Regression-quantile graduation of Australian life tables, 1946–1992

Esther Portnoy, University of Illinois at Urbana–Champaign
portnoy@math.uiuc.edu

Australia has conducted ten national censuses since World War II: in 1947, 1954, 1961 and at five-year intervals through 1996. Life tables [1] have been prepared in connection with each of these censuses (except the most recent one, for which data is still being collected). For ages past infancy, the life tables are based on census counts and counts of deaths in the three-year period surrounding the census. The Australian Government actuaries provided me with data relating to these life tables, namely the death counts and central exposures, by sex and single year of age. The exposure figures are combinations of the adjusted census counts P_x (adjusted for example for migration and international travel); since 1971 (and for ages 4 to about 100) the formula has been

$$E_x = {}^1/_8\, P_{x-2} + {}^7/_8\, P_{x-1} + P_x + {}^7/_8\, P_{x+1} + {}^1/_8\, P_{x+2}\,.$$

The census population counts for any given year exhibit considerable irregularity. Although age misstatement is presumably a factor, perhaps especially at older ages, a more fundamental cause is the significant fluctuation in birth and immigration rates from year to year. The coefficients in the central-exposure formula were chosen to give a more accurate estimate than 3 P_x for the average number of persons aged x (last birthday) at various times during the three-year period. I have not tried to analyze either the accuracy of counts or the validity of adjustments, but have simply taken the ratios of numbers provided to me as initial estimates of age- and sex-specific central mortality rates, $m_x = D_x/E_x$.

The graduation step in producing national life tables is frequently an ad–hoc blend of different techniques in different age ranges. There are methods that could handle the entire age span at once — Whittaker's method, or fitting a multi-parameter model such as that of Heligman & Pollard [4], for example — but they present other difficulties. In comparison to other options I claim that L_1 regression methods (generalizing Schuette graduation) offer the following advantages:

- They produce in one step a graduation for the whole age range.
- They do not require assumption of a particular parametric form either for the pattern of mortality *or for the distribution of errors* (which can be an important consideration, for example when graduating mortality rates measured in dollars instead of lives).
- They are robust in the sense of being resistant to outliers.
- They *quickly*[1] produce a number of graduations among which the graduator can choose one (or more).
- The also produce "regression quantiles", curves that estimate the (conditional) quantiles at various ages and can be interpreted as providing some indication of the variability of rates.

We begin as usual with a series of initial estimates u_x, and obtain revised estimates v_x that minimize the composite measure

$$M = \Sigma\, w_x\, |v_x - u_x| \;+ \lambda\, \Sigma\, |\Delta^2\, v_x| \;.$$

Using the second difference in the smoothness portion of M forces the revised values to vary linearly over moderate stretches of ages. Although the mechanics of Schuette graduation work just as well with higher degrees of difference, there are good statistical reasons for choosing $z = 2$ (see Koenker, Ng & Portnoy, [5]). As an example, consider Figure 1, which shows part of two graduations of the 1990–92 male data, one using the Schuette method (with z=2) and the other the Whittaker method (with z=3). The sharp corners on the Schuette graduation may initially be considered an unpleasant feature; but they represent the data more faithfully than the curves with multiple inflection points that are a common result of Whittaker graduation or cubic splines.

Because mortality rates differ so much over the full span of years (the rate at age 100 being about 2000 times that at age 10), a transformation is helpful for visual interpretation. Popular choices have been the log and logit, which reduce the span to a ratio of about 7.5 and 10 respectively. The logit has the advantage of producing a more nearly linear series of initial estimates from about age 60 upwards (in the sense of having smaller total absolute deviation from

---

[1] I used a collection of Splus programs, some with FORTRAN subroutines. Speed is not their primary advantage. Recent work by Steve Portnoy and Roger Koenker [7] indicates that L_1 methods can challenge or even surpass least–squares methods in speed, particularly for very large data sets.

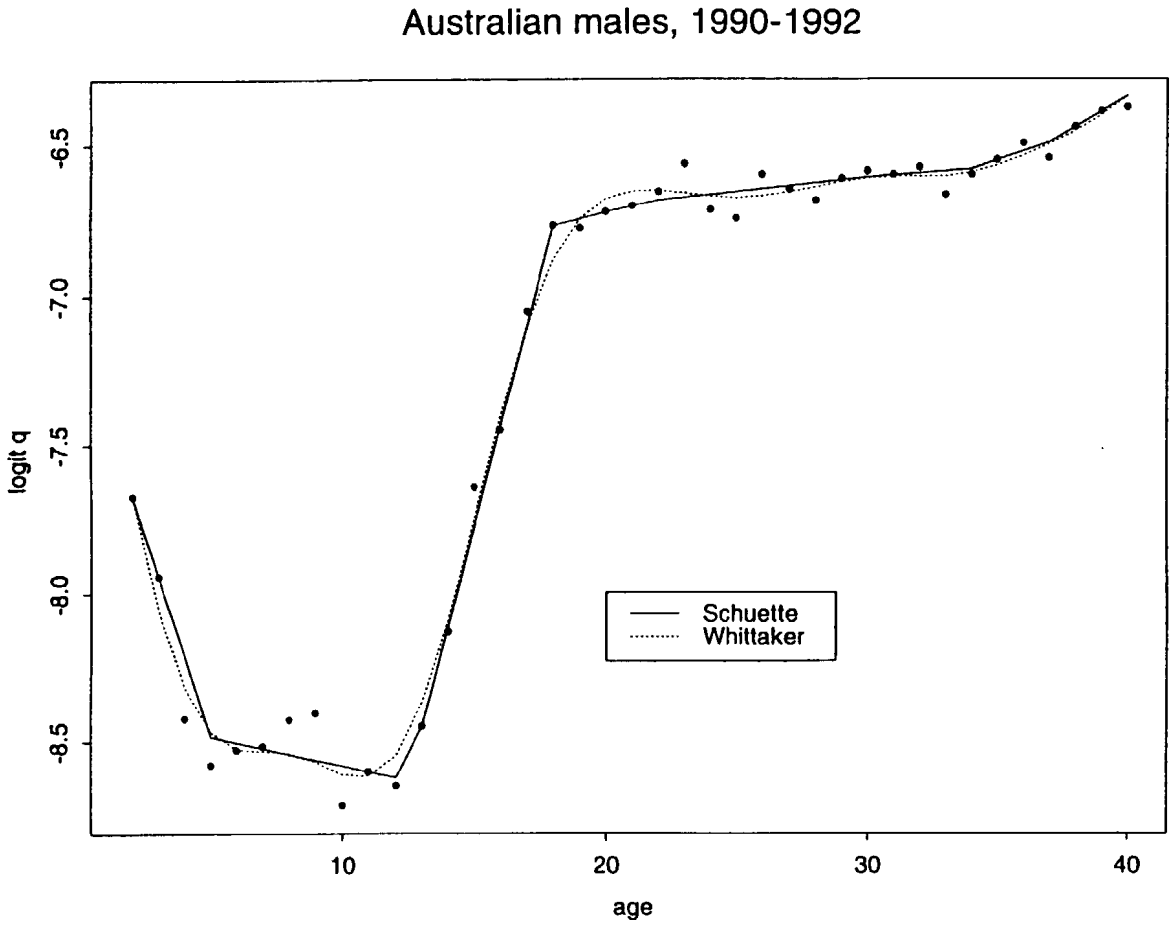
Australian males, 1990-1992
logit q
-6.5
-7.0
-7.5
-8.0
-8.5
10
20
30
40
age
Schuette
Whittaker

Figure 1

a best–fitting line). Since the graduated values are likely to vary linearly in moderately long stretches at these ages, it seemed best to begin with the more linear series. Accordingly, the following analysis is all done on initial estimates

$$u_x = \text{logit}\ (D_x/E_x) = \log\ (D_x/E_x) - \log\ (1 - D_x/E_x).$$

Weights

We turn next to the question of weights. The most common choices (for example in Whittaker graduation) have been uniform weights, weights proportional to exposures, and weights proportional to inverse variance.

The use of inverse variance would be more appropriate if the penalty for lack of fit were based on squared error rather than on absolute errors. Under the logit transformation the inverse variance is approximately $m_x(1-m_x)E_x$, which is much smaller in youth than at higher ages. (For example, in the 1991 series for females, the weight assigned at age 50 would be about 22 times that assigned at age 11, where mortality is lowest, and the weight assigned at age 85 about 65 times that at age 11.) This is a consequence of the fact that the logit spreads out the values where m_x is near 0 (or near 1) and compacts them near 1/2. Assigning higher weights at ages with much lighter exposure violates our intuitive sense of the purpose of differential weighting; more importantly, it results here in graduations that seem less satisfactory than those obtained with other weighting schemes.

The asymptotic theory for L_1 methods has been better developed for uniform weighting. One desirable feature that is not necessarily retained under differential weighting is the fact that the results provide *local* as well as global medians (and other quantiles) — that is, about half of the observed points near some age x will lie above, and about half below, the regressed values. But in experiments with the series under consideration here, which have only about 100 points to be graduated, uniform weighting led to graduations that seemed, depending on the smoothing parameter, either insufficiently smooth at high ages, or poor fits at lower ages. Figure 2 gives an example, showing the younger ages from the 1945–47 male data.

Of the options mentioned, the most satisfactory seems to be weighting proportional to the given central exposures, which drop off with increasing age. This weighting would have more appeal if we

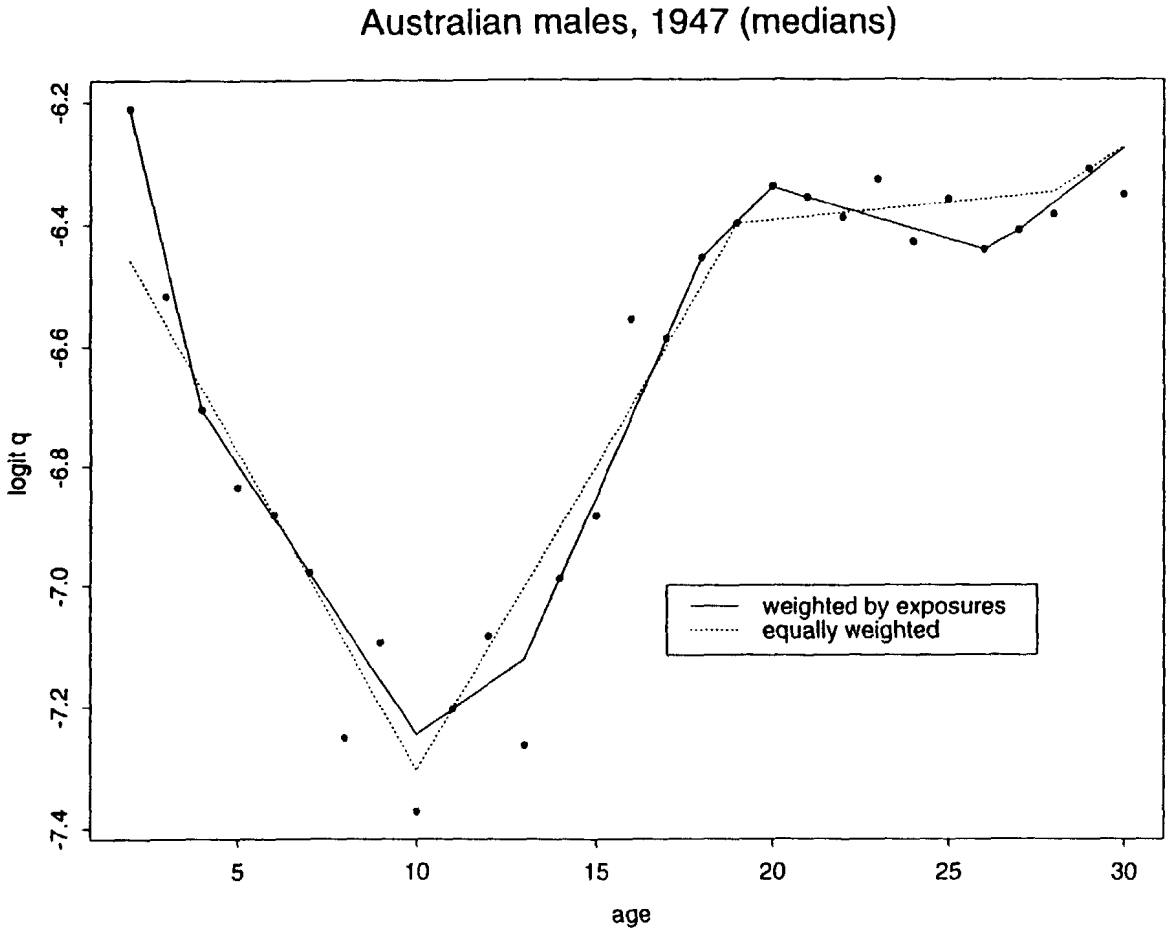
Australian males, 1947 (medians)
logit q
-6.2
-6.4
-6.6
-6.8
-7.0
-7.2
-7.4
5
10
15
20
25
30
age
weighted by exposures
equally weighted

Figure 2

were not doing the logit transformation, because then the "fit" part of the measure M would be the sum of absolute deviations in numbers of deaths, which seems a natural quantity to control. Weights were normalized (so that the sum was 1) in order to cope with some numerical stability problems.

Smoothing parameter

The last question is the choice of the smoothing parameter λ. Here the L_1 methods have a very important practical advantage. Using parametric linear programming, we can generate rather quickly the graduated values for many (or even all) values of λ. The programs[2] used here (most developed by Pin Ng, see [6]) begin with the best (L_1–fit) line, which is the optimal solution for all λ above some finite value, and then "pivot" back through successive solutions with lower values of λ. (Schuette [8] identified the linear-programming aspects of this problem, but was probably not aware of the parametric-programming method that expedites solution.) As λ decreases, the graduation *tends* to match more of the initial estimates, and have more "breaks" or turning points; but progress is not strictly monotonic. Figures 3a and 3b show the graduated values for just a few different λ values, for the 1990–92 female data. It is possible to continue through all the (finitely many) optimal solutions, but to save time one can stop when, say, 30 or 35 of the initial estimates are matched exactly. These graduations are generally too irregular to be satisfactory, so we will not be ignoring any good candidates.

Now we need only select among 150 to 200 graduations. The asymptotic theory (see Koenker, Ng & Portnoy, [5]) suggests that a good choice for λ will give a relatively small value for the Schwartz information number,

$$\mathrm{SIC}(\lambda) = \log\left(\frac{1}{n}\,\Sigma\, w_x |\, v_x - u_x |\right) + \frac{p}{2n}\log n\,,$$

where n is the number of initial estimates (in these examples, about 100) and p is the number exactly matched ($v_x = u_x$). Of course the absolute minimum is SIC $= -\infty$ when $v_x \equiv u_x$; but we find that the

---

[2]Most of the examples were generated using **qsbs**, which is available from statlib. Ng now recommends using **cobs**, developed more recently with some additional features suggested by this and other experiments; but I did not test **cobs** on these data.

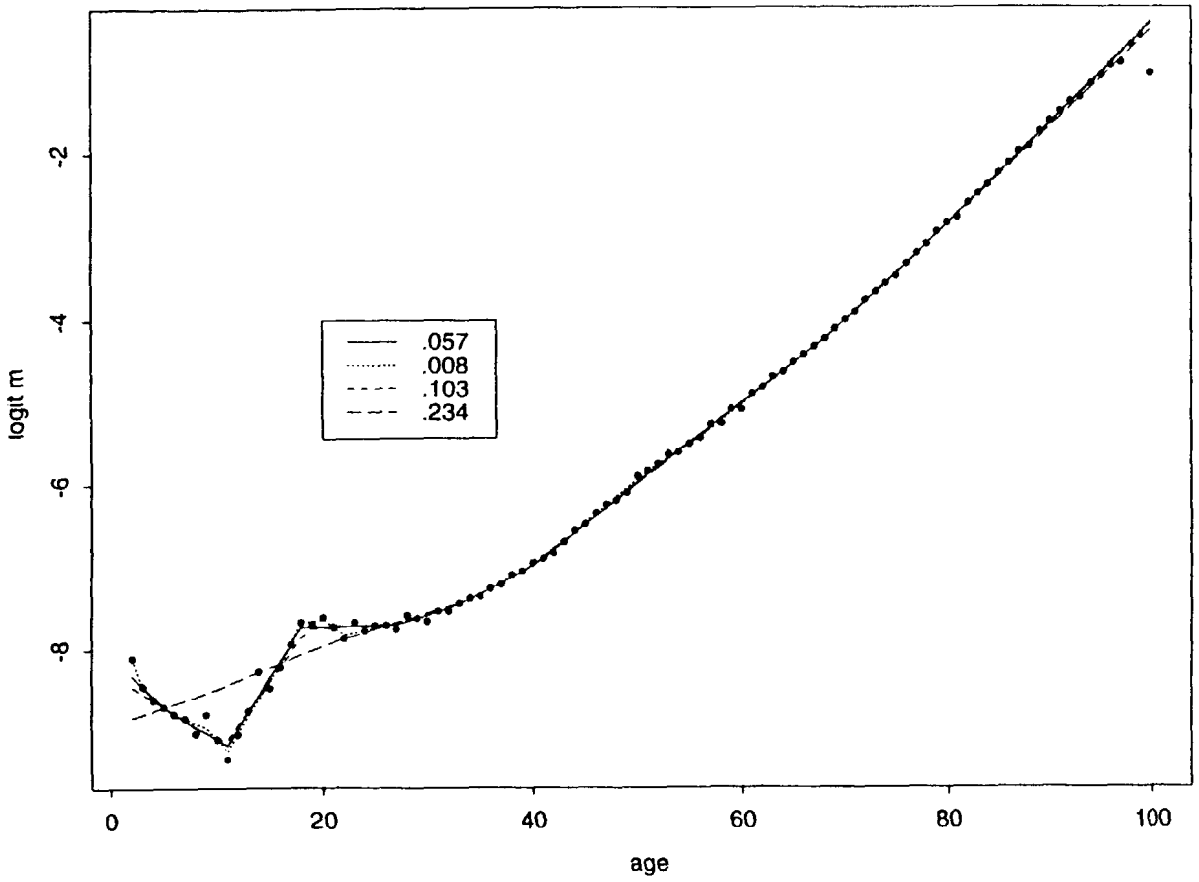
Australian females, 1990-92, various lambdas
logit m
-2
-4
-6
-8
.057
.008
.103
.234
0
20
40
60
80
100
age

Figure 3a

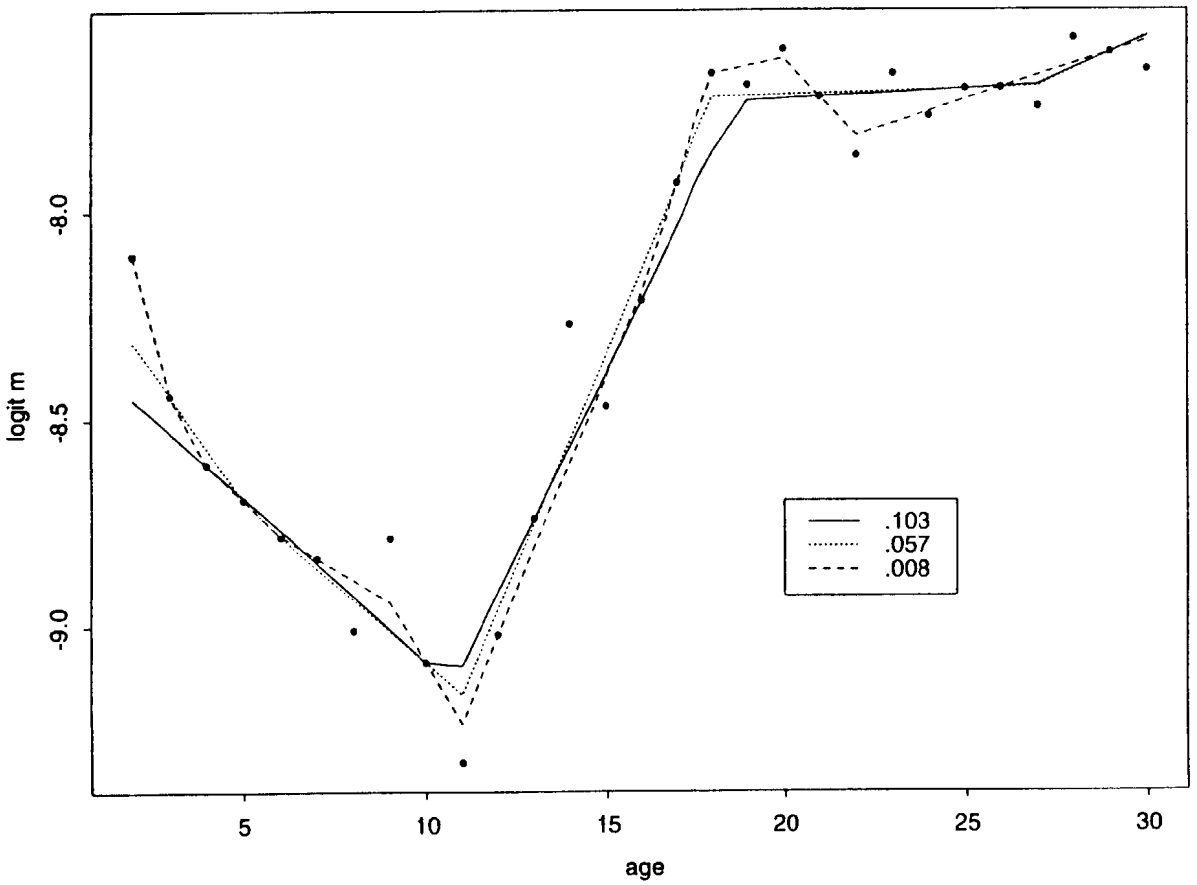
Australian females, 1990-92, various lambdas
logit m
-8.0
-8.5
-9.0
5
10
15
20
25
30
age
.103
.057
.008

Figure 3b

SIC has several *local* minima as λ varies. Figure 4 shows the SIC values for some of the optimal solutions for the 1990–92 female data. The horizontal axis is, unfortunately, only the index of the optimal solution, with #1 being the L_1–fit line, and λ decreasing as the index increases. The four graduations of Figure 3a are identified by arrows.

The Splus program **dynplt** (written by Stephen Portnoy) permits the graduator to view the solutions in succession. Visual inspection is not a reliable guide to the "best" graduation, but it can be used to limit the candidates. I generally found satisfactory graduations matching 15 to 20 of the initial estimates, and chose among them one that gave a local minimum for the SIC. Usually the plots of graduations near the one selected showed little discernible difference, indicating that the solution had a sort of stability.

Regression quantiles

A significant strength of the L_1 methods is that they permit us to generate *regression quantiles*, estimating the curves of conditional quantiles. If we alter the measure to be minimized to

$$M_\theta = \Sigma\, w_x\, \rho_\theta(v_x - u_x) \;+\; \lambda\, \Sigma\, |\, \Delta^2 v\, |$$

where $\rho_\theta(y)$ is the "check" function $2\,\{\theta \cdot y^+ + (1-\theta)\cdot y^-\}$, then the optimizing v_x satisfy the following "quantile" condition:

$$\mathbf{Q}: \qquad \Sigma_{v_x > u_x} w_x \le \theta\, \Sigma\, w_x \le \Sigma_{v_x \ge u_x} w_x$$

(that is, the v values lie above the u values at most a fraction θ of the time, by weight, and below the u values at most a fraction $1-\theta$ of the time). The Schuette graduation, obtained with $\theta=0.5$, estimates the curve of conditional medians. The Splus programs used will generate solutions for a given θ and all (or many) λ, or for fixed λ and all θ. The main advantage of this method over most other ways of estimating quantile curves is that it is non–parametric: we need not assume a particular functional form for the curve itself, or a particular distribution for errors.

The regression quantile method is very useful when we are really interested in some (non–median) quantile curve, say the 75th or 90th quantile for medical costs as a function of age. More commonly we want simultaneous estimates of several quantile curves.

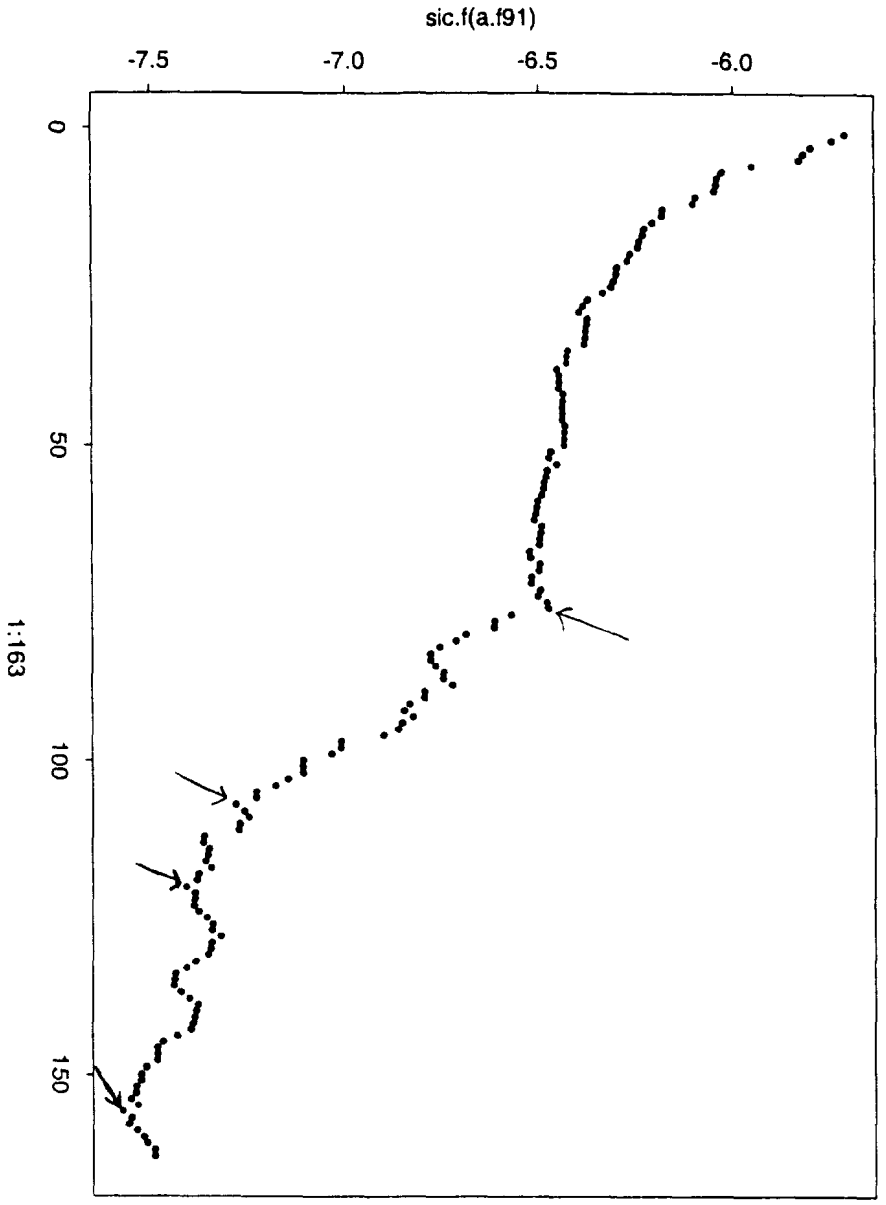

Figure 4

Unfortunately, the method above does not readily provide consistent selections. For example, if θ is very near 0 or 1, the optimal solution will be a straight line for modest values of λ, whereas for the more central quantiles one reaches the linear case only for quite large λ. More embarrassingly, it is difficult to avoid crossings. We have no assurance, for example, that the 75% curve does not drop below the median curve at some points. Such crossings do occur in practice, as Figures 5a and 5b demonstrate for two different data sets.

What we need is a way of looking at the other quantiles with reference to a selected median. An obvious first step is to consider the residuals. In very general terms, let us suppose we have a model

$$y(x) = f(x) + \epsilon(x)$$

where ϵ is random error. Having observed $y(x_i)$ and obtained in some fashion an estimate $\hat{f}(x)$, the residuals ($r_x = u_x - v_x$, in the notation of graduation) provide estimates of $\epsilon(x)$.

If we thought the errors were IID we could simply calculate quantiles of the set $\{r_x\}$, and then displace $\hat{f}(x)$ by the appropriate amount to obtain an estimate of a particular quantile curve. However, visual inspection in this case (as in many others) strongly suggests that the errors are not IID, but have larger dispersion in some age ranges than in others (Figure 6). We might even try to draw smooth curves through the residual plots to estimate, say, the quartile curves of the residual function; but we ought to have a more objective method.

My initial idea was to apply the regression quantile method to the residuals. Indeed the median obtained this way is close to the 0–function for most values of the smoothing parameter, but there are some surprises involving other quantiles — the most important being that the method does not eliminate crossings.

The following method of *restricted regression quantiles* suggested by Xuming He [3] offers a way around the crossing problem. He restricts the model by supposing that $y(x) = f(x) + s(x) \cdot e$, where e represents error, and $s(x) \geq 0$. That is, we assume the error distributions at any two ages differ only by a nonnegative multiplicative scalar. After selecting an estimate $\hat{f}$ for f and setting $r_x = y_x - \hat{f}(x)$, we smooth the *absolute* residuals $|r_x|$ to obtain a

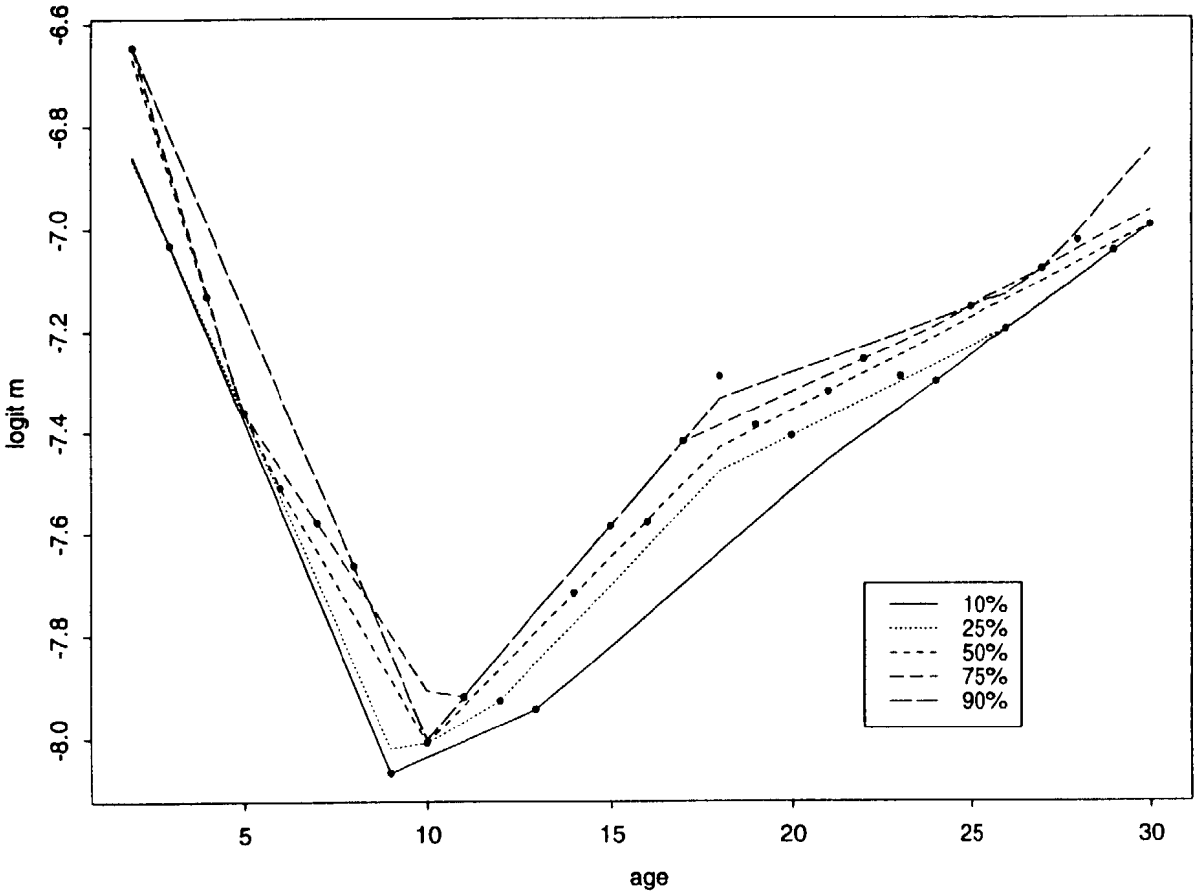
Regression quantiles for lambda = .044 (Aus F 1954)
logit m
-6.6
-6.8
-7.0
-7.2
-7.4
-7.6
-7.8
-8.0
5
10
15
20
25
30
age
10%
25%
50%
75%
90%

Figure 5a

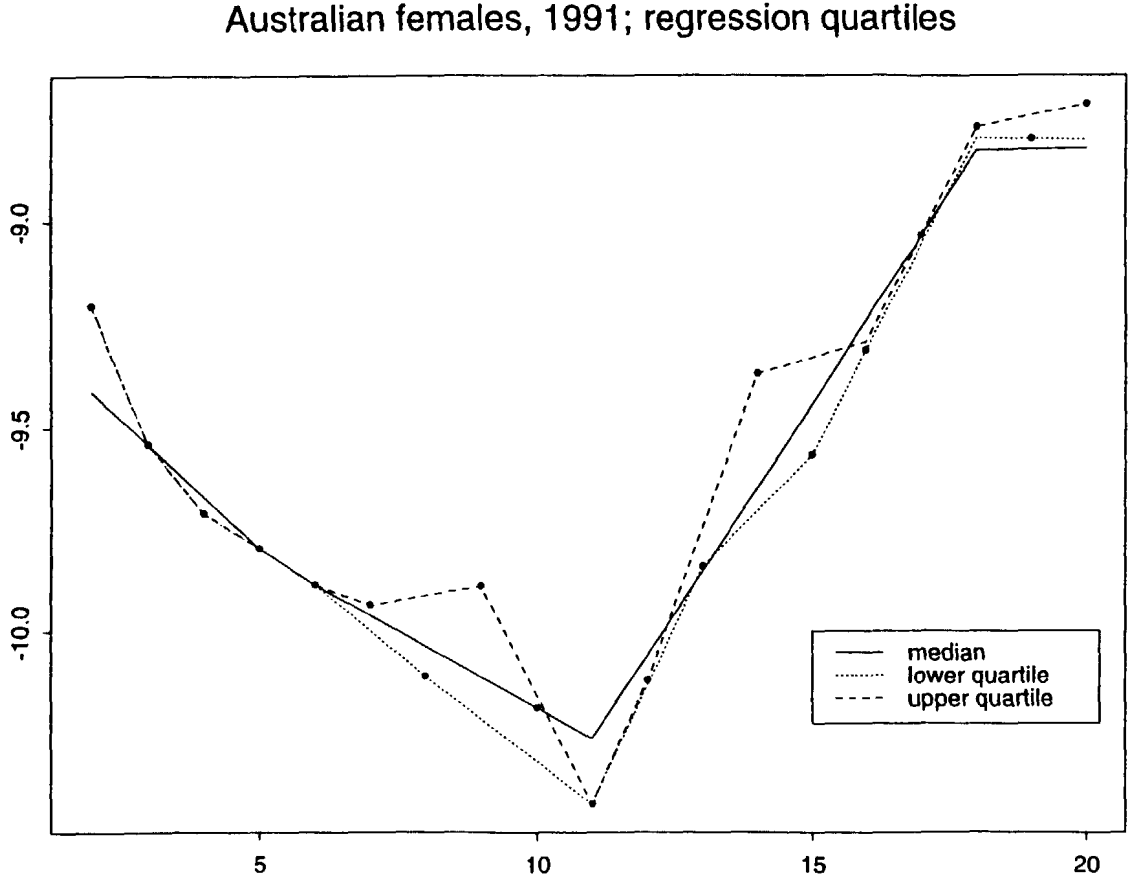
Australian females, 1991; regression quartiles
-9.0
-9.5
-10.0
5
10
15
20
median
lower quartile
upper quartile

Figure 5b

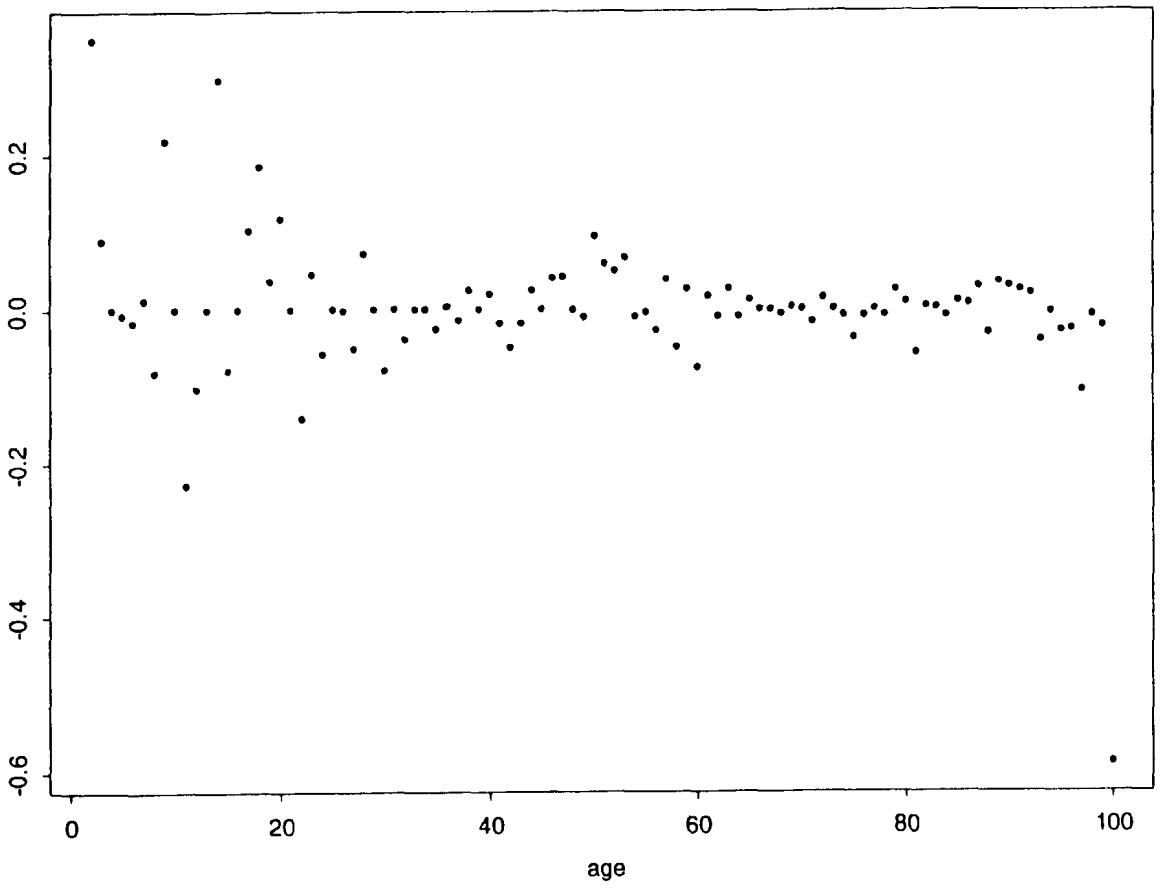

Figure 6

(nonnegative) function $\hat{s}(x)$. Finally, He suggests we estimate the quantile curves by the (non–crossing) functions $\hat{f} + c_\theta \hat{s}$, where the c_θ are chosen to minimize $S(\theta) = \Sigma\, w_x \cdot \rho_\theta(r_x - c_\theta \cdot \hat{s}(x))$. The fundamental idea here can be used no matter how the $\hat{f}$ and $\hat{s}$ are selected; but I use the L_1 (Schuette) method in both steps. One advantage is that this usually gives us a strictly positive function for $\hat{s}$, though zeros can occur if there are several consecutive zero residuals (possibly indicating a poor selection of the median estimate); and, rarely, one might get negative values at the endpoints.

Unfortunately, with the constants c_θ defined as above, the curves $\hat{f} + c_\theta \hat{s}$ do not (quite) satisfy the quantile condition Q defined earlier. As is shown in the appendix, there are a finite number of critical values c_θ, each "optimal", *i.e.* minimizing $\Sigma\, w_x \cdot \rho_\theta(r_x - c \cdot s_x)$, in an interval of θ values, namely those satisfying the inequality

$$\Sigma_{r_x < c s_x}\, w_x s_x \le \theta \cdot \Sigma\, w_x s_x \le \Sigma_{r_x \le c s_x}\, w_x s_x .$$

Condition Q, however, requires

$$\Sigma_{r_x < c s_x}\, w_x \le \theta \cdot \Sigma\, w_x \le \Sigma_{r_x \le c s_x}\, w_x .$$

The intervals defined by these two conditions will not generally coincide, though the set of critical values of c are the same. He's functions will satisfy condition Q asymptotically (as the number of graduated points tends to infinity); also, his method may be more efficient for large samples and may give more realistic estimates when there are outliers among the x values. But for the application here, with relatively small samples and x values at successive integers, it seems preferable to select the c_θ so that condition Q is satisfied. In particular, $c_{.5}$ is necessarily 0 when chosen in this fashion, not necessarily so by He's method.

In either case the selection of the c_θ is simple, essentially equivalent to finding the quantiles of a set of numbers (details in the Appendix). The c_θ are related to (estimates of) the quantiles of the

error distribution, but the relationship is not entirely simple. Quartile curves for the 1990–92 female data, generated by this method, are shown in Figures 7a and 7b; Figure 7c has the quartile curves for the 1985–87 female data (younger ages only), which demonstrate more dramatically the possibility of obtaining quite different shapes for the upper and lower quantiles.

Figure 8 shows the absolute residuals and the function $\hat{s}(x)$ for the 1990–92 female data. To eyes accustomed to least–squares methods the error function does not seem to increase at the extreme ages as much as it "ought". But this is an inevitable consequence of the robustness of the L_1 method. The outlier at age 100 has not been ignored, but it is recognized only as lying above the regression line, its distance being unimportant. The function is linear beyond age 82, and if its slope were much steeper then it would pass above 7 of the 10 values at the end, and not be a local median estimate.

The germ of these L_1 ideas goes back to 1760 and a Jesuit named Roger Boscovich[3]. Gauss knew Boscovich's work; and when he wrote about curve–fitting [in **2**] he discussed Boscovich's idea and discarded it in favor of least–squares. One of the reasons given was the above characteristic of insensitivity to outliers. Gauss may have been justifiably confident of his ability to avoid (or recognize) mistakes in data, but most of us are glad to have robust methods that offer some protection against them.

Well into my analysis of these data, it turned out that in some of the sets the last–age data was contaminated: the number of deaths was correct for that (single) age, but the exposure figure was for that and all higher ages. Consequently the central mortality rate calculated was considerably lower than it should have been. When I regraduated the data leaving out that last age (because it was not clear how to estimate the exposure), I found that the change was minimal. This would not have been the case with least–squares methods.

---

[3]A man of many talents, whom I admire so much I carry his picture i my wallet, on a ten–dinar note issued by the Republic of Croatia.

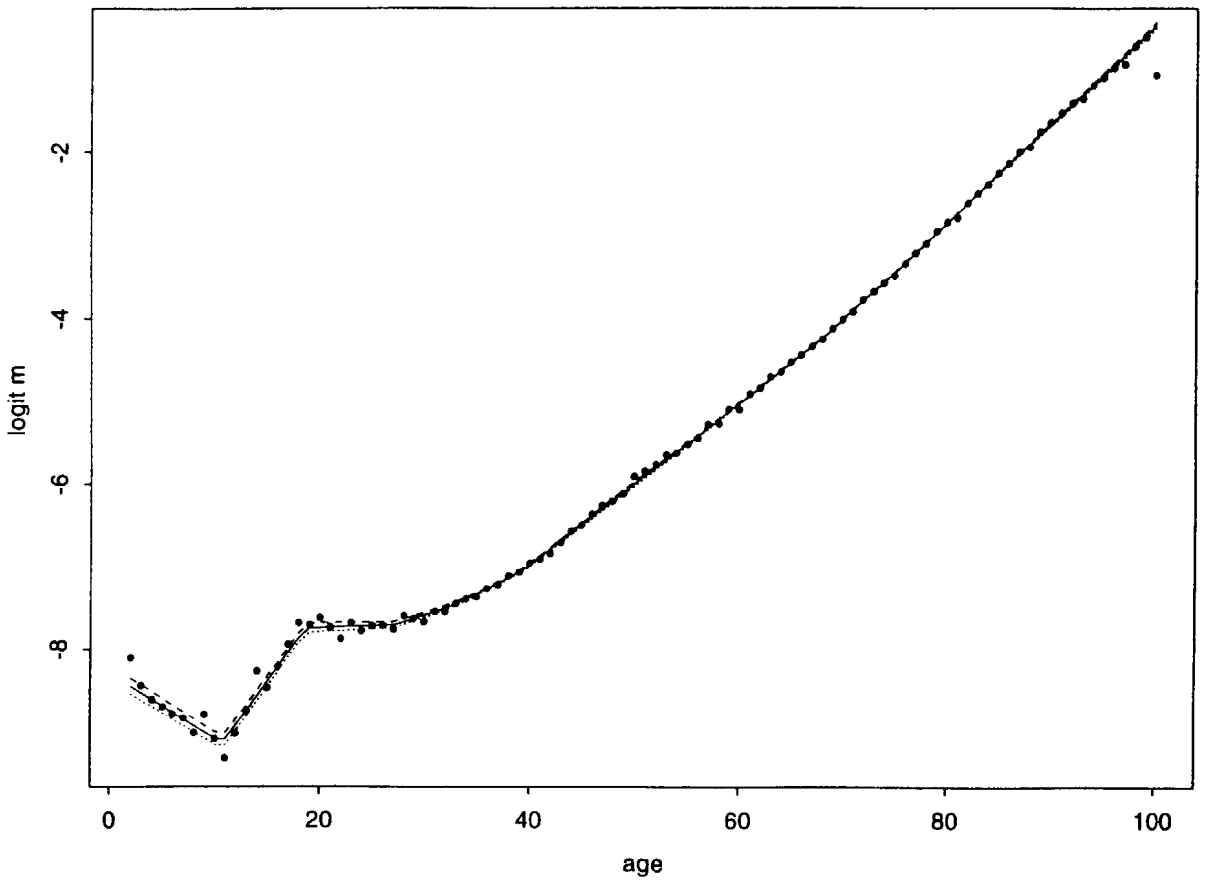

Figure 7a

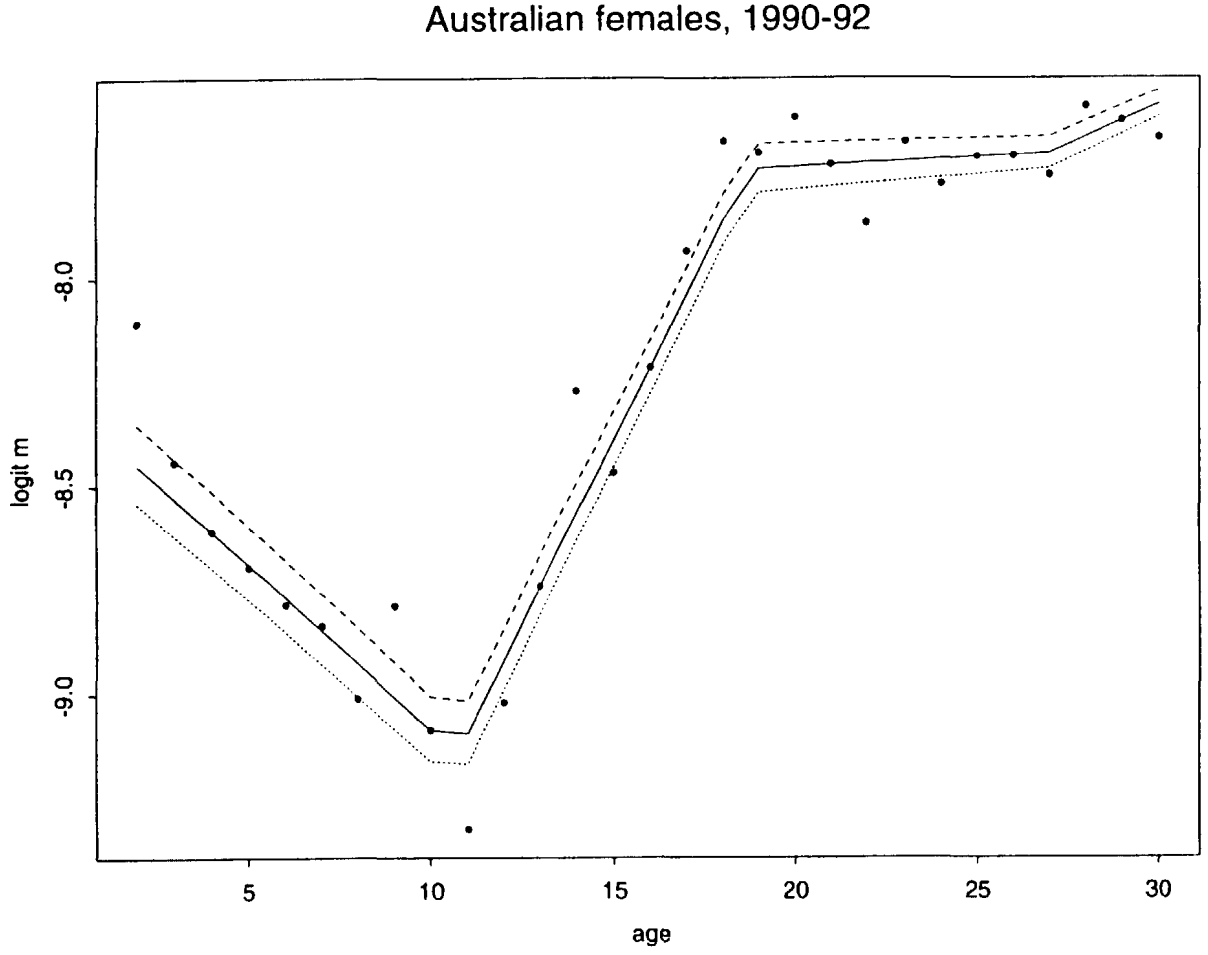

Figure 7b

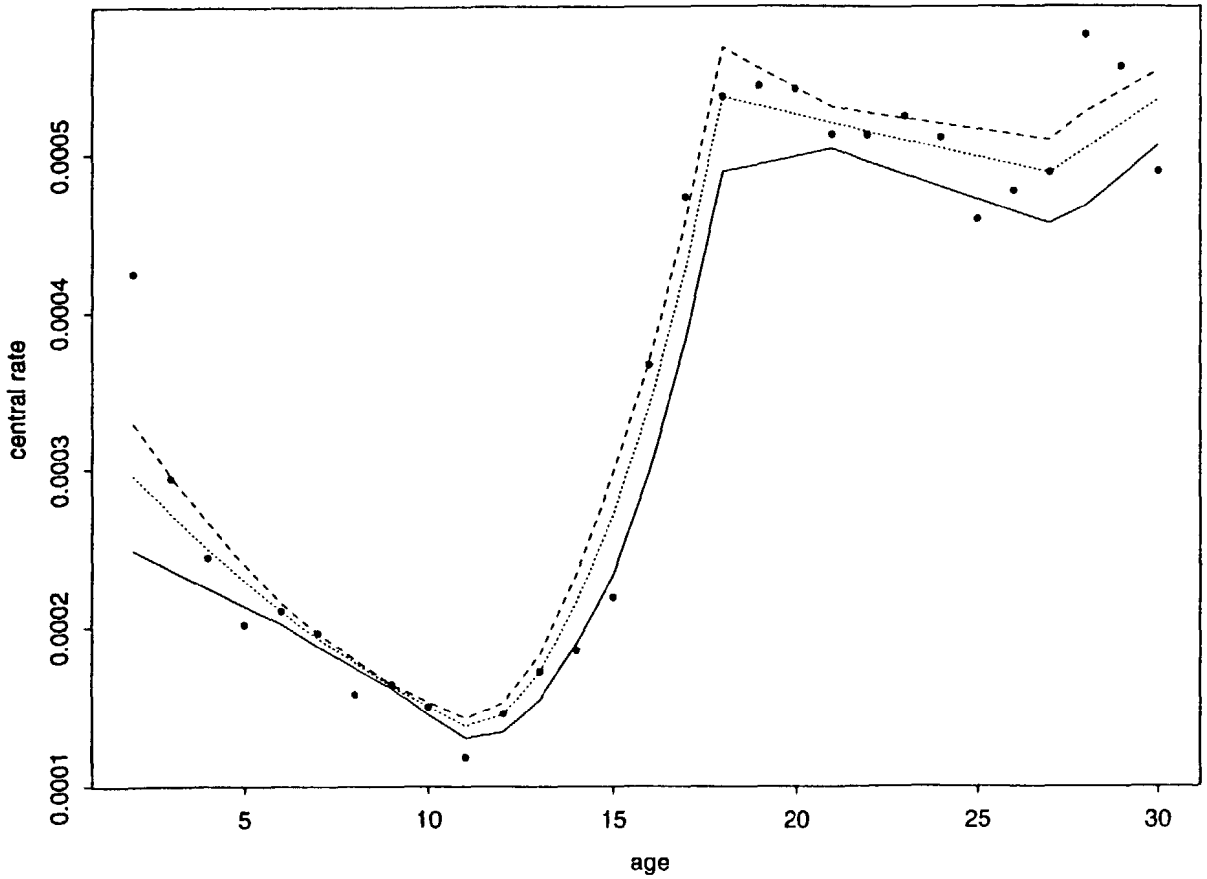
Australian females, 1986
central rate
0.0001
0.0002
0.0003
0.0004
0.0005
5
10
15
20
25
30
age

Figure 7c

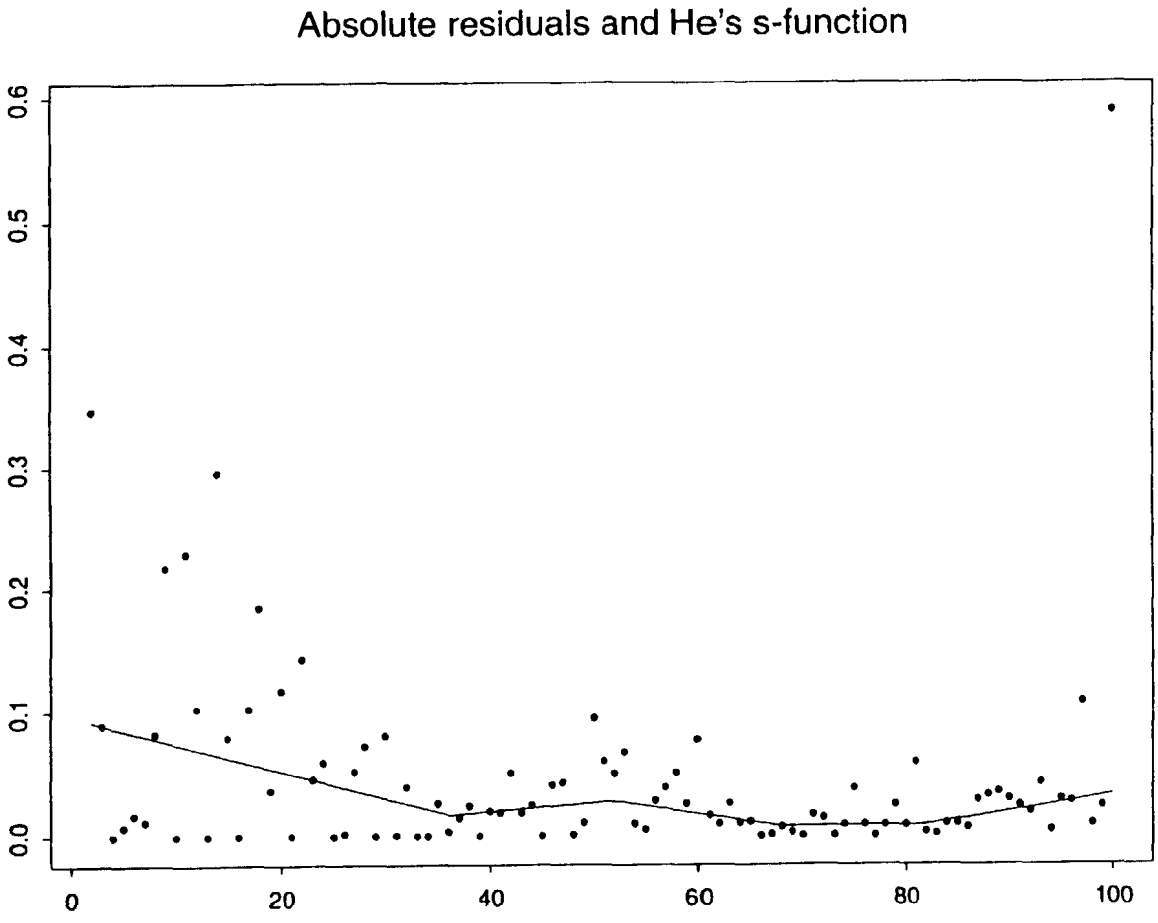
Absolute residuals and He's s-function
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0
20
40
60
80
100

Figure 8

References

1 *Australian Life Tables 1990–92* (and earlier dates), Office of the Australian Government Actuary; Australian Government Publishing Service, GPO Box 84, Canberra ACT 2601, Australia

2 Gauss, C.F., *Theoria Motus Corporum Coelestium* (1806), Section 186. (Reprinted, for example, in Volume VII of the 1906 Göttingen edition of Gauss's *Werke*.)

3 He, Xuming, "Quantile curves without crossing," to appear in *American Statistician*, 1996.

4 Heligman, L.M.A, and J.H. Pollard, "The age pattern of mortality", *Journal of the Institute of Actuaries* 107 (1980), 49–82.

5 Koenker, R., P. Ng and S. Portnoy, "Quantile smoothing splines" in *Biometrika* 81 (1994), 673–680.

6 Ng, Pin "Algorithms for quantile smoothing splines", *Computational Statistics and Data Analysis* 22 (1996), 99–118.

7 Portnoy, S., and R. Koenker, "The Gaussian hare and the Laplacian tortoise: Computability of squared-error *vs* absolute-error estimators", to appear.

8 Schuette, Donald, "A linear programming approach to graduation", *TSA* XXX (1978), 407–445.

APPENDIX: Determining the constants c_θ

Normally He would obtain the constants c_θ that minimize

$$\Sigma\, w_x \cdot \rho_\theta(r_x - c \cdot s_x)$$

(or some small set of them, such as $c_{.25}$, $c_{.5}$ and $c_{.75}$) by L_1 regression through the origin, accomplished rapidly for example by a standard Splus program. However, the following algebraic presentation clarifies the relation between He's method and the "direct" method based on the quantile criterion Q.

Suppose that $\hat{f}$ and $\hat{s} > 0$ have been chosen, and let c be any real number. For what (if any) values of θ will c minimize

$$\Sigma\, w_x \cdot \rho_\theta(r_x - c \cdot s_x)\ ?$$

We begin by dividing the set of indices into $I^+ = \{x: r_x > c\ s_x\}$, $I^- = \{x: r_x < c\ s_x\}$, and $I^0 = \{x: r_x = c\ s_x\}$. Of course the division depends on c and will be different for different values; but in the following discussion we will hold the division fixed. Now

$$S(\theta) = \Sigma\, w_x \cdot \rho_\theta(r_x - c \cdot s_x) =$$

$$2\theta\, \Sigma_{I^+}\, w_x \cdot (r_x - c \cdot s_x) + 2(1-\theta)\, \Sigma_{I^-}\, w_x \cdot (c \cdot s_x - r_x) + \Sigma_{I^0}\, 0\ .$$

If c increases by a positive amount Δc which is small enough so that $r_x > (c + \Delta c)\, s_x$ for all $x \in I^+$, then the change in $S(\theta)$ is

$$\Delta c\, \{-2\theta\, \Sigma_{I^+}\, w_x s_x + 2(1-\theta)\, \Sigma_{I^-}\, w_x s_x + 2(1-\theta)\, \Sigma_{I^0}\, w_x s_x\}$$

$$\stackrel{def}{=} -2\Delta c\{\theta\, S^+ - (1-\theta)(S^- + S^0)\}.$$

If c is optimal at θ, this change must be nonnegative; that is,

$$\theta\, (S^+ + S^- + S^0) \leq S^- + S^0.$$

On the other hand, if c decreases by a positive amount Δc small enough so that $r_x < (c - \Delta c)\, s_x$ for all $x \in I^-$, the change in $S(\theta)$ is

$$\Delta c\, \{2\theta\, \Sigma_{I^+}\, w_x s_x - 2(1-\theta)\, \Sigma_{I^-}\, w_x s_x + 2\theta\, \Sigma_{I^0}\, w_x s_x\}$$

$$= 2\Delta c\{\theta\, S^+ - (1-\theta) S^- + \theta S^0)\}.$$

Now optimality requires $\theta(S^+ + S^- + S^0) \geq S^-$. Combining these inequalities we obtain

$$S^- \leq \theta(S^+ + S^- + S^0) \stackrel{def}{=} \theta \cdot S \leq S^- + S^0\ .$$

For most values of c, I^o will be the empty set, so $S^o = 0$ and the continued inequality reduces to the equation $\theta = S^-/S$. I^o is nonempty if and only if $c = r_x/s_x$ for one or more values of x; this provides the finite collection of critical values of c. To carry out the determination for a moderately small sample, we can sort the ratios r_x/s_x by size and calculate the S^-, S^o quantities. Note that if $\hat{f}$ was obtained by L_1 regression, then $r_x=0$ for at least two (and probably more) values of x, so $c=0$ is a critical value; and if $\hat{s}$ was chosen by L_1 regression of $|r_x|$ on x then $s_x = |r_x|$ at least twice, so that at least one of $+1$, -1 is a critical value (and probably both are).

A critical value c is optimal for θ in the interval $[S^-/S, (S^- + S^o)/S]$. The smallest critical value has $S^- = 0$ (because I^- is empty). As we pass from one critical value to the next, the former I^o becomes part of I^-, while one or more indices move from I^+ into I^o. Thus the upper limit of the interval of optimality for one critical value is the lower limit for the next. Finally, at the largest critical value, I^+ is empty, thus $S^- + S^o = S$, and the upper limit is 1. Thus the entire interval $[0,1]$ is covered.

The same approach applies to determine the constants c_θ to satisfy condition Q. Setting $v_x = \hat{f}(x) + c \cdot \hat{s}_x$, then $v_x > u_x$ if and only if $c \cdot \hat{s}_x + \hat{f}(x) - u_x = c \cdot \hat{s}_x - r_x > 0$; that is, $x \in I^-$. Condition Q is thus equivalent to

$$\Sigma_{I^-} w_x \leq \theta \Sigma w_x \leq \Sigma_{I^- \cup I^o} w_x ,$$

which we might write as

$$W^- \leq \theta W \leq W^- + W^o.$$

The critical values of c (the ones for which I^o is nonempty) are the same as under He's criterion; and if all the s_x had the same value they would have the same intervals of optimality.

Normally the s_x have different values, and the intervals of optimality $[S^-/S, (S^- + S^o)/S]$ and $[W^-/W, (W^- + W^o)/W]$ will be different. This means that for some θ values the constants c_θ (and therefore the regression quantile curves) will be different under the two methods, although they are asymptotically equivalent.