e 11 rows).

. quietly xi: poisson deaths c_year i.agegroup i.gender, exposure(personyears)

[quietly suppresses screen output so results are not shown]

. adjust, by(c_year) exp ci nooffset format(%9.8f) replace

```
----------------------------------------------------------------------------------------
     Dependent variable: deaths     Equation: deaths     Command: poisson
   Variables left as is: _Iagegroup_2, _Iagegroup_3, _Igender_1
----------------------------------------------------------------------------------------

-------------------------------------------------
   c_year |     exp(xb)           lb           ub
----------+--------------------------------------
        0 |  0.00013445  [0.00013091  0.00013809]
        1 |  0.00012914  [0.00012614  0.00013220]
        2 |  0.00012403  [0.00012150  0.00012662]
        3 |  0.00011913  [0.00011696  0.00012135]
        4 |  0.00011442  [0.00011249  0.00011639]
        5 |  0.00010990  [0.00010807  0.00011176]
        6 |  0.00010556  [0.00010371  0.00010743]
        7 |  0.00010138  [0.00009943  0.00010337]
        8 |  0.00009738  [0.00009526  0.00009954]
        9 |  0.00009353  [0.00009122  0.00009589]
       10 |  0.00008983  [0.00008732  0.00009242]
-------------------------------------------------
     Key:  exp(xb)    =  exp(xb)
           [lb , ub]  =  [95% Confidence Interval]
```

The results are the same as those obtained manually using the 'lincom' command for 1990 (c_year=0) and 1995 (c_year=5). The exponentiated linear predictions can be multiplied by 100,000 to give age- and sex- adjusted incidence rates per 100,000 persons (and the precursor estimates, collapsed variables of age-group and gender and any dummy variables dropped).

. gen adjrate = exp*100000
. gen adjub =ub*100000
. gen adjlb =lb*100000
. drop exp lb ub gender agegroup _I*

Save age- and sex-adjusted mortality data for 1990–2000 as a new file (with 11 rows) after firstly sorting by c_year in preparation for producing a graph of the data.
. sort c_year
. save "C:\DATA\poisson traffic injury adjust 1990-2000.dta"

Return to the original data.

. clear

[and reopen the original saved data file of 66 rows]

. gen c_year = year-1990

The estimated crude mortality rates and 95% confidence intervals for each year can be obtained using the Poisson model and the 'adjust' command.

. quietly xi: poisson deaths c_year, exposure(personyears)

. adjust, by(c_year) exp ci nooffset format(%9.8f) replace

```
-------------------------------------------------------------------------------------
     Dependent variable: deaths     Equation: deaths     Command: poisson
-------------------------------------------------------------------------------------
-------------------------------------------------
   c_year |     exp(xb)          lb           ub
----------+--------------------------------------
        0 |  0.00015994  [0.00015603  0.00016394]
        1 |  0.00015317  [0.00014995  0.00015646]
        2 |  0.00014669  [0.00014406  0.00014938]
        3 |  0.00014048  [0.00013830  0.00014270]
        4 |  0.00013454  [0.00013265  0.00013645]
        5 |  0.00012885  [0.00012708  0.00013064]
        6 |  0.00012340  [0.00012157  0.00012525]
        7 |  0.00011817  [0.00011618  0.00012021]
        8 |  0.00011317  [0.00011094  0.00011545]
        9 |  0.00010839  [0.00010589  0.00011094]
       10 |  0.00010380  [0.00010104  0.00010664]
-------------------------------------------------
     Key:  exp(xb)    =  exp(xb)
           [lb , ub]  =  [95% Confidence Interval]
```

The exponentiated linear predictions can be multiplied by 100,000 to give crude incidence rates per 100,000 persons (and the precursor estimates and collapsed variables dropped).

. gen erate = exp*100000
. gen eub =ub*100000
. gen elb =lb*100000
. drop exp lb ub gender agegroup

Save the crude mortality data for 1990–2000 as a new file (with 11 rows) after firstly sorting by c_year in preparation for merging with other files.
. sort c_year
. save "C:\DATA\poisson traffic injury crude 1990-2000.dta"

Return to the original data.

. clear

[and reopen the original saved data file of 66 rows]

Use 'collapse' to condense the 66 rows to 11 rows to summarise the crude mortality rate for all persons for 1990–2000.

. collapse (sum) deaths personyears, by (year)

. gen rate = (deaths/personyears)*100000

To rescale calendar year as a continuous variable, subtract the initial year (1990):

. gen c_year = year-1990

The ‘merge’ command can be used to horizontally merge the variables in the two files on the matching variable, c_year, so that each observation from one data set is merged with the corresponding observation in the other data set. The crude mortality data file remains open and in use, while the ‘merge’ command imports data from the adjusted mortality data file.

```
. sort c_year
. merge c_year using "C:\DATA\poisson traffic injury crude 1990-2000.dta"
```

As a check, Stata adds a system variable to the data set named _merge, which should have a value of 3 for each row if the observations in both files are correctly paired. Once checked, merge can be dropped.

```
. drop _merge
```

The file now contains the crude mortality rate and its trend line estimates predicted from the data using the Poisson model.

The estimates for the age- and sex- adjusted injury mortality rate can be merged into the file as follows:

```
. sort c_year
. merge c_year using "C:\DATA\poisson traffic injury adjust 1990-2000.dta"
. drop _merge
```

The crude and age- and sex- adjusted mortality rates are now in one file, which allows them to be graphically depicted together according to year.

To report the estimated trend in the age- and sex- adjusted injury mortality rate as annual percent change, the expected number of deaths is calculated for each year from the estimates predicted using the ‘adjust’ command.

```
 . gen adjdeaths = adjrate*personyears/100000
. save "C:\DATA\poisson traffic injury crude and adjust 1990-2000.dta"
```

The estimated trend is obtained from the fit of the Poisson regression model, and can be reported as annual per cent change as $100\left[\exp(\beta_1)-1\right]$.

. poisson adjdeaths c_year, exposure(personyears)

```
note: you are responsible for interpretation of non-count dep. variable

Iteration 0:   log likelihood = -50.552467
Iteration 1:   log likelihood = -50.552467  (backed up)

Poisson regression                                Number of obs   =         11
                                                  LR chi2(1)      =     280.06
                                                  Prob > chi2     =     0.0000
Log likelihood = -50.552467                       Pseudo R2       =     0.7347

------------------------------------------------------------------------------
     edeaths |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      c_year |   -.040328   .0024132   -16.71   0.000    -.0450578   -.0355982
       _cons |  -8.914306   .0137182  -649.82   0.000    -8.941193   -8.887419
 personyears | (exposure)
------------------------------------------------------------------------------
```

. di 100*(exp(-.040328)-1)

-3.9525648

. di 100*(exp(-.0450578)-1), 100*(exp(-.0355982)-1)

-4.4057773 -3.4972036

[lower CI, upper CI]

The average annual per cent change in age- and sex- adjusted mortality rate from traffic accidents during 1990–2000 is -4.0% [95% CI: -4.4% to -3.5%]. Changes over a longer period of time can similarly be estimated. For 10 years onwards from 1990, using the formula $100\left[\exp(10\cdot\beta_1)-1\right]$ a decline of 33.2% is expected.

. di 100*(exp(10*-.040328)-1)

-33.1875

If the estimated age- and sex- adjusted mortality rate in 1990 is 13.4 per 100,000 person-years [. display exp(-8.914306)*100000], then the model predicts that the mortality rate would decline to 8.9 per 100,000 person-years in 2000.

Figure 3.13 shows the crude mortality rate and its estimated trend line from 1990–2000, and the age- and sex- adjusted mortality rates and its trend line (obtained using the 'adjust' command).

Both the crude and the age- and sex-adjusted models show a gradually flattening downward trend in fatalities from motor accidents during the decade. The values obtained from the adjusted model differ from those from the crude model because they are for a "what if?" hypothetical population in which each age-group and each sex is equally represented.

The method used in this chapter to age- and sex-adjust is an example of internal standardisation. The road vehicle injury mortality data for males and females for each year of the decade is standardised using mortality data for the total population of Australia for the years under study. It is also possible to age- and/or sex-adjust using an external standard, that is, a population drawn from sources outside the analysis (e.g. the 2001 census of the Australian population). The advantages and disadvantages of internal versus external standardisation are detailed in Chapter 4.

The absolute values of the adjusted rates should not be interpreted as measuring real risk in a population. Their value is that they can be compared validly with other rates adjusted in the same way, to assess changes or differences in rates, after allowing for differences in the age and sex composition of the populations under study. Hence, the trend shown in Figure 3.13 for the age- and sex-adjusted rate is a more valid estimate of change in road vehicle injury mortality risk than the trend in crude rates, because it removes confounding due to any changes in the age or gender distribution of the population over the years of study.

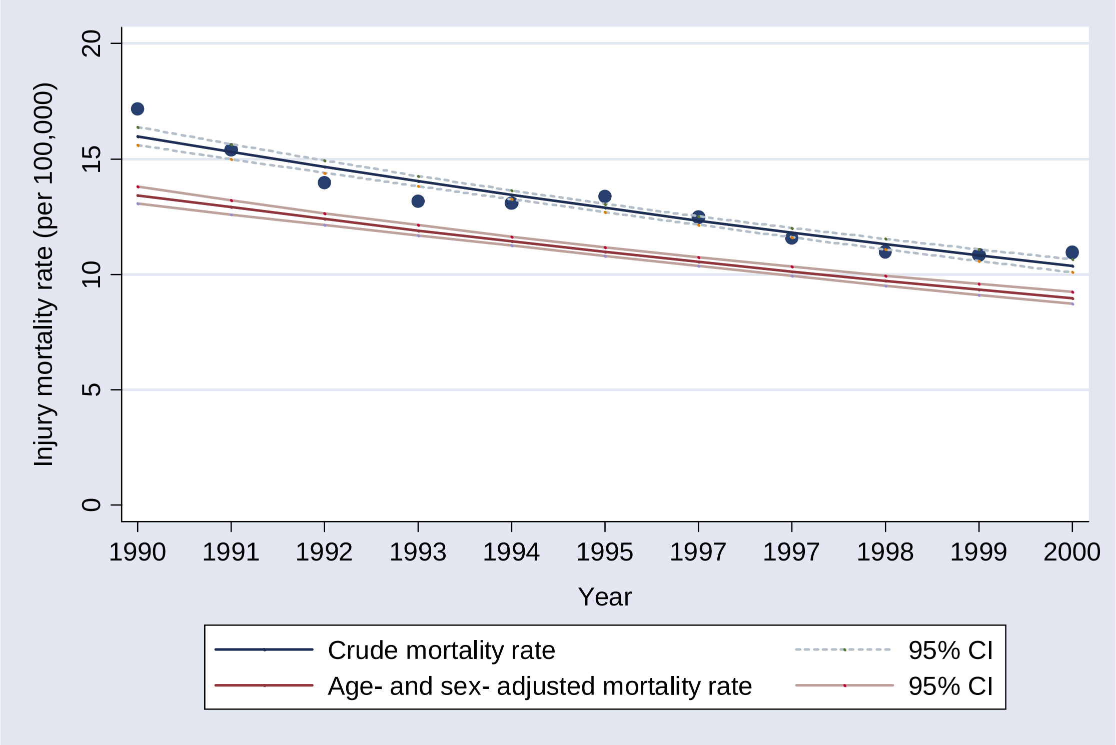

Figure 3.13: Crude and age- and sex- adjusted motor vehicle injury mortality rates in Australia, 1990–2000

The ‘glm’ command will give the same results as the ‘poisson’ command, but provides a detailed output of the diagnostics of model fit. To demonstrate the use of the ‘glm’ command, we can specify Poisson regression by using the natural log as the link function and Poisson as the family distribution. The exposure (for adjustment of the counts to reflect person-time) is entered as the log person-years using the offset option. To display incidence rate ratios, we can use the option ‘eform’ which will exponentiate the regression coefficients. Using the original saved data file of 66 rows:

. xi: glm deaths c_year i.agegroup i.gender, family(poisson) link(log) lnoffset(personyears) eform

```
i.agegroup        _Iagegroup_1-3      (naturally coded; _Iagegroup_1 omitted)
i.gender          _Igender_0-1        (naturally coded; _Igender_0 omitted)

Iteration 0:   log likelihood = -482.68636
Iteration 1:   log likelihood = -439.47428
Iteration 2:   log likelihood = -439.37014
Iteration 3:   log likelihood = -439.37014

Generalized linear models                          No. of obs      =        66
Optimization     : ML: Newton-Raphson              Residual df     =        61
                                                   Scale parameter =         1
Deviance         =   394.520583                    (1/df) Deviance =  6.467551
Pearson          =  398.7112116                    (1/df) Pearson  =  6.536249

Variance function: V(u) = u                        [Poisson]
Link function    : g(u) = ln(u)                    [Log]
Standard errors  : OIM

Log likelihood   = -439.3701411                    AIC             =  13.46576
BIC              =  138.9516438

------------------------------------------------------------------------------
      deaths |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      c_year |   .9604744   .0021422   -18.08   0.000     .9562849    .9646822
_Iagegroup_2 |   .5188357   .0096047   -35.45   0.000     .5003484    .5380062
_Iagegroup_3 |     .94036   .0165802    -3.49   0.000     .9084185    .9734247
  _Igender_1 |   .4005581   .0061953   -59.15   0.000     .3885978    .4128865
 personyears | (exposure)
------------------------------------------------------------------------------
```

The results are the same as the Poisson model.

. xi: poisson deaths c_year i.agegroup i.gender, exposure(personyears) ir

```
i.agegroup        _Iagegroup_1-3      (naturally coded; _Iagegroup_1 omitted)
i.gender          _Igender_0-1        (naturally coded; _Igender_0 omitted)

Iteration 0:   log likelihood = -439.51166
Iteration 1:   log likelihood = -439.37014
Iteration 2:   log likelihood = -439.37014

Poisson regression                                Number of obs   =         66
                                                  LR chi2(4)      =    5643.68
                                                  Prob > chi2     =     0.0000
Log likelihood = -439.37014                       Pseudo R2       =     0.8653

------------------------------------------------------------------------------
      deaths |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      c_year |   .9604744   .0021422   -18.08   0.000     .9562849    .9646822
_Iagegroup_2 |   .5188357   .0096047   -35.45   0.000     .5003484    .5380062
_Iagegroup_3 |     .94036   .0165802    -3.49   0.000     .9084185    .9734247
  _Igender_1 |   .4005581   .0061953   -59.15   0.000     .3885978    .4128865
 personyears | (exposure)
------------------------------------------------------------------------------
```

We should check whether the Poisson model is a reasonable fit for the data. With the model in memory, we use a goodness-of-fit test for the model. This tests whether there is any variation in the counts over and above what would be expected from a Poisson regression model. If the test is significant, then the data are overdispersed. The consequences of using a Poisson model with overdispersion is that the confidence intervals are too narrow, p-values are too small, and there is overstated statistical significance.[8]

.poisgof

```
         Goodness-of-fit chi2  =  394.5372
         Prob > chi2(61)       =    0.0000
```

The test is significant for overdispersion, so a Negative Binomial Distribution (NBD) regression model is a better alternative.

3.9 Negative Binomial Distribution regression

The Negative Binomial Distribution (NBD) accommodates heterogeneity of risk so this model allows for overdispersion—resulting in higher p-values and wider CIs.

. xi: nbreg deaths c_year i.agegroup i.gender, exposure(personyears) ir

```
i.agegroup        _Iagegroup_1-3      (naturally coded; _Iagegroup_1 omitted)
i.gender          _Igender_0-1        (naturally coded; _Igender_0 omitted)

Fitting Poisson model:

Iteration 0:   log likelihood = -439.51166
Iteration 1:   log likelihood = -439.37014
Iteration 2:   log likelihood = -439.37014

Fitting constant-only model:

Iteration 0:   log likelihood =  -438.2113
Iteration 1:   log likelihood = -429.76719
Iteration 2:   log likelihood = -416.01926
Iteration 3:   log likelihood = -415.98642
Iteration 4:   log likelihood =  -415.9864

Fitting full model:

Iteration 0:   log likelihood = -386.90879
Iteration 1:   log likelihood = -385.26897  (not concave)
Iteration 2:   log likelihood = -349.33947
Iteration 3:   log likelihood = -334.73441
Iteration 4:   log likelihood = -332.91503
Iteration 5:   log likelihood = -332.90584
Iteration 6:   log likelihood = -332.90583

Negative binomial regression                      Number of obs   =         66
                                                  LR chi2(4)      =     166.16
                                                  Prob > chi2     =     0.0000
Log likelihood = -332.90583                       Pseudo R2       =     0.1997

------------------------------------------------------------------------------
      deaths |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      c_year |   .9591253   .0057162    -7.00   0.000     .9479869    .9703945
_Iagegroup_2 |   .5480804   .0253926   -12.98   0.000     .5005045    .6001786
_Iagegroup_3 |    1.02583   .0472495     0.55   0.580     .9372802    1.122747
  _Igender_1 |   .4362934   .0167283   -21.63   0.000     .4047082    .4703436
 personyears | (exposure)
-------------+----------------------------------------------------------------
    /lnalpha |  -3.951466   .2106765                     -4.364385   -3.538548
-------------+----------------------------------------------------------------
       alpha |   .0192265   .0040506                      .0127225    .0290555
------------------------------------------------------------------------------
Likelihood-ratio test of alpha=0:  chibar2(01) =  212.93 Prob>=chibar2 = 0.000
```

In the output above, the likelihood-ratio test is listed after the estimates of the parameters. In the NBD, the parameter 'alpha' (α) determines the degree of dispersion in the predicted counts.[9] If $\alpha = 0$, the NBD regression model reduces to the Poisson regression model, which is the key to testing for overdispersion.[9] The likelihood-ratio test indicates there is significant evidence of overdispersion (G^2=212.93, p<0.001) so the NBD regression model is preferred to the Poisson regression model.

To test whether including gender in the model is worthwhile over the simpler model with calendar year only, we can use a likelihood ratio test. The log-likelihood is a measure of the fit of a model.[4] It compares the log-likelihood of two models: a full (larger) model, and a reduced (smaller) model, that is nested within the full model (i.e. it contains the same parameters). The likelihood ratio test statistic is calculated as two times the difference in log-likelihoods of the two models. The degrees of freedom for this test are found by subtracting the degrees of freedom used to generate each model. The larger the log-likelihood ratio test statistic, the smaller the associated probability and the better the fit.[4]

```
. quietly xi: nbreg deaths c_year i.gender, exposure(personyears) ir
. estimates store A
. quietly nbreg deaths c_year, exposure(personyears) ir
. lrtest A
```

```
likelihood-ratio test                                 LR chi2(1)  =     64.91
(Assumption: . nested in A)                           Prob > chi2 =    0.0000
```

The test is significant on 1 degree of freedom, so gender is important to the model.

To test whether including the age categories in the model is worthwhile over the simpler model with calendar year and gender only:

```
. quietly xi: nbreg deaths c_year i.agegroup i.gender, exposure(personyears) ir
. estimates store A
. quietly xi: nbreg deaths c_year i.gender, exposure(personyears) ir
. lrtest A
```

```
likelihood-ratio test                                 LR chi2(2)  =     96.86
(Assumption: . nested in A)                           Prob > chi2 =    0.0000
```

The test is significant on 2 degrees of freedom, so age is important to the model.

To test for interaction between age and year (i.e. whether there is evidence that age affects mortality risk differently in different years), we can generate an interaction term and test whether the model is improved with its inclusion using the likelihood ratio test.

```
. quietly xi: nbreg deaths c_year i.agegroup i.agegroup*c_year gender,
exposure(personyears)
. est store A
. quietly xi: nbreg deaths c_year i.agegroup i.gender, exposure(personyears)
. lrtest A
```

```
likelihood-ratio test                                 LR chi2(2)  =      0.95
(Assumption: . nested in A)                           Prob > chi2 =    0.6209
```

The test is not significant, so the interaction term for age and year is not important to the model.

Note: The syntax of the test for interaction between age and year can be simplified by leaving out the variables ‘age’ and ‘year’ when the interaction term ‘age*year’ is included in the model. This is because the variables ‘age’ and ‘year’ are dropped due to collinearity when the interaction term ‘age*year’ is specified. The example below demonstrates that using the simplified syntax gives the same results as the longer syntax in the example above.

```
. quietly xi: nbreg deaths i.agegroup*c_year i.gender, exposure(personyears)
. est store A
. quietly xi: nbreg deaths c_year i.agegroup i.gender, exposure(personyears)
. lrtest A
```

```
likelihood-ratio test                                 LR chi2(2)  =      0.95
(Assumption: A nested in .)                           Prob > chi2 =    0.6209
```

To test for interaction between age and gender (i.e. whether there is evidence that age affects mortality risk differently for males compared to females), we can generate an interaction term and test it.

```
. quietly xi: nbreg deaths c_year i.agegroup*gender, exposure(personyears)
. est store A
. quietly xi: nbreg deaths c_year i.agegroup i.gender, exposure(personyears)
. lrtest A
```

```
likelihood-ratio test                                 LR chi2(2)  =     97.71
(Assumption: . nested in A)                           Prob > chi2 =    0.0000
```

The test is significant, so the full model should include the interaction term for age and gender.

. xi: nbreg deaths c_year i.agegroup*gender, exposure(personyears) ir

```
i.agegroup        _Iagegroup_1-3      (naturally coded; _Iagegroup_1 omitted)
i.ageg~p*gender   _IageXgende_#       (coded as above)

Fitting Poisson model:

Iteration 0:   log likelihood = -326.98529
Iteration 1:   log likelihood =  -286.1917
Iteration 2:   log likelihood = -286.13113
Iteration 3:   log likelihood = -286.13113

Fitting constant-only model:

Iteration 0:   log likelihood =  -438.2113
Iteration 1:   log likelihood = -429.76719
Iteration 2:   log likelihood = -416.01926
Iteration 3:   log likelihood = -415.98642
Iteration 4:   log likelihood =  -415.9864

Fitting full model:

Iteration 0:   log likelihood =  -384.5463
Iteration 1:   log likelihood = -338.26829  (not concave)
Iteration 2:   log likelihood = -305.56226
Iteration 3:   log likelihood = -285.90909
Iteration 4:   log likelihood = -284.16917
Iteration 5:   log likelihood = -284.05275
Iteration 6:   log likelihood = -284.05172
Iteration 7:   log likelihood = -284.05172

Negative binomial regression                      Number of obs   =         66
                                                  LR chi2(6)      =     263.87
                                                  Prob > chi2     =     0.0000
Log likelihood = -284.05172                       Pseudo R2       =     0.3172

------------------------------------------------------------------------------
      deaths |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      c_year |   .9601216   .0025958   -15.05   0.000     .9550473    .9652228
_Iagegroup_2 |   .4747157   .0124038   -28.51   0.000     .4510167    .4996599
_Iagegroup_3 |   .7568098   .0201228   -10.48   0.000     .7183799    .7972955
      gender |   .3200503   .0082321   -44.29   0.000     .3043157    .3365985
_IageXgend~2 |   1.386656    .063363     7.15   0.000     1.267866    1.516576
_IageXgend~3 |   1.903146   .0804995    15.21   0.000     1.751733    2.067647
 personyears | (exposure)
-------------+----------------------------------------------------------------
    /lnalpha |  -6.809526   .6760116                     -8.134485   -5.484568
-------------+----------------------------------------------------------------
       alpha |   .0011032   .0007458                      .0002933    .0041503
------------------------------------------------------------------------------
```

```
Likelihood-ratio test of alpha=0:  chibar2(01) =    4.16 Prob>=chibar2 = 0.021
```

Adjusting for age and sex in the NBD regression model, there is a small but significant decline in mortality from unintentional motor vehicle accidents over the years 1990–2000 (RR=0.96; p<0.001). The mortality rate in those aged 40–59 years is about half that of the reference age-group 15–39 years (RR=0.47; p<0.001). The mortality rate in those aged 60+ years is about one-quarter less than the 15–39 year age-group (RR=0.76 p<0.001). There is a strong protective effect of being female, with a markedly lower mortality from motor vehicle accidents compared to males (RR=0.32; p<0.001).

With the NBD model (with the interaction term) in memory, the expected count of deaths by age, gender and year can be computed.

. predict edeaths

(option n assumed; predicted number of events)

The table below displays by gender and age-group—the sum of the deaths, the sum of the expected deaths predicted by the model, and person-years (going down the columns). The agreement between observed and expected counts is reasonable, and is better than when no interaction term is included in the NBD model (data not shown). The NBD model with the interaction term included seems to be a good model for these data.

. table gender agegroup, c(sum deaths sum edeaths sum personyears) row col format(%9.0f)

```
--------------------------------------------------------------
          |                        agegroup
   gender | 15-39 years  40-59 years    60+ years        Total
----------+---------------------------------------------------
     male |        9081         2696         2532        14309
          |        9088         2696         2532        14316
          |    38535615     24336899     14295401     77167915
          |
   female |        2868         1170         1882         5920
          |        2868         1170         1881         5919
          |    38013741     23827474     17413903     79255118
          |
    Total |       11949         3866         4414        20229
          |       11955         3866         4413        20234
          |    76549356     48164373     31709304    156423033
--------------------------------------------------------------
```

One of the most common methods of interpretation for count models (Poisson, NBD) is looking at the factor change in the rate.[9] The option ‘ir’ computes these coefficients, but alternatively they can be displayed (with more detail) using the ‘listcoef’ command. Re-running the NBD model so that it is in memory, use ‘listcoef’ to compute the coefficients:

. quietly xi: nbreg deaths c_year i.age*gender, exposure(personyears) ir

. listcoef year, help

```
nbreg (N=66): Factor Change in Expected Count

 Observed SD: 244.41971

----------------------------------------------------------------------
      deaths |      b         z     P>|z|    e^b    e^bStdX      SDofX
-------------+--------------------------------------------------------
      c_year |  -0.04070  -15.052   0.000   0.9601   0.8784     3.1865
-------------+--------------------------------------------------------
    ln alpha |  -6.80953
       alpha |   0.00110   SE(alpha) = 0.00075
----------------------------------------------------------------------
 LR test of alpha=0: 4.16     Prob>=LRX2 = 0.021
----------------------------------------------------------------------
       b = raw coefficient
       z = z-score for test of b=0
   P>|z| = p-value for z-test
     e^b = exp(b) = factor change in expected count for unit increase in X
 e^bStdX = exp(b*SD of X) = change in expected count for SD increase in X
   SDofX = standard deviation of X
```

The coefficient can be interpreted as follows: Each year the expected number of deaths decreases by a factor of 0.96, holding all other variables constant.

To compute per cent change, we add the option ‘percent’.

. listcoef year, percent help

```
nbreg (N=66): Percentage Change in Expected Count

 Observed SD: 244.41971

----------------------------------------------------------------------
      deaths |      b         z     P>|z|      %      %StdX      SDofX
-------------+--------------------------------------------------------
      c_year |  -0.04070  -15.052   0.000     -4.0    -12.2     3.1865
-------------+--------------------------------------------------------
    ln alpha |  -6.80953
       alpha |   0.00110   SE(alpha) = 0.00075
----------------------------------------------------------------------
 LR test of alpha=0: 4.16     Prob>=LRX2 = 0.021
----------------------------------------------------------------------
       b = raw coefficient
       z = z-score for test of b=0
   P>|z| = p-value for z-test
       % = percent change in expected count for unit increase in X
   %StdX = percent change in expected count for SD increase in X
   SDofX = standard deviation of X
```

The coefficient can be interpreted as follows: Each year the expected number of deaths decreases by 4.0%, holding all other variables constant.

This chapter demonstrates that the NBD model is a better fit for the motor vehicle injury mortality data than a Poisson model due to data being overdispersed. The method for internal standardisation as shown in Chapter 3.8 (and external standardisation as shown in Section 4.6.2) can easily be applied to a NBD model by replacing the 'poisson' command with 'nbreg', as follows:

```
. xi: nbreg deaths c_year i.agegroup i.gender, exposure(personyears)
```

By using the 'adjust' command once the NBD model is in memory, estimated age- and sex- adjusted mortality rates and their 95% confidence intervals will be calculated (as per Figure 3.13).

3.10 Modelling SMRs

Instead of a measure of person-years at risk as the exposure in the model, we could use the predicted number of deaths to enable us to model the standardised mortality ratio (SMR). The predicted number of deaths can be derived from external data.[4, 17] The model would include the observed number of deaths as the counts, the exposure would be the expected number of events from the external data, and the measure would be the natural log(observed/expected), that is, ln(SMR).

As an example of external standardisation, we could use the age- and sex- specific mortality rates for road traffic injury for the World Health Organisation member States in 2002 to predict the expected number of deaths in each group [see equation (9) in Section 2.6.2].[18] This would allow international comparisons of mortality from motor vehicle accidents. If we were looking at mortality rates for one State, and wanted to compare it with the rates for all Australia, we could use the age-specific death rates for all Australians to predict the expected numbers of deaths for males and females by age group. If the aim was to examine time trends, special consideration would be required to decide which year(s) to use as the standard, but a common choice would be the census year ending in one (i.e. 1991, 2001 etc.).

3.11 Artefacts of measurement

A consideration when interpreting trends over time must be whether any artefacts of the statistical measurement process exist. Improvements in case ascertainment for morbidity and mortality data over time could contribute to an observed increase in trend. Likewise, changes to coding practice may mean some causes of morbidity and mortality are coded differently to previous years and this could contribute to an erroneous increase or decrease in trend over time. The effect of misclassification bias on trend analysis must also be considered, as misclassification of some types of injury is likely to occur. For example, a certain proportion of road traffic deaths are likely to be attributable to suicide.

3.12 Conclusion

This section has demonstrated that a Pearson's chi-square is a suitable test for difference for nominal (unordered) data as it is able to test for departures from the null hypothesis that can occur in various ways, but is not suitable as a test for trend for ordered data. For dose-response or trend data, we need to use different statistics. Suitable chi-squared tests are the Cochran-Armitage test for trend or Mantel-Haenszel test for trend. Sribney[3] evaluated various tests for trend (including Mantel-Haenszel chi-square test for trend, Pearson's correlation, Cochran-Armitage test ['ptrend'], and 'nptrend' command) and demonstrated that these tests are simply a Pearson's correlation coefficient (a test suitable to determine the correlation between two variables and its significance). Although these tests can detect a linear trend, they are not powerful against non-linear trends.[3] In these instances, the Pearson's chi-square statistic can detect the association with more power.

Regression procedures are more powerful tests of association than chi-square statistics and allow confounders and effect modifiers to be included in the model and adjusted for simultaneously. Regression procedures can test for non-linear trends (e.g. curvilinear). Ordinary Least Squares regression should be avoided for trend analysis as it assumes homogeneity of variance in the residuals (errors) and can be biased for count data due to the assumptions that the data are normally distributed. Variance-weighted least squares regression is preferable to Ordinary Least Squares as it accounts for non-homogeneity of variance and can be used to measure trend over time for directly standardised mortality rates (if numerator and denominator data are not available). For binary data, logistic regression is a powerful test for trend. Count data often follow a Poisson distribution, so if numerator and denominator data are available, Poisson analysis is most appropriate, and can also be used to model Standardised Mortality Ratios. However, if the data prove to be overdispersed, Negative Binomial Distribution regression is a more appropriate method.

3.13 References

1. Armitage P, Berry G, Matthews JNS. Statistical methods in medical research (4th edition). UK: Blackwell Publishing, 2002.
2. Stata statistical software, release 8.2. College Station, Texas: Stata Corporation, 2001.
3. Sribney W. A comparison of different tests for trend. College Station, Texas: Statacorp LP; 1996.
4. Campbell MJ. Statistics at Square Two: Understanding Modern Statistical Applications in Medicine. London: BMJ Books, 2001.
5. Fleiss JL. Statistical methods for rates and proportions. NY, USA: John Wiley & Sons, 2000.
6. Sasieni P. Stratified test for trend across ordered groups. Stata Technical Bulletin Reprints 1996; 6: 196-200. Available at (www.stata.com) as STB-33 snp12 (help npt_s if installed) 9/96 (accessed Aug 3004).
7. Rosenberg D. Trend analysis and interpretation. Maryland, USA: Maternal and Child Health Bureau; 1997.
8. Ryan, P. A short course in elementary biostatistics (2nd edition, revised). Adelaide, Australia: Department of Public Health, University of Adelaide, 2004. {Available at: (www.public-health.adelaide.edu.au/staff/ryan_phil.html) (accessed Aug 2004).
9. Long JS, Freese J. Regression models for categorical dependent variables using Stata (revised edition). College Station, Texas: Stata Corporation, 2003.
10. Australian Bureau of Statistics. Australian historical population statistics—5. Deaths. Table 46: Infant mortality rates, states and territories, 1991 onwards (cat. no. 3105.0.65.001) Canberra: Commonwealth of Australia, 2003.
11. Statacorp. Stata base reference manual, S-Z. Release 8. College Station, Texas: Stata Corportation, 2004; 4: 312–317.
12. ABS unit-record mortality data for deaths registered 1997-2002. [Datafile] Analysed by the Australian Institute of Health and Welfare National Injury Surveillance Unit, 2004.
13. Hosmer DW, Lemeshow S. Applied Logistic Regression (2nd edition). NY, USA: John Wiley & Sons, 2000.
14. Campbell MJ, Machin D. Medical Statistics: A commonsense approach (3rd edition). Chichester, England: John Wiley & Sons, Inc; 2002.
15. ABS unit-record mortality data for deaths registered 1990–2000. [Datafile] Analysed by the Australian Institute of Health and Welfare National Injury Surveillance Unit, 2004.
16. Selvin S. Practical Biostatistical Methods. USA; Duxberry Press, 1995.
17. Breslow NE, Day NE. In Statistical methods in cancer research. Volume II—the design and analysis of cohort studies. Lyon; International Agency for Research on Cancer, 1987.
18. World Health Organization. World Report on Road Traffic Injury Prevention. WHO: Geneva. 2004.

4 Regression analyses of mortality rates

4.1 Analysis of mortality rates

A common type of observational study uses population data to assess the mortality rates which are specific to certain groups. An example is a comparison of the mortality rates from suicide over calendar year by gender. A related type of observational study involves an exposure or treatment that is more prevalent in some regions or groups than in others. The relationship between the extent of exposure and the outcome is studied in order to assess the effects of exposure. Examples include (i) a study of socioeconomic determinants of mortality, and (ii) studies which examine mortality rates from a particular cause in various countries and their relationship to exposure/environmental factors.

Statisticians have pointed out the need for caution in regression analyses of standardised rates that have (age-specific) population denominators in both dependent and independent variables.[1] A common mistake is to undertake the analysis when the outcome variables used in the analyses, such as mortality rates in various regions, have been age-adjusted, but the predictor variables have not been age-adjusted. If mortality is adjusted for age in regression analyses, then any covariates (e.g. income, co-morbidities) must be adjusted for age as well.

The use of crude regional death rates as the outcome variable, with crude covariates and age as predictors, can avoid the problems encountered with regression of age-adjusted mortality rates.

4.2 Age-adjustment and regression

There is a need for caution when performing regression analysis of age-adjusted rates. If we designate:

Y	= an age- and state-specific mortality rate
X_2	= the corresponding age
X_1	= the per capita personal income (a variable that varies with both age and state)

Suppose we wish to estimate the regression coefficient, $\beta YX_1 * X_2$, of Y on X_1 in the multiple regression with two predictors, X_1 and X_2. The least squares estimate of this coefficient may be found by, first, regressing Y on X_2 and calculating the residuals $Y*X_2$ (to age-adjust mortality), then regressing X_1 on X_2 and calculating the residual X_1*X_2 (to age-adjust income), and finally calculating the estimate of $\beta YX_1 * X_2$ as the estimated slope in the regression of the first set of residuals $Y*X_2$ on the second X_1*X_2.[1]

To find the least squares estimate of $\beta YX_1 * X_2$, the age-adjusted mortality, $Y*X_2$, should be regressed on age-adjusted income, X_1*X_2. Often, however, age-adjusted mortality, $Y*X_2$, is regressed on income, X_1. This mistaken approach will give a biased estimate, unless income, X_1, and age, X_2, are statistically unrelated.[1]

4.3 Multivariate methods of age-standardisation

4.3.1 Biased methods

The methods that generally lead to a biased least squares estimate of $\beta YX_1 * X_2$ are:

1. Weighted regression of age-adjusted rates on crude predictors
2. Weighted regression of age-adjusted rates on age and crude predictors
3. Weighted regression of age-specific rates on crude predictors

Method 1 is a popular technique, but should be avoided except in situations in which the bias can be shown to be negligible for the purposes of the study. Method 2 can only produce an unbiased least squares estimate under restrictive conditions on the relationship between adjusted and unadjusted covariates which limits its general applicability. [1] In addition to it leading to a biased least squares estimate, Method 3 would not suit the format of data presentation for NISU reports. For further discussion of why these three methods lead to biased estimates, see Rosenbaum et al. [1]

4.3.2 Unbiased methods

There are a number of methods that will give an unbiased least squares estimate of $\beta YX_1 * X_2$. These are:

1. Weighted regression of crude response rates on age and crude predictor averages
2. Weighted regression of age-adjusted rates on age-adjusted predictors
3. Weighted regression of age-specific rates on age and on age-specific predictor averages
4. Regression of the responses of individuals on the age of individuals and the predictors describing individuals

For the purposes of this report, we will concentrate on Methods 1 (internal standardisation) and 2 (external standardisation), which have the most practical application for multivariate analyses in NISU reports. Method 3 does not suit the format of data used in NISU reports. Method 4 does not suit the format of NISU reports and is not commonly used due to restrictions in obtaining the necessary data from official sources e.g. NISU data are limited to aggregate population data and usually lack person-level data for non-cases.

Two types of standard population can be chosen for age-adjustment: external and internal. Usually an internal standard consists of combining the groups to be compared to form one standard group, but another valid approach is to select a specific group among those sampled to serve as a standard (e.g. males could be chosen). The internal 'standard' group is used as a source of age-specific counts (direct standardisation) or rates (indirect standardisation).[2] For external standardisation, a population external to the study cohort is used as a source of age-specific counts (direct standardisation) or rates (indirect standardisation).

4.4 Method 1: Internal standardisation

Regression with crude rates and crude covariate averages, including age

This method uses crude age (a pooled combination of all study groups) to internally standardise each group in order to make comparisons between them. As an example, if the study populations consisted of Indigenous and non-Indigenous Australians, this method would internally standardise each study population (Indigenous yes or no) to the total population of Australia for the year(s) of the study. The method of internal standardisation has a strong following due to its relative simplicity and intuitive appeal.[3] By regressing crude rates on the independent predictor variable while simultaneously adjusting for age (added to the model as a covariate), age-adjusted mortality rates are generated. Internal standardisation can be preferable to external standardisation (Method 2) if there are questions about the appropriateness of using a particular external standard population. [3]

A limitation of internal standardisation is that it tends to yield mildly conservative tests and estimates in typical practice.[3] The conservatism increases if there is a high degree of association between the stratum-variables (e.g. age, year) and the exposures (i.e. if the stratification variables strongly confound the exposure-disease relationship).[3]

In general, the analysis of crude rates with age as a covariate can lead to unbiased estimates, and the simplicity of this method can make it a good choice of method to use.

4.5 Method 2: External standardisation

Regression with age-adjusted rates and age-adjusted covariate averages

This method relies on standard weights (direct standardisation) or standard rates (indirect standardisation) that are external to the study cohort in order to make comparisons between study groups. External standards are conventionally chosen from large populations published for use in the adjustment process.[2] An example would be the regression of age-standardised rates for Indigenous and non-Indigenous Australians by year, with rates directly standardised to the total Australian population in the census year 2001. For indirect standardisation, SMRs can be modelled externally as shown in the example below. Any covariates entered into the model that are age-specific (e.g. income, comorbidities) would need to be age-adjusted. A consideration before deciding to use Method 2 is the appropriateness of the particular standard population. [3]

Another potential limitation is that although age-adjusted mortality rates can be calculated easily, age-adjusting covariates (such as income) may not be a straightforward process. Likewise, if the data is derived from published sources, age-adjusted mortality rates are commonly published, but it is uncommon to find covariates that have been age-adjusted before tabulation.

4.6 Direct standardisation by Methods 1 and 2

4.6.1 Internal standardisation

Chapter 3.8 demonstrates internal standardisation using the direct method to adjust simultaneously for age and sex in motor vehicle injury mortality rates (see Figure 3.13). This method can be reduced to adjustment on one variable only e.g. rates are often standardised by adjusting for age alone (see Chapter 2).

4.6.2 External standardisation

To demonstrate external standardisation, we will use the data from Chapter 3.8 on deaths due to unintentional motor vehicle traffic injury in Australia (1990–2000) and standard population weights from the 2001 census of the Australian population (Table 4.1).[4] We have opted to perform age- and sex- standardisation to be consistent with the analyses in Chapter 3.8, but this method can be simplified to age-adjustment alone.

[As an aside, the NBD model was shown in Chapter 3.9 to be a better fit for the motor vehicle injury mortality data. However, we have opted to use the Poisson regression model here simply to be consistent with the analyses in Chapter 3.8. The model can be changed from Poisson to NBD by replacing the 'poisson' command with 'nbreg'.]

Table 4.1: Standard weights from the 2001 census of the Australian population.

	Standard population in each of the age bands, divided by the total standard population					
	Males		Females		Persons	
Age range (years)	(P_{si})	(w_{si})	(P_{si})	(w_{si})	(P_{si})	(w_{si})
15–39	3,520,707	0.22823	3,500,350	0.22691	7,021,057	0.45514
40–59	2,574,919	0.16692	2,572,508	0.16676	5,147,427	0.33368
60+	1,490,654	0.09663	1,766,904	0.11454	3,257,558	0.21117
Total	**7,586,280**	**0.49178**	**7,839,762**	**0.50822**	**15,426,042**	**1.00000**

Open the motor vehicle traffic injury data file of 66 rows.

To perform age- and sex- standardisation, the age- and sex- specific rates for the study populations (each year of motor vehicle traffic injury, 1990–2000) are multiplied by the age- and sex- specific weights of the 2001 Australian standard population.

[As an aside, to perform age-standardisation alone, the age-specific rates of the study populations are multiplied by the age-specific weights for persons in the 2001 Australian standard population.]

Firstly, create a variable which is the age- and sex- specific weight of the 2001 Australian standard population that corresponds to each age- and sex- stratum in the study population. Full decimal places are used for weighting to reduce rounding error.

```
. gen dirweights =.
. replace dirweights = 0.228231389 if agegroup==1 & gender==0
. replace dirweights = 0.166920264 if agegroup==2 & gender==0
. replace dirweights = 0.096632305 if agegroup==3 & gender==0
. replace dirweights = 0.226911738 if agegroup==1 & gender==1
. replace dirweights = 0.166763970 if agegroup==2 & gender==1
. replace dirweights =0.114540334 if agegroup==3 & gender==1
```

The stratifying variables (year, age-group and gender) must be sorted on prior to the use of the analytical commands, so that the age- and sex- adjusted rates for each strata are calculated correctly. The data can be sorted by either using the 'sort' command, or by using the 'collapse' command, as follows:

```
. sort year agegroup gender
```

OR

```
. collapse (sum) deaths personyears, by(year agegroup gender dirweights)
```

The age- and sex- specific rates for each study population (e.g. 1990, 2000, 2001 etc) are calculated.

```
. gen rate = (deaths/personyears)*100000
```

The age- and sex- specific rates for each year are multiplied by the corresponding age- and sex- specific weights in the 2001 Australian standard population.

```
. gen adjrate2001 = rate*dirweights
```

For each year of the study, the directly standardised rate is calculated by summing the 'adjrate2001' variable for its component age- and sex-strata [see equation (1) in Section 2.4.2]. The 'collapse' command sums the data by collapsing the 66 rows (11 rows for males by 3 age-groups, and the same for females) to 11 rows to summarise the age- and sex- adjusted mortality rate for all persons for 1990–2000.

```
. collapse (sum) adjrate2001 deaths personyears, by(year)
```

For each year of the study, the number of deaths for all persons can be adjusted (using the directly standardised rate calculated for the corresponding year). This enables a Poisson regression model to be fitted to the data to produce estimated age- and sex-adjusted mortality rates and their 95% confidence intervals for each year.

```
. gen deaths2001 =  adjrate2001* personyears/100000
```

Calendar year is entered as a continous variable, scaled by subtracting the initial year (1990) so that the intercept can be interpreted as the log of the baseline mortality rate in the first year (1990) (see Chapter 3.8 for detail).

```
. gen c_year= year-1990
```

The estimated trend line for the age- and sex-adjusted injury mortality rate and its 95% confidence intervals for each year can be obtained using the Poisson model and the 'adjust' command.

```
. xi: quietly poisson deaths2001 c_year, exp(personyears)
[quietly suppresses screen output so results are not shown]
. adjust, by(c_year) exp ci nooffset format(%9.8f) replace
```

The exponentiated linear predictors can be multiplied by 100,000 to give age- and sex-adjusted incidence rates per 100,000 persons (and the precursors dropped).

```
. gen rate2001 = exp*100000
. gen ub2001 = ub*100000
. gen lb2001 = lb*100000
. drop exp lb ub
```

Save the externally standardised age- and sex-adjusted mortality data for 1990–2000 as a new file (with 11 rows) after firstly sorting by c_year in preparation for merging with other files.

```
. sort c_year
. save "C:\DATA\poisson traffic injury external adjust 1990-2000.dta"
. clear
```

Return to the data file created in Chapter 3.8 which contains both the crude and internally standardised age- and sex- adjusted injury mortality rate estimates.

```
. use "C:\DATA\poisson traffic injury crude and adjust 1990-2000.dta"
. sort c_year
```

The 'merge' command can be used to horizontally merge the variables in the two files on the matching variable, c_year, so that the observations from one data set are merged with the corresponding observations in the other data set. The combined crude and internally age- and sex- standardised file remains open and in use, while the 'merge' command imports data from the externally age- and sex- standardised file.

```
. merge c_year using "C:\DATA\poisson traffic injury external adjust 1990-2000.dta"
 . save "C:\DATA\poisson traffic injury crude internal & external adjust 1990-2000.dta"
```

Figure 3.14 shows the crude mortality rate and its estimated trend line from 1990–2000, and the trend line obtained using internal and external age- and sex- adjustment of the mortality rate. The estimates of the trend line for the externally age- and sex- adjusted mortality rate are slightly lower than the crude mortality rate but the values converge in the latter years of the decade. This is as to be expected given that the population used for external standardisation was the 2001 census of the Australian population. The estimates of the trend line for the internally age- and sex- adjusted mortality rate are lower than those for external standardisation.

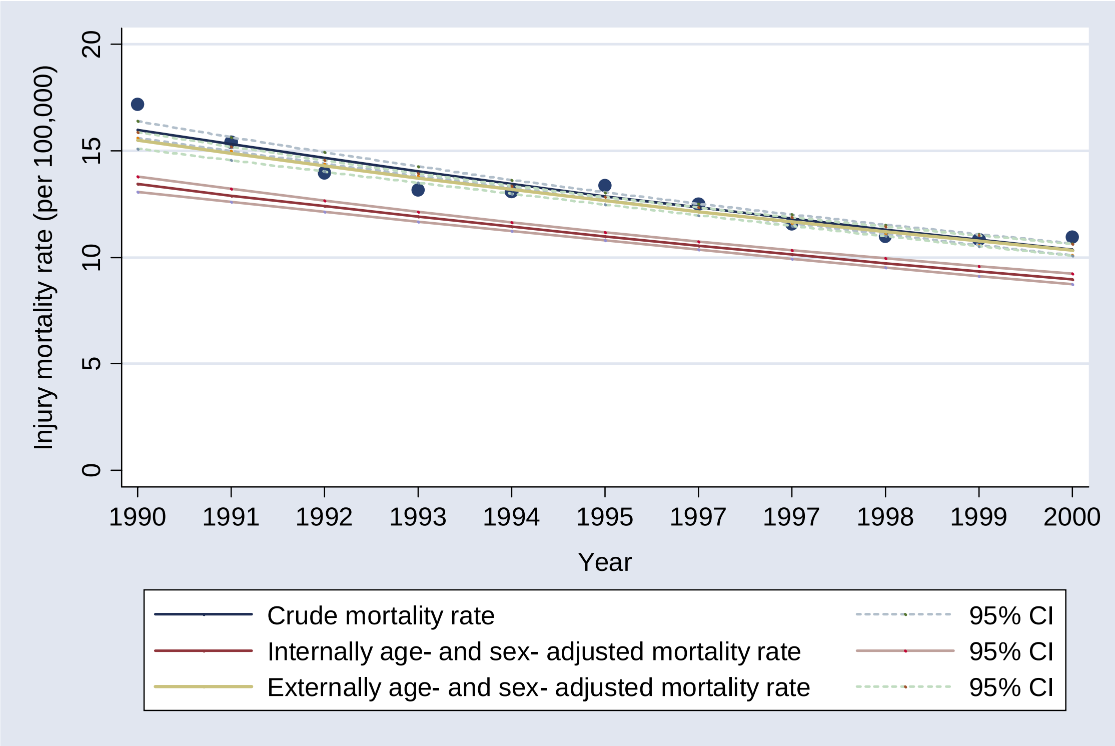

Figure 3.14: Crude and internally and externally age- and sex- adjusted motor vehicle injury mortality rates in Australia, 1990–2000

4.7 Indirect standardisation by Methods 1 and 2

4.7.1 Internal standardisation

To demonstrate internal standardisation, we will use ‘Grouped data from the Montana Smelter Workers Study’.[5] The dataset consists of respiratory cancer deaths and person-year denominators classified by age, calendar period, period of hire and estimated years of exposure to arsenic. For further information on the study and the analyses, refer to Breslow and Day.[3]

We will use Poisson regression to adjust respiratory cancer deaths for age and calendar period and look only at period of hire prior to 1925 (when the arsenic exposure was highest).

. xi: poisson resp_ca i.age i.calendar if periodhire==1, exp(personyrs)

```
i.age             _Iage_1-4           (naturally coded; _Iage_1 omitted)
i.calendar        _Icalendar_1-4      (naturally coded; _Icalendar_1 omitted)

Iteration 0:   log likelihood = -87.515486
Iteration 1:   log likelihood = -87.470217
Iteration 2:   log likelihood = -87.470092
Iteration 3:   log likelihood = -87.470092

Poisson regression                                Number of obs   =         52
                                                  LR chi2(6)      =      62.86
                                                  Prob > chi2     =     0.0000
Log likelihood = -87.470092                       Pseudo R2       =     0.2643

------------------------------------------------------------------------------
     resp_ca |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     _Iage_2 |   1.054763   .5501958     1.92   0.055    -.0236011    2.133127
     _Iage_3 |   1.870064   .5290349     3.53   0.000      .833175    2.906954
     _Iage_4 |   1.794457   .5461125     3.29   0.001     .7240963    2.864818
_Icalendar_2 |   .6323841    .256347     2.47   0.014     .1299532    1.134815
_Icalendar_3 |   .8435844   .2687129     3.14   0.002     .3169168    1.370252
_Icalendar_4 |   1.039504   .3287077     3.16   0.002     .3952485    1.683759
       _cons |  -7.473619   .5093356   -14.67   0.000    -8.471898   -6.475339
   personyrs | (exposure)
------------------------------------------------------------------------------
```

The expected numbers of respiratory cancer deaths are calculated for each exposure category by multiplying the pooled rates by the appropriate number of person-years and summing counts for each exposure category.

. predict edeaths

(option n assumed; predicted number of events)

To tabulate observed and expected numbers for arsenic exposure levels:

. table periodhire arsenic if periodhire==1, c(sum resp_ca sum edeaths sum personyrs) col format(%9.1f)

To determine relative risks for arsenic exposure levels:

```
. xi: poisson   resp_ca  i.arsenic i.age i.calendar if  periodhire==1, exp(personyrs) ir
```

These results are tabulated below:

Table 4.2: Dose-response analysis of respiratory cancer deaths among Montana smelter workers, based on internal standardisation.

	Cumulative years of moderate/heavy arsenic exposure				
Workers employed before 1925	**< 1 year**	**1–4 years**	**5–14 years**	**15+ years**	**Total**
No. of observed deaths	51	17	13	34	115
No. of expected deaths	77.5	10.6	10.2	16.8	115.0
(adjusted for age and calendar period)					
Relative risk	1.00	2.44	1.95	3.08	
(using ratio of Observed/Expected)					

The results in this table using Stata statistical software are similar to those in Table 3.11 in Breslow and Day.[3] Compare these results using internal standardisation to Table 4.4 in the next section which uses external standardisation.

4.7.2 External standardisation

To demonstrate external standardisation, we will use 'Grouped data from the Montana Smelter Workers Study'[5] and standard population rates (Table 4.3 [abridged from Table 3.2 in Breslow and Day[3]]).

Table 4.3: Standard respiratory cancer death rates used for comparative analyses of the Montana smelter workers data.

	No. of deaths per 1,000 person-years Calendar period			
Age range (years)	**1938–1949**	**1950–1959**	**1960–1969**	**1970–1977**
40–49	0.14817	0.21896	0.28674	0.37391
50–59	0.47412	0.80277	1.05824	1.25469
60–69	0.73136	1.55946	2.33029	2.90461
70–79	0.73207	1.63585	2.85724	4.22945

Remembering that:

$$\mathrm{SMR} = \frac{\text{observed deaths (in study population)}}{\text{expected deaths (in study population)}} = \frac{\mathrm{D_i}}{\sum \mathrm{r_{si}}\ \mathrm{P_i}}$$

To generate expected deaths we need to multiply the standard respiratory cancer death rates for each age- and calendar- stratum by the stratum's person-years denominator and then sum them for each exposure group. In Stata:

```
gen istdrate =.
replace  istdrate=0.14817 if age==1 & calendar==1
replace  istdrate=0.21896 if age==1 & calendar==2
replace  istdrate=0.28674 if age==1 & calendar==3
replace  istdrate=0.37391 if age==1 & calendar==4
replace  istdrate=0.47412 if age==2 & calendar==1
replace  istdrate=0.80277 if age==2 & calendar==2
replace  istdrate=1.05824 if age==2 & calendar==3
replace  istdrate=1.25469 if age==2 & calendar==4
replace  istdrate=0.73136 if age==3 & calendar==1
replace  istdrate=1.55946 if age==3 & calendar==2
replace  istdrate=2.33029 if age==3 & calendar==3
replace  istdrate=2.90461 if age==3 & calendar==4
replace  istdrate=0.73207 if age==4 & calendar==1
replace  istdrate=1.63585 if age==4 & calendar==2
replace  istdrate=2.85724 if age==4 & calendar==3
replace  istdrate=4.22945 if age==4 & calendar==4
gen iedeaths =  istdrate*personyrs/1000
```

To tabulate observed and expected numbers for arsenic exposure levels:

```
. table  periodhire arsenic if  periodhire==1, c(sum  resp_ca sum  iedeaths sum
personyrs) col format(%9.1f)
```

To determine the SMRs for each level of arsenic exposure, exponentiate the regression cooefficents:

```
. bysort  arsenic: poisson   resp_ca  if  periodhire==1, exp(iedeaths)
. di exp(.8650619)*100
. di exp( 1.752841)*100
. di exp( 1.551161)*100
. di exp( 2.034669)*100
```

To determine relative risks for arsenic exposure levels by defining the exposure as the expected number of events from the external data:

```
. xi: poisson   resp_ca  i.arsenic if  periodhire==1, exp(iedeaths) ir
```

These results are tabulated below:

Table 4.4: Dose-response analysis of respiratory cancer deaths among Montana smelter workers, based on external standardisation.

	Cumulative years of moderate/heavy arsenic exposure				
Workers employed before 1925	**< 1 year**	**1–4 years**	**5–14 years**	**15+ years**	**Total**
No of observed deaths	51	17	13	34	115
No. of expected deaths (standard population)	21.5	3.0	2.8	4.4	31.6
SMR (%)	237.5	577.1	471.7	765.0	
Relative risk (ratio of SMRs)	1.00	2.43	1.99	3.22	

The relative risks obtained using external standardisation are similar to those obtained from internal standardisation (and similar to those obtained by Breslow and Day[3]) (Table 4.2). However, external standardisation gives values that are slightly more extreme with a steeper dose-response relationship (for the highest exposure category the relative risk is 3.10 for internal standardisation vs 3.22 for external standardisation).

4.8 Comparing Methods 1 and 2

Internal standardisation by Method 1 can avoid the problems caused by the non-comparability of external standard rates, and its simplicity can provide an advantage over Method 2. However, while the ability of multivariate modelling to accommodate the internal estimation of baseline rates is desirable, incorporation of external rates may be advantageous in some circumstances.[3] One advantage is that it provides an overall measure of how the baseline cohort rates compare with those for the general population. If the external standard population chosen is commonly used, it may allow comparisons with other studies.

Method 2 is generally more complicated to perform correctly, and requires a more thorough critical appraisal of the multivariate model. For indirect standardisation, non-comparability of external standard rates may mean that the relative risk estimates (i.e. ratios of the SMRs using the first exposure level as baseline) for different exposure groups will fail to summarise adequately the stratum-specific rate ratios.[3] However, the process of model fitting encourages the investigator to evaluate the assumptions of proportionality that are essential in order that the estimated parameters have the intended interpretation.[3] The usual goodness-of-fit machinery may be applied to validate these assumptions. Additional interaction terms may also be incorporated into the model to account for confounding by any of the stratification variables. In instances where confounding has occurred as a result of using external standard rates, incorporating interaction terms to adjust for age, year and age x year interactions in externally estimated models can yield identical results to internally estimated models (see Section 4.8 of Breslow and Day).[3]

4.9 Conclusion

For multivariate analysis, it is valid to use either internal standardisation (where age-adjustment occurs through weighting by a pooled combination of all study groups) or external standardisation (where age-adjustment occurs using an external standard population) by either direct or indirect standardisation. If external standardisation is the method chosen, then the covariates in the model must also be adjusted for age, in order to yield unbiased estimates of the parameters of the model. The methods of internal and external standardisation have different strengths and weaknesses. The simplicity of internal standardisation (the analysis of crude rates with age as a covariate), can make this method more practical to use.

4.10 References

1. Rosenbaum PR, Rubin DB. Difficulties with regression analyses of age-adjusted rates. Biometrics 1984; 40(2): 437–443.
2. Selvin S. Practical Biostatistical Methods. USA; Duxberry Press, 1995.
3. Breslow NE, Day NE. ‘Comparisons among exposure groups’ and ‘Fitting models to grouped data’. In Statistical methods in cancer research. Volume II — the design and analysis of cohort studies. Lyon; International Agency for Research on Cancer, 1987.
4. Australian Bureau of Statistics 2003. Population by age and sex, Australian states and territories, 2001 Census edition—Final. Canberra: ABS (Cat. No. 3201.0).
5. Datasets from Breslow, N.E. and Day, N.E. Statistical Methods in Cancer Research: Vol I (1980) The Analysis of Case-Control Studies, and Vol II (1987) The Design and Analysis of Cohort Studies. Available at: http://faculty.washington.edu/norm/datasets.html (accessed October, 2004).

Appendix

Appendix 1: Lower and upper 95% confidence limit factors for a death rate (age-adjusted, crude or age-specific) based on a Poisson variable of 1 through to 99 deaths (D_i)

D_i	Lower confidence factor	Upper confidence factor	Di	Lower confidence factor	Upper confidence factor
1	0.02532	5.57164	51	0.74457	1.31482
2	0.12110	3.61234	52	0.74685	1.31137
3	0.20622	2.92242	53	0.74907	1.30802
4	0.27247	2.56040	54	0.75123	1.30478
5	0.32470	2.33367	55	0.75334	1.30164
6	0.36698	2.17658	56	0.75539	1.29858
7	0.40205	2.06038	57	0.75739	1.29562
8	0.43173	1.97040	58	0.75934	1.29273
9	0.45726	1.89831	59	0.76125	1.28993
10	0.47954	1.83904	60	0.76311	1.28720
11	0.49920	1.78928	61	0.76492	1.28454
12	0.51671	1.74680	62	0.76669	1.28195
13	0.53246	1.71003	63	0.76843	1.27943
14	0.54671	1.67783	64	0.77012	1.27698
15	0.55969	1.64935	65	0.77178	1.27458
16	0.57159	1.62394	66	0.77340	1.27225
17	0.58254	1.60110	67	0.77499	1.26996
18	0.59266	1.58043	68	0.77654	1.26774
19	0.60207	1.56162	69	0.77806	1.26556
20	0.61083	1.54442	70	0.77955	1.26344
21	0.61902	1.52861	71	0.78101	1.26136
22	0.62669	1.51401	72	0.78244	1.25933
23	0.63391	1.50049	73	0.78384	1.25735
24	0.64072	1.48792	74	0.78522	1.25541
25	0.64715	1.47620	75	0.78656	1.25351
26	0.65323	1.46523	76	0.78789	1.25165
27	0.65901	1.45495	77	0.78918	1.24983
28	0.66449	1.44528	78	0.79046	1.24805
29	0.66972	1.43617	79	0.79171	1.24630
30	0.67470	1.42756	80	0.79294	1.24459

Continued

Appendix 1 (continued): Lower and upper 95% confidence limit factors for a death rate (age-adjusted, crude or age-specific) based on a Poisson variable of 1 through to 99 deaths (D_i)

D_i	Lower confidence factor	Upper confidence factor	D_i	Lower confidence factor	Upper confidence factor
31	0.67945	1.41942	81	0.79414	1.24291
32	0.68400	1.41170	82	0.79533	1.24126
33	0.68835	1.40437	83	0.79649	1.23965
34	0.69253	1.39740	84	0.79764	1.23807
35	0.69654	1.39076	85	0.79876	1.23652
36	0.70039	1.38442	86	0.79987	1.23499
37	0.70409	1.37837	87	0.80096	1.23350
38	0.70766	1.37258	88	0.80203	1.23203
39	0.71110	1.36703	89	0.80308	1.23059
40	0.71441	1.36172	90	0.80412	1.22917
41	0.71762	1.35661	91	0.80514	1.22778
42	0.72071	1.35171	92	0.80614	1.22641
43	0.72370	1.34699	93	0.80713	1.22507
44	0.72660	1.34245	94	0.80810	1.22375
45	0.72941	1.33808	95	0.80906	1.22245
46	0.73213	1.33386	96	0.81000	1.22117
47	0.73476	1.32979	97	0.81093	1.21992
48	0.73732	1.32585	98	0.81185	1.21868
49	0.73981	1.32205	99	0.81275	1.21746
50	0.74222	1.31838			

Table from: Anderson RN, Rosenburg HM. Age standardisation of death rates: implementation of the year 2000 standard. National Vital Statistics Report 1998; 47(3). Hyattsville, Maryland: National Center for Health Statistics. 1998.