

# Estimating detailed distributions from grouped sociodemographic data: ‘get me started in’ curve fitting using nonlinear models

Paul Norman · Alan Marshall · Chris Thompson ·  
Lee Williamson · Phil Rees

Published online: 3 April 2012  
© Springer Science & Business Media B.V. 2012

**Abstract** In much demographic analysis, it is important to know how occurrence-exposure rates or transition probabilities vary continuously by age or by time. Often we have coarse or fluctuating data so there can be a need for estimation and smoothing. Since the distributions of rates or counts across age or another variable are often curved, a nonlinear model is likely to be appropriate. The main focus of this paper is on the estimation of detailed information from grouped data such as age and income bands; however, the methods we outline could also be applied to other settings such as smoothing rates where the original data are ragged. The ability to carry out curve fitting is a very useful skill for population geographers and demographers. Curve fitting is not well covered in statistics textbooks, and whilst there is a large literature in journals thoroughly discussing the detail of functions which define curves, these texts are likely to be inaccessible to researchers who are not specialists in mathematics. We aim here to make nonlinear modelling as accessible as possible. We demonstrate how to carry out nonlinear regression using SPSS, giving stepped-through hypothetical and research examples. We note other software in which nonlinear regression can be carried out, and outline alternative methods of curve fitting.

**Keywords** Sociodemographic data estimation · Curve fitting · Nonlinear models · Census and social survey data

---

P. Norman (✉) · A. Marshall · C. Thompson · P. Rees  
School of Geography, University of Leeds, Leeds LS2 9JT, UK  
e-mail: p.d.norman@leeds.ac.uk

L. Williamson  
School of Geography and Geosciences, University of St. Andrews,  
St. Andrews KY16 9AL, UK

## Introduction

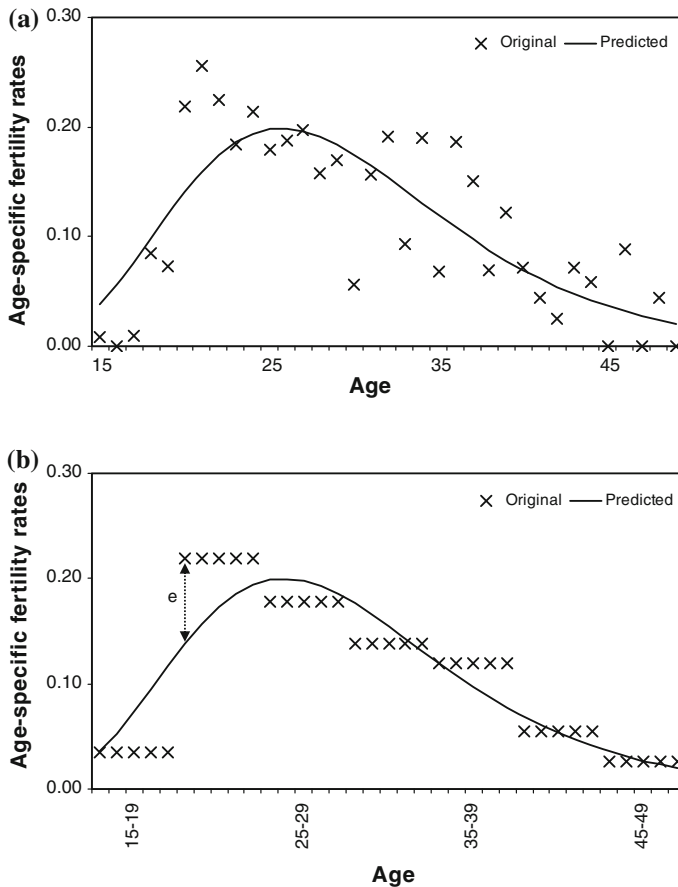
In relation to a demographic schedule, the distribution of rates across age, Keyfitz (1982), Congdon (1993) and Wilson (2010) usefully summarize the circumstances in which the graduation of schedules has become established. These include: smoothing, when rates are ragged or when age heaping has occurred through age misreporting; estimation, where data are missing or unreliable; comparative analysis and data reduction, by replacing full schedules with a smaller number of parameters; and projection, whereby the shape of a curve informs a forecasting model. The focus of this paper overlaps these situations: the estimation of detailed information from grouped data. Relevant circumstances in which it may be necessary to carry out the modelling of data include

1. The estimation of single-year-of-age rates or counts when only grouped age information is available. The outputs of this would then inform an annual series of a population estimate or projection so that the age-time plan of the demographic model matches that of the rates input;
2. The harmonization of categories so that data are comparable if the groupings in which a variable has been released are either inconsistent from different sources or change over time from the same supplier;
3. The matching of numerators and denominators when they have been released with incompatible categorizations. For example, age-groupings may need to be different from the standard 5-year bands around school-leaving age to calculate employment rates.

The distributions of rates or counts across age or another variable are often curved (e.g. concave, convex, exponential growth or decay, sigmoidal), and nonlinear models can be used to provide estimates of rates across the distribution in such situations.

Many sociodemographic data sources such as censuses and social surveys include variables which have grouped data which are an aggregate of more detailed information. Examples include age-groups and income bands. Data are often grouped (banded, binned, categorized) before release because more detailed information may risk breaching the confidentiality of respondents or because the data supplier has aggregated the information into application-relevant groupings. There may, though, be some situations wherein data are needed in more detailed units than the grouped data or when different aggregations are needed from those previously published.

As an example, Fig. 1a shows original age-specific fertility rates by single year of age which are based on sparse data (Williamson and Norman 2011). The rates are derived from a small population and are ragged, but the relationship between age and rate is clearly not linear. The predicted values in Fig. 1a have been estimated using a nonlinear model. The modelling predicts a plausible set of rates even when some of the original observations are zero. Figure 1b demonstrates the situation on which this paper concentrates. Here the original data have been expressed as 5-year age-specific fertility rates (15–19, 20–24 ... 45–49) and represent a common way in which data are available from suppliers. A nonlinear model has been used to



**Fig. 1** Modelling fertility rates using nonlinear regression, Bangladeshi women living in 'urban deprived industrial areas' in Bradford, West Yorkshire, 1991. **a** Original and predicted rates based on single year of age information. **b** Original and predicted rates based on 5-year grouped age information. *Note:* Authors' calculations based on Bradford Birth Statistics Database (after Williamson and Norman 2011)

estimate predicted fertility rates by single year of age (15, 16, 17 ... 47, 48, 49). The predicted rates based on the original single-year-of-age data and those derived from the 5-years-of-age information are very close in value.

Wilson (2010) very usefully demonstrates the modelling of migration schedules in Excel and is right to suggest that the spreadsheet environment is accessible and within the experience of people involved in demographic research. Given its widespread use, here we use SPSS (a.k.a. PASW) to demonstrate curve fitting using nonlinear models in as straightforward and nontechnical a manner as possible. Later we note other statistical software which can be used.

The terminology in this topic area can be confusing. An internet search on 'curve fitting', 'curve estimation', 'graduation' and 'smoothing' will reveal their use in a wide variety of methods and applications. Benjamin and Pollard (1980, pp. 239–242) provide very useful discussion about the terms 'graduation' and

‘smoothing’ which they find have widespread and often undefined usage. Below, we use the term ‘nonlinear regression’ since this is the term used in software such as SPSS and because the models are expressing the relationship between two variables using a nonlinear function where parameters are estimated by a regression technique, minimizing the errors between predicted and observed values.

Regarding nonlinear regression, Congdon (1993) notes that graduation using a parametric formula over an entire curve, as a technique can be contrasted with nonparametric graduation (such as the use of kernel density methods) and graduation using spline functions. We focus here on nonlinear regression and note its main advantage over nonparametric graduation which is the potential for substantive interpretability of parameters and the associated scope for comparison of curves over time and place (Congdon 1993; Ratkowski 1983). If the only aim of a curve fitting exercise is to obtain a good fit for the purposes of representation, then nonparametric graduation or other techniques may be preferred.

So the task here is to demonstrate how to estimate information for each unit of a grouped variable. The detailed estimates may then be used ‘as is’ or re-aggregated into application-relevant groups. After an explanation of carrying out nonlinear regression in SPSS, three research examples are explored below. First, the estimation of fertility rates by single years of age, second the estimation of single-year mortality probabilities from an abridged life table and third, the harmonization of a time-series of information on expenditure in different supermarkets which has been categorized in different ways over time.

## Where do you start?

The starting point is a dataset with grouped data which needs to be disaggregated to unit level. We assume that underlying the grouped data a more detailed distribution exists, and for many ordinal variables this will be the case. Whilst we are focusing here on the disaggregation of grouped data, the process is equivalent when data are already in the desired units but are in need of smoothing, as in Fig. 1a. A scatterplot of the relationship between the independent ( $x$ ) and dependent ( $y$ ) variables will therefore give a guide to the shape of your data. As with many modelling situations, there is as much art here as science.

To run a nonlinear regression model you need to select a formula to act as a model expression. The formula includes some parameters the values of which the model will estimate. You may need to provide the model with some starting values for these parameters, though some statistical packages will set default starting values. Table 1 describes and illustrates some curves likely to be relevant to sociodemographic data. The associated model expressions are defined. The parameters here are given descriptions to aid their use in this type of application. The parameter values initially provided are indicative about aspects of the data distribution, so knowledge of the maximum, minimum and median values will help inform the setting of initial parameter values which are in the right ‘ball park’ for the data. Similarly, the proportional change between low and high values can give a useful indication of an initial rate of change. When the model has run, the estimates

Table 1 Shapes of curves and nonlinear model expressions

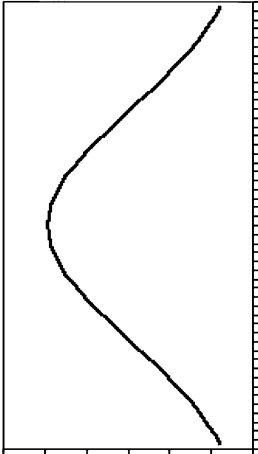
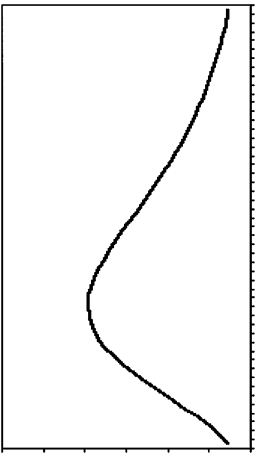
Description of curve	Shape of curve	Model expression (algebraic version)	Model expression (SPSS syntax version)
1. Curves which are symmetrical with the peak in the middle of the distribution and an even ascent and descent to the extremes of the distribution		$f(x) = \left( h \times e^{\left( \frac{-(x-p)^2}{r} \right)} \right)$	$(h * \exp(-(x - p)**2/r))$ where $x$ = the increments (e.g. age), $h$ = height of the curve, $p$ = position on the $x$ axis, $r$ = rate of ascent and descent, $\exp$ = exponential function
2. Curves with a peak towards the lower end of the distribution with a steeper ascent from the low end of the distribution up to the peak and a slower descent to the high end		$f(x) = (h \times e^{(-r_d \times (x - p) - e^{(-r_a \times (x - p)))})})$	$(h * \exp(-r_d * (x - p) - \exp(-r_a * (x - p))))$ where $r_a$ = rate of ascent, $r_d$ = rate of descent

Table 1 continued

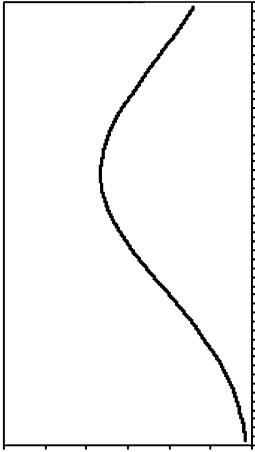
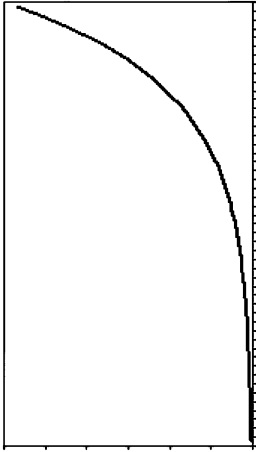
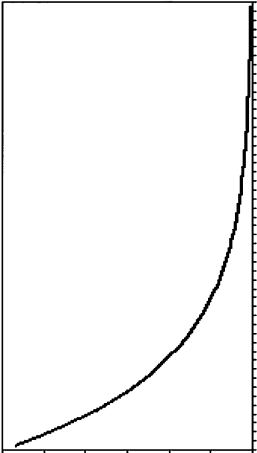
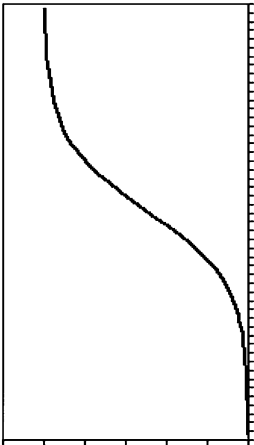
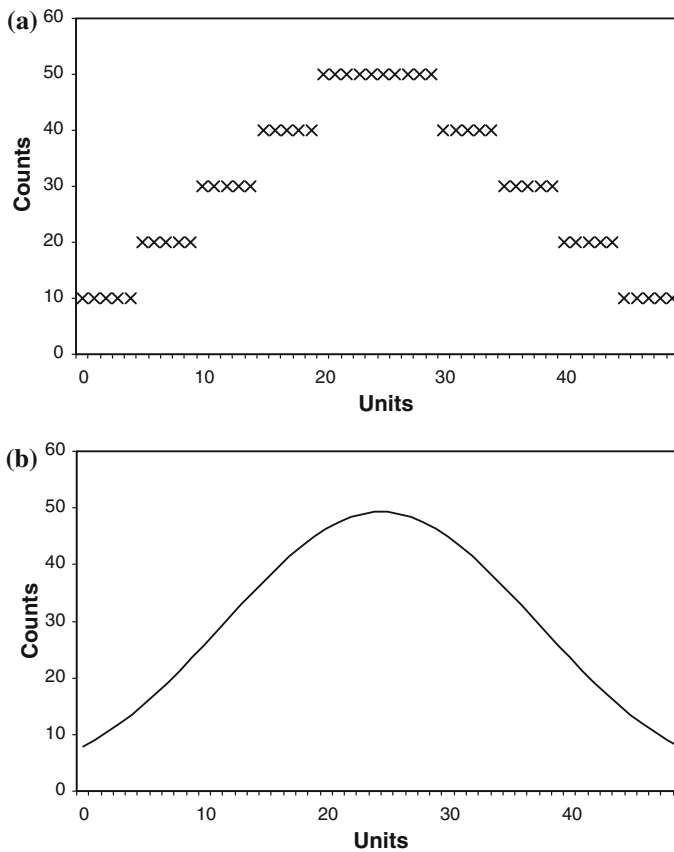
Description of curve	Shape of curve	Model expression (algebraic version)	Model expression (SPSS syntax version)
3. Curves with a peak towards the higher end of the distribution with a slower ascent from the low end of the distribution up to the peak and a steeper descent to the high end		$f(x) = (h \times e(-r_d \times (x - p)) - e(-r_a \times (x - p)))$	$(h * \exp(-r_d * (x - p)) - \exp(-r_a * (x - p)))$
4. From low at to high with a gradually steepening rate of ascent with the peak at the highest values		$f(x) = (h \times e(r \times x))$	$(h * \exp(r * x))$

Table 1 continued

Description of curve	Shape of curve	Model expression (algebraic version)	Model expression (SPSS syntax version)
5. Curves from a peak at the low end of the distribution falling to a low at the high end		$f(x) = (h \times e(-r \times x))$	$(h * \exp(-r * x))$
6. Growth curves with a 'sigmoidal' shape (whereby the units (x) are often time) Variations in functions relate to flatness of the tails and steepness of the middle of distribution		<p>Growth curve examples:</p> <p><math>f(x) = a/(1 + e(b + c \times x))</math> (Logistic)</p> <p><math>f(x) = a \times e(-e(b - c \times x))</math> (Gompertz)</p> <p><math>f(x) = a + (b - a) \times e(-c \times x^d)</math> (Weibull)</p>	<p><math>a/(1 + \exp(b + c * x))</math></p> <p><math>a * \exp(-\exp(b - c * x))</math></p> <p><math>a + (b - a) * \exp(-c * x^{**d})</math></p>

Note: \*\* 2 means raise to the power of 2 in SPSS syntax

Note: Units of interest are on the x axis  
Counts or rates on the y axis



**Fig. 2** Distribution of grouped and estimated unit level data. **a** Grouped data distributed evenly across within-group units. **b** Data predicted using nonlinear regression for unit values

of the parameters which are output may have an interpretable relationship relevant to the phenomenon of interest.

Here we take data grouped in fives as an example. A section of the data is in Table 2. You need to spread the grouped data across the within-group units. In this example, we have taken the total for the group and divided by five and these are taken as initial values. If you had some information from another source on the within-group distribution (e.g. national-level data applied to subnational areas), you could use this as a first-step estimation. In itself, this can be used to disaggregate grouped information but there tend to be discontinuities at group boundaries. The shape of the 5-year grouped information graphed in Fig. 2 suggests that curve 1 and associated model expression in Table 1 are likely to represent the distribution well.

In the SPSS data editor, you would have variables with the units and the initial values. It is possible to run a nonlinear regression in SPSS using menu selection and dialogue boxes (Analyse > Regression > Nonlinear ...) but this route requires



**Table 2** Example grouped data to be disaggregated to values for each unit

Group	Original	Units	Initial	Predicted
0–4	50	0	10	7.79
		1	10	9.03
		2	10	10.40
		3	10	11.91
		4	10	13.55
5–9	100	5	20	15.32
		6	20	17.22
		7	20	19.23
		8	20	21.35
		9	20	23.56
10–14	150	10	30	25.83
		11	30	28.15
		12	30	30.49
		13	30	32.83
		14	30	35.12
15–19	200	15	40	37.35
		16	40	39.47
		17	40	41.46
		18	40	43.29
		19	40	44.91
etc.	etc.	etc.	etc.	etc.

somewhat fiddly selections to be made. Here we use SPSS syntax since for this procedure it is more efficient and enables the parameters to be specified and changed more readily than by using dialogue boxes. The variables in the dataset are ‘unit’ and ‘curve1’. Explanations of this SPSS syntax and the commands below are at [Appendix](#)

```

MODEL PROGRAM h=50 p=25 r=0.5 .

COMPUTE curve1_pr = (h * exp( - (unit - p)**2 / r)) .

NLR curve1

/PRED curve1_pr

/SAVE PRED

/CRITERIA ITER 1000 .

TSLOT VARIABLES = curve1 curve1_pr

/ID= unit

/NOLOG .

```

The predicted values in the SPSS data editor can be used as single units, if that suits the application, or aggregated into desired groups. Note that the sum of the predicted values should be the same as the sum of the original values. However, there tend to be slight differences which can be recovered by constraining the sum of the predicted values to agree with the sum of the original values should this be critical.

The SPSS output includes the estimated parameter values. If the research need is to be able to predict values or to use the parameters in place of a full schedule, the estimated parameters ( $h = 49.29$ ;  $p = 25.50$ ;  $r = 325.45$ ) from the model of curve 1 can be used as follows. If the unit value of interest is 25, then using the model expression ( $h * \exp(-(unit - p)**2/r)$ ) becomes ( $49.29 * \exp(-(25 - 25.50)**2/325.45)$ ). The value predicted by the model is 49.32.

The suitability of a model can be assessed in: (a) statistical; (b) visual, (c) interpretive; and (d) practical ways.

### Statistical

The SPSS output includes an  $R^2$  value of how well the modelled curve fits the original data. In Fig. 1b the distance between one of the original data points and the estimated curve is marked  $e$ . This represents the error or residual which may be positive, negative or zero. The estimation of the parameters in a nonlinear regression involves an iterative process whereby an algorithm repeatedly adjusts the curve being fitted so that the sum of the squares of the errors is reduced at each step. In the SPSS Output window, the iteration history reports the successive parameter values used in the model and the 'residual sum of squares' (the sum of the squared differences between the original and predicted data points). The  $R^2$  value for the overall fit of the final model is calculated as  $1 - (\text{Residual Sum of Squares})/(\text{Total Sum of Squares})$ . The total sum of squares is the squares of the distances between the observed data points and their mean; in the SPSS output, the total sum of squares is named the 'Corrected Sum of Squares'. When graduating the unit data grouped into fives for curve 1 from Table 1, the  $R^2$  value is  $0.962 = 1 - (381.921)/(10,000)$ . SPSS reports the 95 % confidence intervals of the parameter estimates. To be significant at the 95 % level, this confidence interval should not include zero.

If the fits of different curve functions are being compared to see which performs best, in addition to  $R^2$  values further tests can be carried out on outputs to determine the level of error:

$$\text{Mean Absolute Error(MAE)} = \sum |y_i - \hat{y}_i|/n \quad (1)$$

$$\text{Mean Squared Error(MSE)} = \sum (y_i - \hat{y}_i)^2/(n - p) \quad (2)$$

where  $n$  = the number of observations and  $p$  = the number of model parameters.

As an example, the asymmetrical curve 2 in Table 1 has a four-parameter function defined with explicit allowance for different rates of ascent and descent. Fitting this model to 5-unit grouped data returns a reported  $R^2$  value of 0.918 with MAE of 2.70 and MSE of 12.54. If the simpler three-parameter function (used for

curve 1 above) is fitted, the model converges to a solution, but the  $R^2$  value is lower at 0.846 and the error measures are larger with MAE of 3.94 and MSE of 23.07.

It is invariably the case that a more complicated model will fit the data better (i.e. have a lower residual sum of squares) than a simpler one. However, parsimony of the number of parameters is generally preferred, as simple models are statistically more stable and offer a better basis for comparison over time and place (Coale and Trussell 1996; Congdon 1993). An F test can be used to determine whether the improvement in model fit associated with a more complex model reflects a better representation of the underlying distribution being modelled, or a quirk of the sample with which we are working (Motulsky and Christopoulos 2004). So, whilst above, the four-parameter model appears to give a better fit than the three-parameter model, we can assess whether the improvement in model fit associated with the more complicated model is statistically significant. The F test is the ratio of the relative improvement in fit (through the decrease in residual sum of squares) to the relative decrease in loss of degrees of freedom (through the increase in number of parameters) and is defined as

$$F = \frac{(SS1 - SS2)/SS2}{(DF1 - DF2)/DF2} \quad (3)$$

where  $SS$  is the residual sum of squares and  $DF$  is the degrees of freedom in a simpler model (1) and a more complex model (2).

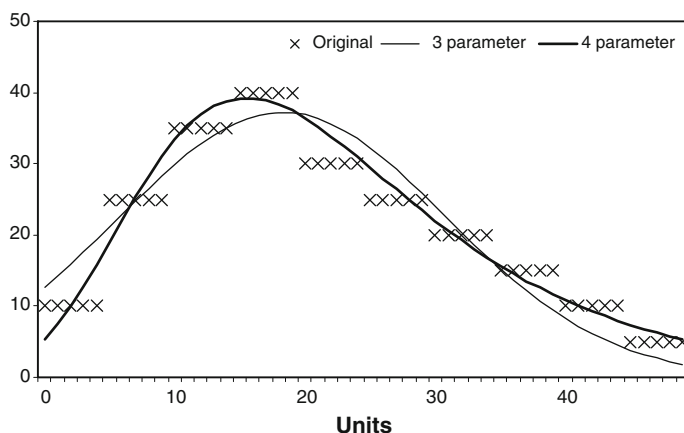
If the resulting  $F$ -ratio is near 1 then the simpler model is acceptable. If the ratio is greater than 1 and when compared with an F distribution is significant ( $p < 0.05$ ), then the more complex model can be taken as a justifiable improvement. The F ratio for the three and four parameter models fitted above to the asymmetric curve is calculated as

$$49.88 = \frac{(798.04 - 382.87)/382.87}{(47 - 46)/46}$$

Since the  $F$ -ratio is greater than 1 and when compared with an F distribution the  $p$  value is less than 0.05, we can conclude that the more complex model is a significantly better fit.

## Visual

A visual inspection of the modelled curves compared with the original observations provides a good indication of how well the model represents the data. A smooth curve will be estimated, but systematic sets of adjacent positive or negative residuals may indicate that experimentation with a different function would be advised. Figure 3 illustrates the predicted values from the three and four-parameter models on the asymmetric grouped data. Between unit ( $x$ ) values 20 and 32 successive predictions are overestimated by the three-parameter model, and then between unit values 36 and 50 successive predictions are underestimated. The curve estimated using the four-parameter model does not have this degree of systematic under or overestimation of the predicted values.



**Fig. 3** Comparing curves estimated using different models

### Interpretive

If the literature points towards an interpretation of the relationship between the estimated parameters, then checks for consistency can be made. We give specific examples relating to fertility and mortality below.

### Practical ways

In practical terms, if demographic rates are being estimated, when you apply those rates to the population at risk, it is worthwhile checking whether the number of events calculated to occur is close to the observed number of events. An example using births data is referred to below.

So, having established how to carry out nonlinear regression, we now give some research examples which use the technique.

## Research examples

### Estimating ethnic-group fertility rates by single year of age from 5-year rates

In Fig. 1 we provided an example from Williamson and Norman (2011) on the smoothing of fertility rates from fluctuating single-year-of-age rates. This research used maternity data for 1991 on births for local areas within Bradford, West Yorkshire, and is challenging because of small numbers of births and populations; hence the need for modelling the calculated rates. A more usual source on fertility is the published data from the Vital Statistics which record births by age of mother in 5-year groups for local-authority district geographies but with no information on ethnic group. Norman et al. (2010) and Rees et al. (2011) estimate age-specific fertility rates by 5-year age-groups in 2001 and then interpolate these to single-year-of-age rates

for use in a projection model. Here we provide examples for the White ethnic group; for the Pakistani group, the largest minority group in Bradford who traditionally have high fertility rates; and the Chinese, a smaller group with lower fertility.

A fertility curve has similarities to curve 2 in Table 1 but has a more complex shape. The model expression used here in a nonlinear regression of fertility is the Hadwiger function (Chandola et al. 1999; Hadwiger 1940):

$$f(x) = \frac{ab}{c} \left(\frac{c}{x}\right)^{3/2} \exp\left\{-b^2\left(\frac{c}{x} + \frac{x}{c} - 2\right)\right\} \quad (4)$$

where  $x$  = the age of mother at birth; and  $a$ ,  $b$  and  $c$  are the parameters to be estimated.

Within a nonlinear regression, Eq. 4 above is defined in SPSS syntax as  
MODEL PROGRAM a=1 b=4 c=28 .

```
COMPUTE rate_Pr = (a*b/c)*(c/age)**(3/2)*exp(-b*b*(c/age+age/c
- 2)) .
```

The advice in the literature is that the initial values of the model parameters  $b$  and  $c$  should be set at the maximum likely fertility level and the median age in the distribution. The variable 'rate\_Pr' is the predicted fertility rate which the model will output.

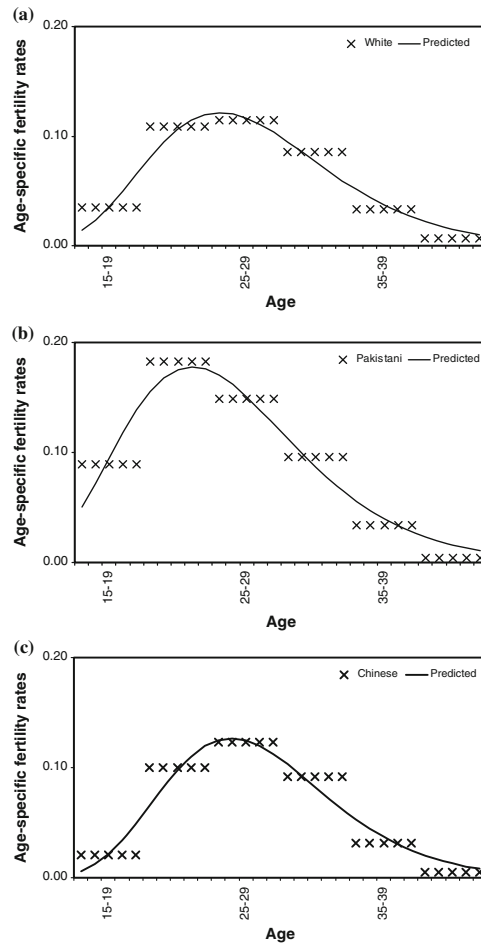
The original 5-year rates and predicted single-year-of-age rates are illustrated in Fig. 4; Total Fertility Rates (TFRs), the sum of the age-specific rates, are a regularly used demographic summary measure. For the rates modelled here, the TFRs are the area under the curve. The resulting TFRs based on the 5-year rates and the single-year-of-age rates are very close: White 1.92 and 1.93; Pakistani 2.77 and 2.79; and Chinese 1.86 and 1.87.

Whilst in general the estimated parameters in a nonlinear regression may not necessarily have any real meaning, parameter  $a$  is found to relate to overall fertility (Chandola et al. 1999). This is the case here with values of 0.22, 0.33 and 0.21 respectively for the White, Pakistani and Chinese groups. Similarly, parameter  $c$  reflects the age at which the fertility rates peak. Here the Pakistani group has a relatively young curve with  $c$  estimated as 25.8. The Chinese curve is somewhat older with  $c$  estimated as 28. The White group curve peak and  $c$  value of 27.6 fall in between.

For each ethnic group, we can apply the modelled rates to the populations at risk of giving birth to give the number of births which would result at each age. We then sum these across all childbearing ages. Regarding the accuracy of the predicted number of births, we can compare with the registered births for Bradford district but without an ethnic breakdown. This reveals that the total number of births in Bradford is predicted to within 1 %.

### Estimating mortality probabilities by single year of age from abridged life table rates

Some curves have distinctive subsections which need explicit handling and this is the case for mortality curves. Following Heligman and Pollard (1980), and the life



**Fig. 4** Modelling age-specific fertility rates by ethnic group, Bradford, West Yorkshire in 2001. **a** Original and predicted rates for the White ethnic group. **b** Original and predicted rates for the Pakistani ethnic group. **c** Original and predicted rates for the Chinese ethnic group. *Note:* Authors' calculations based on 2001 Census, Labour Force Survey and Vital Statistics data (after Norman et al. 2010; Rees et al. 2011)

table probabilities  $q$  of dying between exact ages  $x$  and  $x + 1$ , Congdon (1993) identifies several features of observed mortality which a modelled curve should capture: (i) a high death rate in the first year of life which declines rapidly; (ii) an 'accident hump' for young adults, especially males; and (iii) an accelerating chance of death with increasing age. These components are represented in a full model curve:

$$q(x) = A^{(x+B)^C} + D \exp\left\{-E(\ln x - \ln F)^2\right\} + GH^x / (1 + KGH^x) \quad (5)$$

where,  $A$ ,  $B$ ,  $C$ ,  $D$ ,  $E$ ,  $F$ ,  $G$ ,  $H$  and  $K$  are the parameters to be estimated.

The expansion of an abridged life-table into single-year-of-age information can be achieved using nonlinear regression, and Eq. 5 above is defined in SPSS syntax as

```
MODEL PROGRAM A=1 B=1 C=1 D=1 E=10 F=20 G=1 H=1 K=2 .

COMPUTE dth_Pr = A**(age+B)**C + D*exp(-E*(ln(age)-ln(F))**2) +
G*H**age / (1+K*G*H**age) .
```

This can be applied to the probabilities of dying obtained from an abridged life table for males in England and Wales, 1991. Figure 5 illustrates the log of the original and modelled predicted probabilities. The original data were in the abridged life-table age groups of age 0, 1–4, 5–9 ... 85+ and the predicted probabilities are by single year of age. The overall modelled curve is the sum of the mix of the three functions which constitute Eq. 5 and thereby capture the different mortality features noted by Congdon (1993). Detailed discussion of the expansion of abridged life tables into single year of age detail can be found in Kostaki and Panousis (2001), Debón et al. (2005) and Ibrahim (2008) with the latter two papers discussing the demographic interpretation of the parameters A to H and the former the need for the parameter K. A very useful comparison of different techniques for expanding an abridged life table is provided by Kostaki and Panousis (2001).

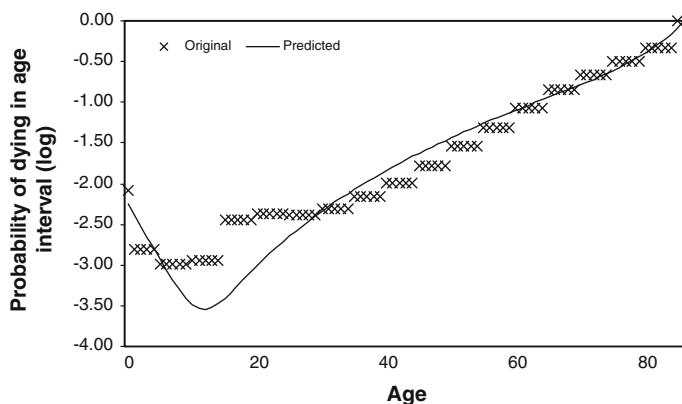
Mixture models, as described above, have widespread use in nonlinear regression modelling. To cope with bimodal distributions, mixture models of fertility curves characterized by relatively low fertility for women in their mid-20 s compared with younger and older women have been developed by Chandola et al. (1999) and Peristera and Kostaki (2007). Reasons for the bimodal fertility distributions can include married and unmarried women and women from different ethnic groups or of different educational achievement.

Mixture models are also used for migration schedules because different age groups within the population have very distinctive migration propensities. The standard model migration schedule (Rogers and Watkins 1987) is the sum of five components comprising curves for childhood, labour force, retirement and elderly populations plus a background constant. For some locations a student curve should be added in (Wilson 2010).

An important advantage of the mixture models discussed above is the interpretability of the parameters. This is particularly useful when comparing schedules of fertility, migration or mortality over time and place.

### Estimating comparable groupings of levels of expenditure in different supermarkets

Usually graduation in demographic analysis involves estimation of a particular characteristic across the age range. There is no reason why such analysis should be restricted to characteristics that follow a strong age pattern. We demonstrate here the potential for wider application of graduation techniques using an example



**Fig. 5** Expanding an abridged life-table for males in England and Wales, 1991 to single-year-of-age information. *Note:* Authors' calculations based on vital statistics and mid-year estimates

involving household grocery expenditure where a curve is fitted to a strong expenditure, rather than age, pattern.

Since the start of the economic downturn in the UK, there has been a need in social science research to undertake reliable time-series analysis to understand the major impacts of the recession. However, for time-series analysis to be reliably conducted, the data being used must have been recorded in a consistent manner over the course of any given time frame. If, for example, variables are categorized in different ways, it makes it difficult to carry out an informative trend analysis. Consequently, we provide a guide for dealing with these inconsistencies through the use of nonlinear regression, using grocery expenditure data for a set of food retailers as an example. Household expenditure is a topical area of study with a growing literature on recent changes in consumption (Leaker 2009; Melvin and Taylor 2009; Mitchell 2009; Thompson et al. 2010b; Vaitilingam 2009), as academics and industry professionals aim to track ongoing trends in the UK economy.

We make use here of grocery expenditure data from the annual Research Opinion Poll (ROP), a commercial lifestyle survey compiled by Acxiom Ltd to capture detailed information about household spending behaviour across Great Britain (GB). The survey is run twice a year in January and September capturing a total annual sample of over 1,000,000 households (Thompson et al. 2010a), a figure which dwarfs the sample of 6,500 households achieved by the government's Living Costs and Food survey (LCF). The ROP, however, is not a random sample. Despite being an excellent source on household expenditure, the ROP survey has undergone some changes over time whereby the bands for the grocery expenditure variable have been altered to include more detail for higher levels of weekly spending (Thompson et al. 2010a); so before time-series analysis can be performed, harmonization of the bands is needed.

Table 3 reveals two issues which require attention so that changes in household expenditure between 2005 and 2010 can be assessed. First, whilst there are six bands of expenditure in all years there is a change in the categories between 2008 and 2009. Second, the class intervals are not evenly spaced in the two versions of



the categories. This may not necessarily be critical but interpretation may well be more logically presented if the bands are of the same size.

Table 4 demonstrates the format of the Acxiom data, displaying the number of households which shop at each retailer cross-tabulated against their total weekly grocery expenditure. The data are for Yorkshire and the Humber region and represent four different retailer types based on their product quality and price range: Retailer A represents one of the more expensive food retailers; Retailer B caters for all ranges; Retailer C characterizes the deep discounters; and Retailer D corresponds to the convenience market. To prepare the data for the nonlinear regression model, the counts in Table 4 are converted into proportions. This is achieved by dividing the number of households in each spending category by the total number of households which shop at that supermarket. Next, the categorical data are disaggregated into single units (£s) of spending. We achieve this by dividing the newly calculated rates by the total number of units in that category. As the data are not equally banded like the fertility example above, care must be taken to divide the units correctly.

The expenditure data having been disaggregated into initial single-pound distributions, the data can be modelled using nonlinear regression. With regard to the model expression, the categorical data have an asymmetrical distribution so the formula used in Table 2 for curve 2 looks appropriate. This function performs well returning a realistic single-unit distribution. Alternative methods used for comparison include the symmetrical model expression (curve 1) and the cubic spline curve estimation function (see below). The model expression defined in SPSS syntax is as follows:

```
MODEL PROGRAM a=1 b=1 c=1 d=50 .
```

```
COMPUTE Retailer_estimate = (a*exp ( - c*(Spend - d) - exp( -  
b*(Spend - d)))) .
```

In this model,  $a$  = the height of the curve;  $b$  = the rate of ascent;  $c$  = the rate of descent;  $d$  = position on the x axis of the peak; and Spend = units of single £s.

The model outputs the predicted distribution of the grocery data in single pounds but, as noted above, there was a need to constrain the modelled outputs to sum to the original data. Figure 6 illustrates the original and estimated data for each of the retailer types in 2009. The curve estimation retains the shape of the original data,

**Table 3** Grocery spend by food retailer: changing spend categories over time in Acxiom's Research Opinion Poll

2005	2006	2007	2008	2009	2010
Under £16	Under £16	Under £16	Under £16	Under £36	Under £36
£16–£34	£16–£34	£16–£34	£16–£34	£36–£49	£36–£49
£35–£49	£35–£49	£35–£49	£35–£49	£50–£69	£50–£69
£50–£59	£50–£59	£50–£59	£50–£59	£70–£99	£70–£99
£60–£89	£60–£89	£60–£89	£60–£89	£100–£149	£100–£149
£90+	£90+	£90+	£90+	£150+	£150+

**Table 4** Grocery spend by food retailer: incompatible data over time

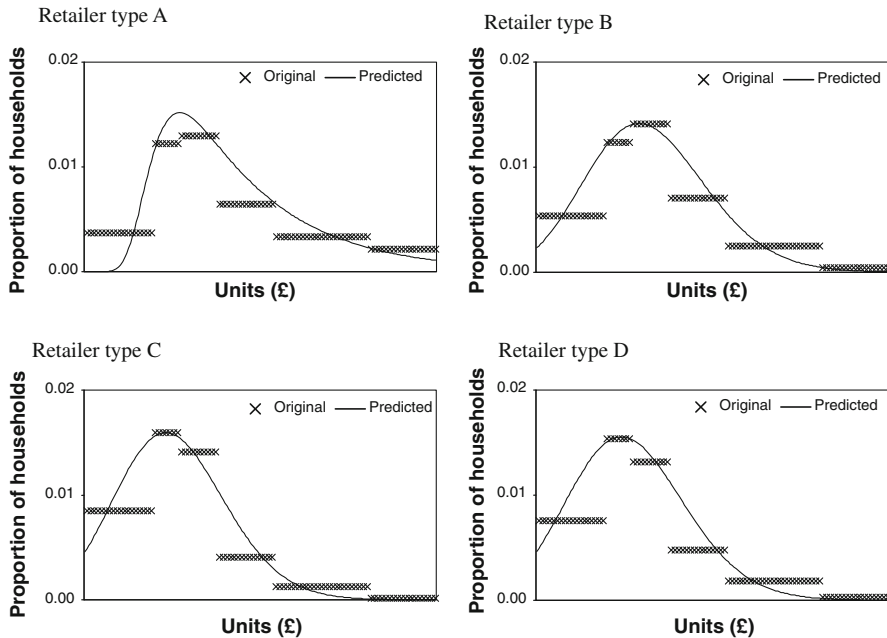
Spend by number of households	High end (A)	Middle (B)	Low end (C)	Convenience (D)	Total
(a) 2008 expenditure by retailer type					
Under £16	7	344	200	106	1,828
£16–£34	42	1,470	664	268	8,027
£35–£49	26	1,489	485	213	7,776
£50–£59	31	1,251	306	146	6,482
£60–£89	42	1,536	265	145	7,602
£90+	43	892	89	93	4,095
Total	191	6,982	2,009	971	35,810
(b) 2009 expenditure by retailer type					
Under £36	14	1,734	634	243	8,433
£36–£49	32	1,315	375	145	6,544
£50–£69	46	1,970	455	185	9,739
£70–£99	45	1,987	327	150	9,355
£100–£149	34	836	78	46	3,380
£150+	10	48	2	2	135
Total	181	7,891	1,870	771	37,585

Original data supplied by Acxiom

with each of the food retailers having distinctive curves which reflect the amount of money being spent. The curves in Fig. 6 show that the rise to the peak for the high-end supermarket (Retailer A) starts later and the curve stays higher to the right than both the discount supermarket (Retailer C) and convenience store (Retailer D). This would indicate, as we would expect, that households which shop at discounters and small convenience stores spend less on groceries than those shopping in more expensive supermarkets with Retailer B occupying the middle ground of expenditure (Thompson et al. 2010b).

To demonstrate change over time in household spending by each retailer type and to highlight any overall changes, we can aggregate the grocery expenditure by single pounds into even categories of spending. Here we opt for £25 groupings from £0–£25 up to £100 and over. Now that the 2008 and 2009 spending categories have been harmonized, it is possible to compare the expenditure in these two years. Table 5 and Fig. 7 show how overall expenditure has changed between 2008 and 2009 with reductions in the lowest two categories, a small rise in the middle, £50–74 bracket and relatively large rises for expenditure of £75 and over. The evidence suggests that household weekly expenditure on groceries has risen between 2008 and 2009 partly because of inflation.

The growth in expenditure on groceries is arguably a result of the recession which started at the beginning of 2008. Since the recession began, consumers have altered their shopping behaviour seeking both low cost and quality. People are eating out less and spending more on food, which has helped sustain growth in the grocery market when other sectors have declined (Mitchell 2009; Thompson et al.



**Fig. 6** Original and predicted rates based on grocery spend data in Yorkshire and the Humber, UK in 2009

2010b). Continuing work in this area will look at the complete 2005–2009 time-series including more up-to-date information as it becomes available.

## Further aspects

### Alternative nonlinear function fitting software

In addition to SPSS, it is possible to fit curves using nonlinear regression techniques through a number of different statistical packages: Stata has the command *nl* (Royston 1993; Stata 2011); for a practical example of the use of Stata to fit exponential curves to disability schedules see Marshall (2009); SAS has a procedure called *NLIN* (Freund and Littell 2000); for users of R, *nls* is the nonlinear regression command (Ritz and Streibig 2008); Minitab also allows users to carry out nonlinear regression: users can supply their own function or select from a list of named functions or from sample graphs illustrating the shape of curves (Minitab 2011); as demonstrated by Wilson (2010), Excel can also be used.

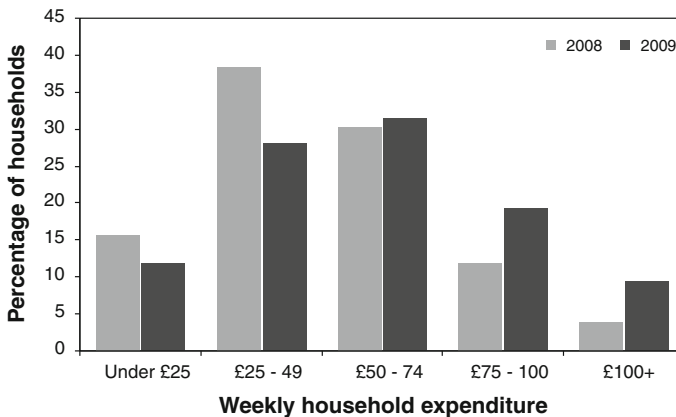
### Alternative curve fitting procedures

There are a number of alternatives to nonlinear regression for curve fitting and two are considered here: curve fitting using functions within linear regression and curve estimation using relational models.

**Table 5** Grocery spend by food retailer: compatible data over time

Spend by number of households	High end (A)	Middle (B)	Low end (C)	Convenience (D)	Total
(a) 2008 expenditure by retailer type					
Under £25	26	1,029	514	243	5,592
£25–49	57	2,601	887	388	13,701
£50–74	45	2,142	467	242	10,865
£75–100	28	892	119	79	4,275
£100+	36	318	23	19	1,376
Total	191	6,982	2,009	971	35,810
(b) 2009 expenditure by retailer type					
Under £25	2	924	366	141	4,438
£25–49	44	2,126	643	247	10,539
£50–74	56	2,405	538	221	11,849
£75–100	36	1,553	243	114	7,245
£100+	43	884	79	47	3,515
Total	181	7,891	1,870	771	37,585

Authors' calculations based on data supplied by Acxiom



**Fig. 7** Weekly household expenditure in categories which are compatible over time. *Note:* Authors' calculations based on data supplied by Acxiom

### Curve fitting using functions in linear regression

In this approach a linear regression is used to relate a set of demographic rates ( $y$ ) to a polynomial function of age ( $x$ ), where the highest power of  $x$  depends on the particular shape of relationship being modelled. A polynomial is the right hand side of the equations in Table 6. Thus, a regression involving a polynomial of age is linear in terms of the parameters which are estimated from the data and is a special

case of linear regression. In demographic applications polynomials have been used, for example, to smooth fertility rates (Brass 1960; Gage 2001; Gilje 1969; Hoem et al. 1981; Williamson 2007).

This approach can be carried out using menu commands in SPSS (Analyse > Regression > Curve estimation) and then the curves to be estimated as the Dependents and the units (such as age) as the Independent variable. A variety of models (curve shapes) can be selected with predicted values saved as new variables. These predictions can then be compared with values output from a nonlinear regression on the same curves, and assessed for model fits as described above. The SPSS output window contains tables which indicate the fit of the modelled curves to the data with the original and predicted values plotted on graphs to provide visual checks on the appropriateness of a particular curve shape. SPSS syntax can also be used, with the following fitting a quadratic curve (a parabola) to curve 1 from Table 1:

```
TSET MXNEWVAR = 1 .

CURVEFIT

/ VARIABLES = curvel WITH unit

/ CONSTANT

/ MODEL = QUADRATIC

/ PLOT FIT

/ SAVE = PRED .
```

Compared with nonlinear regression, disadvantages of the use of polynomials include difficulty in the interpretation of parameter values, and the fact that a change in the predicted curve direction can result at the extremes of the data distribution (McNeil et al. 1977).

### Curve estimation using relational models

Relational models comprise a reliable standard schedule of demographic rates and a mathematical rule that maps the standard to another schedule in a population where

**Table 6** Modelling curves in linear regression

Curve shape	Equation	Polynomial
Straight line	$y = a_0 + a_1x$	First degree
Quadratic/Parabola	$y = a_0 + a_1x + a_1x^2$	Second degree
Cubic	$y = a_0 + a_1x + a_1x^2 + a_1x^3$	Third degree
Quartic	$y = a_0 + a_1x + a_1x^2 + a_1x^3 + a_1x^4$	Fourth degree

information may be incomplete or unreliable (Preston et al. 2001). A key advantage of relational models of mortality is that the complexity of the mortality age pattern is captured in the standard schedule and a small number of parameters quantify the deviation from the standard. These models thus require fewer parameters than many mathematical mortality functions and can flexibly reproduce sets of model life tables using two suitably chosen parameters and a standard (Keyfitz 1982; Preston et al. 2001).

The relational approach was originally developed by Brass (1971) for the modelling of mortality schedules. The Brass model is based on a logit transformation of  $q(x)$  the probability of surviving to age  $x$ .

$$Y(x) = \frac{1}{2} \text{Logit}[q(x)] = \frac{1}{2} \ln \left[ \frac{q(x)}{1 - q(x)} \right] \quad (6)$$

The logit transformation of  $q(x)$  is valuable because the relationship between two logit mortality schedules turns out to be remarkably linear (Newell 1988). On the basis of this linear relationship, Brass proposed a simple relational formula to predict  $\hat{Y}(x)$  from the logit of  $q(x)$  in the standard population  $Y^s(x)$ :

$$\hat{Y}(x) = \alpha + \beta * Y^s(x) \quad (7)$$

Relational models have been developed for particular countries (e.g. for Peru by Kamara and Lamsana 2001) and have also been successfully extended to other demographic characteristics. For example, Brass (1974) developed a relational model for fertility schedules based on the Gompertz function, noting the utility of this approach in its simplicity and the quality of fit of model rates. Zaba (1979) developed a relational model for schedules of immigration and emigration that involves three parameters. Booth (2006) documents the use and development of relational models of migration (Congdon 1993) and fertility (Zeng et al., 2000) in her excellent review of techniques of demographic forecasting. Marshall (2009) developed relational models for the estimation of disability schedules and researchers wanting a practical guide to operationalizing relation models should see Marshall (2010).

Practitioners deciding on which approach to use for graduation and smoothing would be advised to read Keyfitz (1982), Kostaki and Panousis (2001), Peristera and Kostaki (2005) on mortality and de Beer (2011) on fertility since each of these papers differentiates between parametric, spline and relational models and compares the outputs estimated using different methods.

## Conclusion

The ability to carry out curve fitting is a very useful skill for population geographers and demographers. Curve fitting is not well covered in statistics textbooks and, whilst there is a large literature in journals thoroughly discussing the detail of functions which define curves, these texts are likely to be inaccessible to researchers

who are not specialists in mathematics. We have aimed here to make nonlinear models as accessible as possible.

Many data sources make variables available grouped across unit values and there is a need in many situations for more detailed or differently banded information. So we have concentrated on the estimation of unit values from grouped information and how to achieve this. Even if this circumstance is not a particular researcher's focus (which might be the need to smooth ragged data), we hope that the explanations and examples given here allow different purposes to be enabled. We have focused here on nonlinear regression but acknowledge that other techniques can be used and are well worth considering.

We hope that the resources referred to below will be useful both for curve fitting and using SPSS syntax.

### Recommended resources

Advice on various types of curve fitting

Fox, J. (2000). *Nonparametric Simple Regression: Smoothing Scatterplots*. SAGE: Thousand Oaks, California.

Marshall, A. (2010). *Small area estimation using ESDS government surveys: an introductory guide*. Online <http://www.esds.ac.uk/government/docs/smallareaestimation.pdf>.

Preston, S., Heuveline, P. & Guillot, M. (2001). *Demography: Measuring and Modelling Population Processes*. Blackwell: Oxford.

Wilson, T. (2010). Model migration schedules incorporating student migration peaks. *Demographic Research* 23(8): 191–222.

Advice on the use of SPSS syntax

Boslaugh, S. (2004). *An Intermediate Guide to SPSS Programming: Using Syntax for Data Management*. SAGE: London.

**Acknowledgments** Work underpinning this paper was funded by research grants as follows: Paul Norman and Phil Rees by ESRC Research Awards RES-163-25-0032 and RES-189-25-0162, Alan Marshall by ESRC Postdoctoral Fellowship ES/H030328/1, Chris Thompson by ESRC RIBEN CASE studentship with Acxiom, and Lee Williamson by ESRC CASE studentship S42200124001 and Bradford Metropolitan District Council. The case studies reported here have been enabled by Bradford Birth Statistics Database supplied by Bradford Metropolitan District Council; Research Opinion Poll data supplied by Acxiom Ltd; 1991 and 2001 Census data obtained through the MIMAS CASWEB facility, an academic service supported by ESRC and JISC; and midyear estimates and Vital Statistics provided by the Office for National Statistics. The Census and Vital Statistics data are Crown copyright and are reproduced with permission of the Office of Public Sector Information. The authors are grateful to the anonymous referees whose useful comments have helped us to improve this paper.

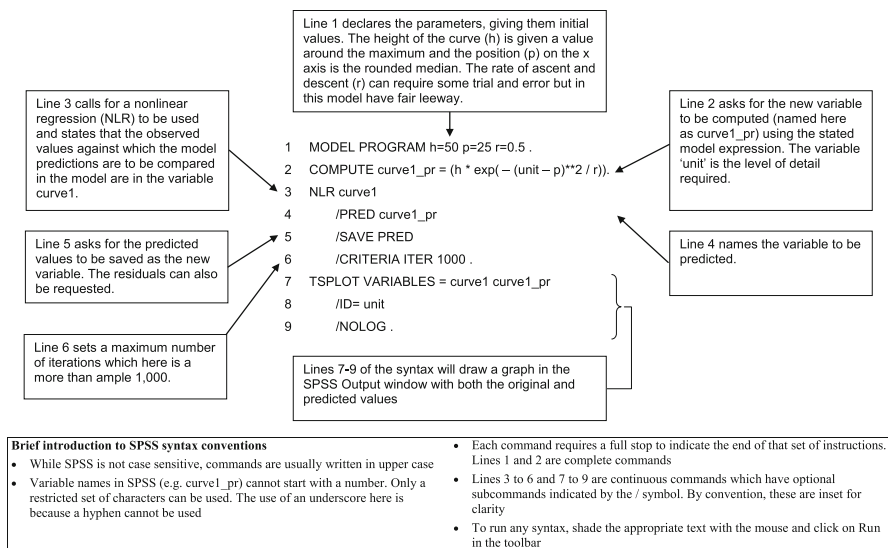
## Appendix: Using SPSS syntax

### (a) Quick guide to SPSS syntax

SPSS syntax is a ‘high level’ programming language which can readily be used to automate tasks without the user being a trained computer programmer. Users of SPSS will be familiar with the Data Editor window (comprising Data View and Variable View) through which you can open and analyse files with the .sav extension. Any analyses you carry out have the results (e.g. tables and graphs) presented in the Output window (which can be saved as .spo or .spv files depending on SPSS version). SPSS syntax files have the .sps extension.

For people who are inexperienced in using syntax, there is no need to type the command instructions from scratch. There are two easy ways to obtain syntax since SPSS will write it for you. If you use dialogue boxes to carry out an analysis, when you have made your variable selections, click on ‘Paste’ and the syntax commands to carry out your mouse click selections will be pasted into a new or previously open syntax file. A better way is to select via the menu File > Options, then the Viewer tab, tick ‘Display commands in the log’ and the click Apply. Any choices you make through ‘point and click’ are then always recorded in the Output window as syntax. You can then copy and paste syntax which you need into a .sps file and save the syntax file for future use. Existing commands can be edited for a new analysis.

### (b) Explanation of the SPSS syntax to carry out nonlinear regression





## References

- Benjamin, B., & Pollard, J. (1980). *The analysis of mortality and other actuarial statistics* (2nd ed.). London: Heinemann.
- Booth, H. (2006). Demographic forecasting: 1980 to 2005 in review. *International Journal of Forecasting*, 22, 547–581.
- Brass, W. (1960). The graduation of fertility distributions by polynomial functions. *Population Studies*, 14, 148–162.
- Brass, W. (1971). On the scale of mortality. In W. Brass (Ed.), *Biological aspects of demography* (pp. 69–110). London: Taylor Francis.
- Brass, W. (1974). Perspectives in population prediction: Illustrated by the statistics of England and Wales. *Journal of the Royal Statistical Society*, 137(Series A), 532–583.
- Chandola, T., Coleman, D. A., & Hiorns, R. W. (1999). Recent European fertility patterns: Fitting curves to ‘distorted’ distributions. *Population Studies*, 54(3), 317–329.
- Coale, A., & Trussell, J. (1996). The development and use of demographic models. *Population Studies*, 50(3), 469–484.
- Congdon, P. (1993). Statistical graduation in local demographic analysis and projection. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 156(2), 237–270.
- de Beer, J. (2011). A new relational method for smoothing and projecting age specific fertility rates: TOPALS. *Demographic Research*, 24(18): 409–454. doi:10.4054/DemRes.2011.24.18.
- Debón, A., Montes, F., & Sala, R. (2005). A comparison of parametric models for mortality graduation. Application to mortality data for the Valencia Region (Spain). *SORT*, 29(2), 269–288.
- Freund, R. J., & Littell, R. C. (2000). *SAS system for regression* (3rd ed.). New York: Wiley.
- Gage, T. B. (2001). Age-specific fecundity of mammalian populations: A test of three mathematical models. *Zoo Biology*, 20, 487–499.
- Gilje, E. (1969). Fitting curves to age-specific fertility rates: Some examples. *Statistisk tidskrift (Statistical Review of the Swedish National Central Bureau of Statistics)*, Series III, 7, 118–134.
- Hadwiger, H. (1940). Eine analytische reproduktionsfunktion für biologische gesamtheiten. *Skandinavisk Aktuarietidskrift*, 23, 101–113.
- Heligman, L., & Pollard, J. (1980). The age pattern of mortality. *Journal of the Institute of Actuaries*, 107, 49–80.
- Hoem, J. M., Madsen, D., Nielsen, J. L., Ohlsen, E.-M., Hansen, H. O., & Rennermaln, B. (1981). Experiments in modelling recent Danish fertility curves. *Demography*, 18, 231–244.
- Ibrahim, R. I. (2008). Expanding an abridged life table using the Heligman-Pollard Model. *Matematika*, 24(1), 1–10.
- Kamara, J., & Lamsana, A. (2001). *The use of model systems to link child and adult mortality levels: Peru data*. Paper presented at International Union for the Scientific Study of Population 24th General Conference, Salvador de Bahia, Brazil. Available online at: [www.iussp.org/Brazil2001/s10/S14\\_03\\_Kamara.pdf](http://www.iussp.org/Brazil2001/s10/S14_03_Kamara.pdf).
- Keyfitz, N. (1982). Choice of function for mortality analysis: Effective forecasting depends on a minimum parameter representation. *Theoretical Population Biology*, 21(3), 329–352.
- Kostaki, A., & Panousis, V. (2001). Expanding an abridged life table. *Demographic Research*, 5(1), 1–22.
- Leaker, D. (2009). Unemployment trends since the 1970 s. *Economic and Labour Market Review*, 3(2), 1–5.
- Marshall, A. (2009). *Developing a methodology for the estimation and projection of limiting long term illness and disability*. Ph.D. Thesis, School of Social Sciences, University of Manchester.
- Marshall, A. (2010). *Small area estimation using ESDS government surveys: An introductory guide*. Online <http://www.esds.ac.uk/government/docs/smallareaestimation.pdf>.
- McNeil, D. R., Trussell, J. T., & Turner, J. C. (1977). Spline interpolation of demographic data. *Demography*, 14(2), 245–252.
- Melvin, M., & Taylor, M. (2009). The global financial crisis: Causes, threats and opportunities. Introduction and overview. *Journal of International Money and Finance*, 28(8), 1–3.
- Minitab. (2011). *Minitab's nonlinear regression tool*. Online: <http://www.minitab.com/en-GB/training/articles/articles.aspx?id=9030&langType=2057>.
- Mitchell, R. (2009). Consumption patterns in recessionary times. *The Yorkshire and Humber Regional Review*, 19(1), 13–14.

- Motulsky, H., & Christopoulos, A. (2004). *Fitting models to biological data using linear and nonlinear regression*. New York: Oxford University Press.
- Newell, C. (1988). *Methods and models in demography*. New York: The Guildford Press.
- Norman, P., Rees, P., Wohland, P., & Boden, P. (2010). Ethnic group populations: The components for projection, demographic rates and trends. In J. Stillwell & M. van Ham (Eds.), *Ethnicity and integration. Series: Understanding population trends and processes* (pp. 289–315). Dordrecht: Springer.
- Peristera, P., & Kostaki, A. (2005). An evaluation of the performance of kernel estimators for graduating mortality data. *Journal of Population Research*, 22(2), 185–197.
- Peristera, P., & Kostaki, A. (2007). Modeling fertility in modern populations. *Demographic Research*, 16(6), 141–194.
- Preston, S., Heuveline, P., & Guillot, M. (2001). *Demography: Measuring and modelling population processes*. Oxford: Blackwell.
- Ratkowsky, D. A. (1983). *Nonlinear regression: A unified practical approach*. New York: Marcel Dekker.
- Rees, P., Wohland, P., Norman, P., & Boden, P. (2011). A local analysis of ethnic group population trends and projections for the UK. *Journal of Population Research*, 28(2–3), 149–183.
- Ritz, C., & Streibig, J. C. (2008). *Nonlinear regression with R*. Dordrecht: Springer.
- Rogers, A., & Watkins, J. (1987). General versus elderly interstate migration and population redistribution in the United States. *Research on Aging*, 9(4), 483–529.
- Royston, P. (1993). Standard nonlinear curve fits. *Stata Technical Bulletin Reprints*, 2, 121.
- Stata. (2011). *Nonlinear regression*. Online: <http://www.stata.com/capabilities/nlreg.html>.
- Thompson, C., Stillwell, J., & Clarke, M. (2010a). *Understanding and validating Axiom's research opinion poll data, working paper 10/6*. Leeds: University of Leeds.
- Thompson, C., Stillwell, J., & Clarke, M. (2010b). The changing grocery market in Yorkshire and Humber. *The Yorkshire and Humber Regional Review*, 20(2), 13–15.
- Vaitilingam, R. (2009). *Recession Britain: Findings from economic and social research, ESRC*. Online: [http://www.esrc.ac.uk/ESRCInfoCentre/Images/Recession\\_Britain1\\_tcm6-33756.pdf](http://www.esrc.ac.uk/ESRCInfoCentre/Images/Recession_Britain1_tcm6-33756.pdf).
- Williamson, L. E. P. (2007). *Population projections for small areas and ethnic groups—developing strategies for the estimation of demographic rates*. Ph.D. thesis, School of Social Science, Faculty of Humanities, University of Manchester.
- Williamson, L., & Norman, P. (2011). Developing strategies for deriving small population fertility rates. *Journal of Population Research*, 28(2–3), 129–148.
- Wilson, T. (2010). Model migration schedules incorporating student migration peaks. *Demographic Research*, 23(8), 191–222.
- Zaba, B. (1979). The four-parameter logit life table system. *Population Studies*, 33(1), 79–100.
- Zeng, Y., Zhenglian, W., Zhongdong, M., & Chunjun, C. (2000). A simple method for projecting or estimating a and h: An extension of the Brass relational Gompertz fertility model. *Population Research and Policy Review*, 19, 525–549.