Public Health Nutrition: 6(5), 513–519 DOI: 10.1079/PHN2003459

A potentially useful distribution model for dietary intake data

JP Myles[1,*], GM Price[2], N Hunter[3], M Day[4] and SW Duffy[1]

[1]Cancer Research UK Department of Epidemiology, Mathematics and Statistics, Wolfson Institute of Preventive Medicine, Charterhouse Square, London EC1M 6BQ, UK: [2]Epidemiology Unit, Department of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, London, UK: [3]MRC Biostatistics Unit Cambridge (now at National Radiological Protection Board, Chiltern, UK): [4]Food Standards Agency, London, UK

Submitted 26 April 2002: Accepted 14 January 2003

Abstract

Background: Conventional mixed models for the analysis of diet diary data have introduced several simplifying assumptions, such as that of a single standard deviation for within-person day-to-day variation which is common to all individuals.
Objective: We developed a model in which the within-person standard deviation was allowed to differ from person to person.
Design: The model was demonstrated using data on daily retinol intake from the Dietary and Nutritional Survey of British Adults. The data were from 7-day weighed dietary diaries. Estimation was performed by Markov chain Monte Carlo. Reliability of the model was assessed from the accuracy of estimation of the percentage of days on which various intakes were exceeded. For levels above the median retinol intake, estimation of percentages of days with excessive intakes was most accurate using the model with varying within-person standard deviation.
Setting: A survey of British adults aged 16–64 years.
Subjects: In total 2197 adults living in the UK, 1087 males and 1110 females.
Results: Under the traditional model, estimated daily intake ranged from 716.4 to 1421.8 μg depending on age and sex, with a within-person standard deviation of 4298.9 μg. Under the new model, estimated average daily intake ranged from 388.9 to 518.3 μg depending on age and sex, but with a within-person standard deviation varying between subjects with a 95% range of 29 to 8384 μg. The new model was shown to predict the percentage of days of exceeding large intakes more successfully than the traditional model. For example, the percentage of days of exceeding the maximum recommended intake (9000 μg for men and 7500 μg for women) was 2.4%. The traditional model predicted no excessive intakes, whereas the new model predicted 2.9%.
Conclusions: This model is potentially useful in dietary research in general and for analysis of data on chemical contaminants in foods, in particular.

Keywords
Retinol intake
Intra-subject variability
Statistical models

Individual daily consumption of various nutrients and additives is of considerable interest in the field of nutrition, especially – in the case of pesticides or other chemical residues – the estimation of the probabilities of exceeding high (and possibly dangerous) values. Much of the data that have been collected on this subject consist of daily intakes for a large group of subjects, with 7 days' data per subject, which suggests the use of multilevel models. However, one feature of such data that has been noted but is often ignored in standard analysis is the large variability between individuals in day-to-day variation of intake, as well as in mean level of intake.

To see this more clearly, consider an example of reported intakes of retinol over several days from the Dietary and Nutritional Survey of British Adults (see below for study details). Figure 1 shows reported intakes over 7 days from nine subjects in the study. Clearly the intakes are highly skewed. Individual average intakes over the 7 days vary from 256 to 773 μg. Individual standard deviations vary from 93 to 694 μg. Clearly logarithmic transformation would be likely to correct the skewed distribution but the heterogeneity between people – of average intakes and individual day-to-day variation in intakes – would remain. A model that allows for such heterogeneity is potentially of some value, particularly for the estimation of excessive values of natural toxins, pesticide residues or other hazardous chemicals in the diet. This is because while the customary assumption of a variability between days that is common to all individuals may be adequate for estimating means, it might lead to severe inaccuracy in predicting the probability of exceeding a given hazardous level of intake.

Although the major aim of this work was to provide more flexible models for intakes of chemical residues in foods, total daily intakes of these were not available at the

**Corresponding author:* Email jonathan.myles@cancer.org.uk

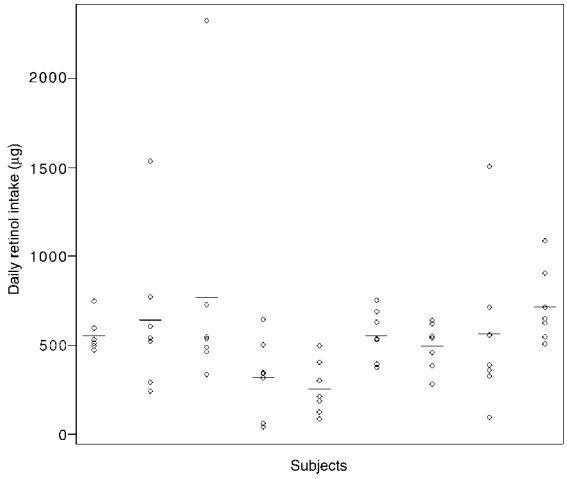

Fig. 1 Reported intakes over 7 days from nine subjects in the study

time of writing. We therefore used retinol intake as a surrogate for such chemicals in demonstration of the method, since this is known to be similarly irregularly distributed.

In this paper we describe a multilevel model for daily intake that allows for variation between individuals in day-to-day variability, and show how this model can be fitted from a Bayesian viewpoint using the BUGS program[1]. We demonstrate the use of this model to study daily intake of retinol and discuss the results.

Background and study

'Habitual diet' and the 7-day diary study

The elusive concept of 'habitual diet' underlies nutritional epidemiology. Widely accepted 'standard methods' of assessing this have emerged[2], of which the most effective is generally considered to be the 7-day weighed record, regarded in many quarters as the 'gold standard' of dietary assessment methods for nutritional assessment[3]. This method is a fully quantified 7-day diary method in which the subject is asked to weigh each item of food and drink, recording the food description and its mass before consuming it. Obviously there can be problems of misreporting, or of subjects modifying their usual diet to make the recording process simpler, but it is arguably the most complete of all the methods in general use.

The study

The Dietary and Nutritional Study of British Adults was commissioned by various branches of the UK Government to achieve a national database of dietary and nutritional information in a representative sample of non-pregnant adults. The rationale, design, methodology and basic results of the survey are described in the official report[4].

A total of 2197 adults were asked to complete a weighed 7-day diary, as well as completing various other questionnaires. The diaries were completed between October 1986 and August 1987, representing the four seasons between these dates.

Selection of covariates

A large number of person-level covariates were recorded, a subset of which was selected for this study as being most likely to affect dietary intake. The variables chosen were age, sex, height, weight and whether the subject was unwell during the 7-day period. Age, weight and height were all categorised into four groups of roughly equal size. Figures 2–5 show, for males and females, histograms of retinol intake with and without truncation at 5000 μg. Table 1 shows the mean daily intake of retinol, by each covariate and by sex. Each of the categorical variables was tested for differences in mean retinol intake by analysis of variance and regression. For the preliminary analysis to establish important covariates, the mean over the 7 days was calculated for each subject and this was used as the 'Y' variable in the equation.

Retinol intakes are known to have a positively skewed distribution, and a logarithmic transformation might be expected to rectify this to some extent. Our aim was to

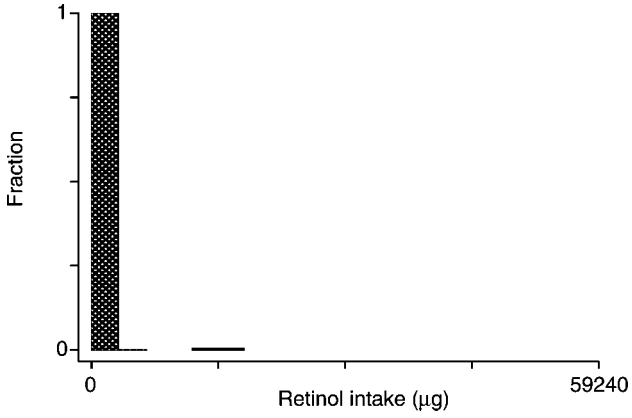

Fig. 2 Histogram of reported intakes for women (without truncation)

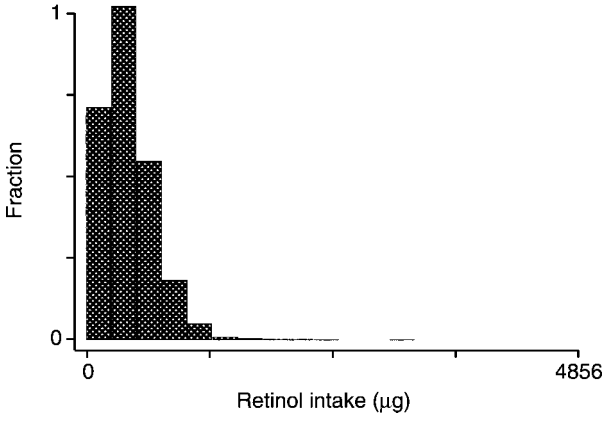

Fig. 3 Histogram of reported intakes for women (truncated at 5000 μg)

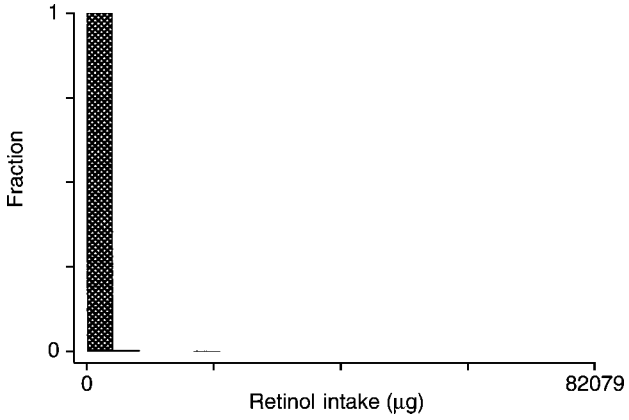

Fig. 4 Histogram of reported intakes for men (without truncation)

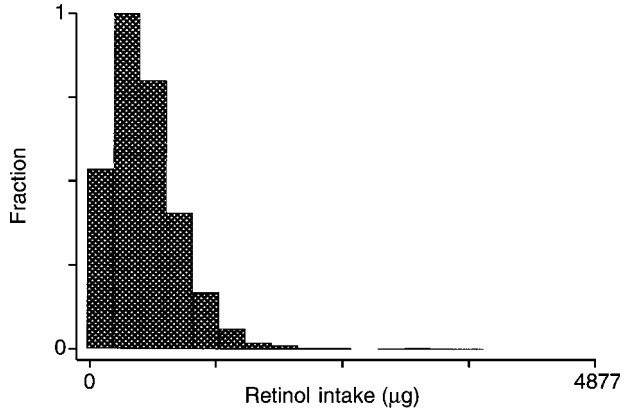

Fig. 5 Histogram of reported intakes for men (truncated at 5000 μg)

ascertain whether innovative models for the standard deviation might confer the same, or better, improvement in fit as a logarithmic transformation. Therefore, our most simple model uses untransformed data, but we did compare the results using a logarithmic transformation with the more complex analysis on untransformed data (see below). A reviewer has pointed out that a model that allowed for outliers should be expected to improve the performance of the model, but experience of data concerning daily intakes of retinol and other substances suggests that this extension would be unlikely to remove the skewness.

Those variables found to have a significant effect on retinol consumption at the 5% level – namely sex, age, height and illness – were used as the starting variables in a backwards stepwise regression, with sex forced to remain in the model. This procedure suggested dropping the illness and height covariates, resulting in a model consisting of sex and age variables which we will call Model I:

$$Y_i = \alpha + \beta_j + \gamma_k + \varepsilon_i, \quad (1)$$

where $i = 1, \ldots, 2197$ represents the subject, Y_i is the mean retinol intake of subject i over the 7 days, j represents the sex of the subject ($1 =$ male, $2 =$ female), $k = 1, \ldots, 4$ represents age category (see Table 2) and ε_i is the random variation around the fitted retinol intake (i.e. the average for all subjects in the particular age/sex category), with mean of zero and constant variance denoted by σ^2. The parameters α, β, γ and σ^2 are estimated by regression. (Throughout this paper we adopt the parameterisation $\beta_0 = \gamma_0 = 0$.)

Such a traditional linear model for the data is, however, inadequate in at least two respects. First, it does not take account of the fact that the data have a hierarchical structure, in which daily intakes are 'nested within subjects', and second, it does not allow for the fact that individual subjects may have different amounts of day-to-day variability about their own mean daily intake. In the next section we discuss models which take into account these aspects of the data.

Modelling the data

A traditional class of model for dealing with the nested structure of data is a multilevel model. In this case, we model not the mean retinol intake for each person but the intake recorded on each day, and we include in the model a parameter that represents the effect of each individual subject. These effects are assumed to be from a common distribution.

In this model (Model II) we have:

$$Y_{im} = \alpha + \beta_j + \gamma_k + \delta_i + \varepsilon_{im}, \quad (2)$$

where i, j and k are as in Model I, $m = 1, \ldots, 7$ represents day within each subject, Y_{im} represents the retinol intake

Table 1 Mean and standard deviation (SD), minimum and maximum of individual within-person SD for men and women

Sex	*n*	Mean	SD	Minimum	Maximum
Male	1087	1867.466	4124.925	31.924388	30745.52
Female	1110	1720.012	3683.809	19.5477	23314.89

Table 2 Mean and standard deviation (SD) of retinol intake (μg) by age, sex, weight, height and illness during diary period

Factor	Category	Mean	SD	*n*
Age (years)	16–24	810	3177.3	403
	25–34	1037	4226.4	507
	35–49	1225	4608.7	731
	50–64	1332	4689.1	556
Sex	Male	1215	4543.0	1087
	Female	1051	4082.0	1110
Weight (kg)	<55	960	4195.7	403
	55–69.9	1100	4116.8	507
	70–89.9	1228	4517.7	731
	≥90	1106	4477.9	556
Height (cm)	<155	752	2906.7	153
	155–169.9	1110	4338.7	1093
	170–184.9	1246	4563.7	861
	>185	963	3516.0	69
Illness	No	1170	4380.6	1816
	Yes	973	4069.6	357

of person i on day m, δ_j represents the random variation in intake from subject to subject (assumed to be normal with mean of zero and variance σ_b^2 to be estimated) and ε_{im} represents the random variation within subject i between days (assumed to be normal with mean of zero and variance σ_w^2 to be estimated). The parameters σ_b and σ_w represent between- and within-subject standard deviations. In this model we allow for within-subject variability but assume each subject to display the same amount of day-to-day variation.

This model was fitted in Stata using the XTREG command[5]. We also fitted Model II to log-transformed values, in view of the skewed distribution.

Fitting such a model allows us to estimate two variances that may be of interest: the *between-person variance* (σ_b^2), which measures the extent to which the mean intakes of individual subjects vary about the mean intake estimated from the entire study population, and the *within-person variance* (σ_w^2), which measures the amount by which subjects' daily intakes vary about their individual means.

Model II rests on the assumption that the within-person variance is constant across different subjects. It has been shown that this may be an unrealistic assumption in the case of daily energy consumption[6], and it would therefore be reasonable to assume that this might also be true of retinol intake (as well as other nutrients and additives). We therefore replace the simple multilevel model with a more complex one in which each individual has their own within-person variance, these variances being drawn from a common distribution.

This is Model III:

$$Y_{im} = \alpha + \beta_j + \gamma_k + \delta_i + \varepsilon_{im}. \qquad (3)$$

In this model, the parameters represent the same quantities as in Model II with the important difference that ε_{im}, the within-person variation from day to day, is assumed to be normally distributed with zero mean and variance $\sigma_w^2(i)$; that is, the within-person variability may be different for different individuals. The values of $\log(\sigma_w(i))$, i.e. the logarithms of the within-person standard deviations, are assumed to come from a normal distribution with mean M and standard deviation S.

The parameter δ_j again represents the difference between the actual mean intake for an individual and the estimated average intake for an individual of the same age and sex, so that a positive value of δ_j indicates that individual i has a larger mean intake than predicted by their age and sex and a negative value of δ_j indicates a smaller intake than predicted. The parameter ε_{im} represents the day-to-day variability of an individual's intakes about their daily mean and its variance is $\sigma_w(i)$: larger-than-average values of this indicate that the individual's daily intakes vary more about their individual mean than the population as a whole, while smaller values indicate less variability. Consider for example two individuals, both of whom have a predicted daily retinol intake (given their age and sex) of 600 μg, but who have different values of δ_j and $\sigma_w(i)$, so that $\delta_1 = 100$, $\sigma_w(1) = 100$, $\delta_2 = -50$ and $\sigma_w(2) = 200$. Thus individual 1 will have a higher mean intake than would be predicted by his/her age and sex alone, of $600 + 100 = 700$ μg, and a standard deviation around this of 50 μg, while individual 2 will have a lower mean intake of $600 - 50 = 550$ μg and a larger standard deviation around this of 200 μg.

This model is difficult to fit using traditional statistical packages. We fitted it using the statistical program BUGS, which estimates the parameters, including M and S, using the Bayesian technique of Markov chain Monte Carlo[1,7].

We compared the fit of the models to the data using the accuracy of prediction of the number of days on which specified levels of intake were exceeded. This strategy was chosen because formal assessment methods for goodness-of-fit for models with complex variance structure are not fully developed, and because the ultimate aim of these models was the analysis of chemical contaminants, where the probability of exceeding hazardous intakes is likely to be of some importance.

Results

Table 3 shows the results from Model I. The results have a relatively simple interpretation. For example, a male subject aged 30 years would be expected, from these results, to have an average daily retinol intake of $893.3 + 231.5 = 1124.8$ μg. A female of the same age would have an estimated average intake of $893.3 - 176.9 + 231.5 = 947.9$ μg. The model tells us nothing about any individual's day-to-day variation in intake.

The results for Model II are given in Table 4. Notice the extra parameters σ_b and σ_w. Apart from this, the interpretation is as before. For example, for a 40-year-old female, the expected mean daily intake is $893.3 - 176.9 + 424.3 = 1140.7$ μg. On the basis of this model, one would expect such an individual's daily average intake to vary about this figure with a 95% range of $1140.7 \pm (1.96 \times 344.8)$ μg, i.e. 465.0–1816.6 μg. For any individual, one would expect their successive

Table 3 Parameter estimates, with their standard error (SE), for Model I

Parameter	Estimate	SE
α intercept	893.3	89.2
β_1, male sex	–	–
β_2, female sex	−176.9	70.9
γ_1, age 16–24 years	–	–
γ_2, age 25–34 years	231.5	110.9
γ_3, age 35–49 years	424.3	103.2
γ_4, age 50–64 years	529.5	108.7

Table 4 Parameter estimates, with their standard error (SE), for Model II

Parameter	Estimate	SE
α intercept	893.3	89.2
β_1, male sex	–	–
β_2, female sex	−176.9	70.9
γ_1, age 16–24 years	–	–
γ_2, age 25–34 years	232.1	110.9
γ_3, age 35–49 years	424.3	103.2
γ_4, age 50–64 years	528.5	108.7
σ_b	344.8	
σ_w	4298.9	

daily intakes to vary with a 95% range of $1140.7 \pm (1.96 \times 4298.9)$ μg, i.e. −7285.0–9566.6 μg. Clearly, negative intakes are impossible, and this suggests a hugely skewed distribution of intake, which might be at least partly corrected by a logarithmic transformation.

Table 5 shows the results of regression analysis after such a transformation. In this case, the expected logarithm of daily retinol intake for a female aged 40 years is $5.87 - 0.29 + 0.46 = 6.04$. This varies between subjects with 95% range of $6.04 \pm (1.96 \times 0.57)$, i.e. 4.94–7.14, and within a given subject with 95% range of $6.04 \pm (1.96 \times 1.10)$, i.e. 3.88–8.20. Transforming to the linear scale, we have an expected value of 419.8 μg with a 95% range between subjects of 139.8–1261.4 μg and a 95% range within a given subject of 48.4–3641.0 μg.

Table 6 shows the results of Model III (untransformed data). The parameters for mean intake are interpreted in a similar way as before. For example, the expected daily intake for a woman aged 40 years is $507.8 - 114.8 + 2.78 = 395.8$ μg, much lower than the results for other models, since now high observed intakes are partly modelled as a result of high within-person variability rather than a high average intake. Estimation of an overall 95% range does not apply since in this model each subject has his or her own within-person standard deviation (SD_w). The extent to which this varies amongst subjects can be expressed as a 95% range of 29—8384, which is obtained in BUGS as the credible interval for the distribution of individual within-person standard deviations via the estimates of M and S. This means that a female subject at the lowest age of this range would have an SD_w value of 29 and day-to-day variation with a 95% range of $395.8 \pm (1.96 \times 29)$ μg, i.e. 338.9–452.5 μg, while one at the upper end would have an SD_w value of 8384 and day-to-day variation with a 95% range of $395.8 \pm (1.96 \times 8384)$ μg, i.e. 0–16 828.3 μg (after truncation at 0) (this very large standard deviation is a result of the long tail of the log-normal prior distribution that we give to the within-person standard deviations).

Table 7 shows the results of fitting Model III to log-transformed data.

The results from the different models above raise the question of which model might be 'best', in some sense, for the data. One method of assessing the different models is to consider the percentage of days, summed over all subjects in the study, on which retinol intake exceeded a given amount. We can compare the predicted percentage of such excessive intakes from each model with the actual percentages. In this example, the amounts we chose are as follows: 9000 μg for men and 7500 μg for women (the approximate maximum amount recommended), 300 μg for men and 250 μg for women (the minimum amount recommended), and twice and half these amounts. The predicted percentage of days on which such excessive intakes occurred for the first two models can be obtained from PREDICT statements in the Stata language, and for the third model by adding some lines to the BUGS code. The actual percentages of days on which excessive intakes occurred, and the percentages predicted by each model, are shown in Tables 8 and 9.

We can also compare histograms of observed daily intakes for the entire study population with those predicted by Model III with (Fig. 6) and without (Fig. 7) log-transformation.

Table 5 Parameter estimates, with their standard error (SE), for Model II with log-transformed data

Parameter	Estimate	SE
α intercept	5.87	0.04
β_1, male sex	–	–
β_2, female sex	−0.29	0.30
γ_1, age 16–24 years	–	–
γ_2, age 25–34 years	0.20	0.05
γ_3, age 35–49 years	0.37	0.04
γ_4, age 50–64 years	0.46	0.05
σ_b	0.56	
σ_w	1.10	

Table 6 Parameter estimates, with their standard error (SE), for Model III

Parameter	Estimate	SE
α intercept	507.8	12.1
β_1, male sex	–	–
β_2, female sex	−114.8	8.8
γ_1, age 16–24 years	–	–
γ_2, age 25–34 years	10.5	14.4
γ_3, age 35–49 years	2.78	13.3
γ_4, age 50–64 years	−4.12	13.7
σ_b	157.8	

Table 7 Parameter estimates, with their standard error (SE), for Model III with log-transformed data

Parameter	Estimate	SE
α intercept	6.18	0.02
β_1, male sex	–	–
β_2, female sex	−2.66	0.02
γ_1, age 16–24 years	–	–
γ_2, age 25–34 years	0.05	0.03
γ_3, age 35–49 years	0.008	0.03
γ_4, age 50–64 years	−0.001	0.03
σ_b	0.3	

Table 8 Observed number of excessive intakes at different levels of intake, and numbers (with 95% confidence intervals) predicted by Model I and Model II with and without log-transformation

Intake level (μg)		Excessive intakes (%)			
			Predicted		
Men	Women	Observed	Model I (untransformed)	Model II (untransformed)	Model II (log-transformed)
18 000	15 000	1.8	0	0	0
9000	7500	2.4	0	0	0.09 (0.08–0.11)
4500	3750	3.0	0	0	7.5 (7.0–7.8)
600	500	34.1	100	24.6 (24.1–25.5)	37.5 (36.7–38.2)
300	250	73.9	100	69.2 (68.5–70.0)	79.2 (78.6–79.9)
150	125	90.6	100	91.9 (91.5–92.3)	97.2 (97.0–97.5)

Table 9 Observed number of excessive intakes at different levels of intake, and numbers (with 95% confidence intervals) predicted by Model III with and without log-transformation

Intake level (μg)		Excessive intakes (%)		
			Predicted	
Men	Women	Observed	Model III (untransformed)	Model III (log-transformed)
18 000	15 000	1.8	1.0 (0.9–1.0)	0.8 (0.8–0.9)
9000	7500	2.4	2.9 (2.8–3.0)	1.6 (1.6–1.7)
4500	3750	3.0	5.3 (5.2–5.3)	3.5 (3.4–3.7)
600	500	34.1	36.8 (36.3–37.3)	41.3 (40.8–41.7)
300	250	73.9	67.3 (66.8–67.7)	66.7 (66.2–67.1)
150	125	90.6	79.1 (78.8–79.5)	84.9 (83.8–85.3)

Clearly, none of the models predicts perfectly, but Model III does appear to be more reliable in estimation of the excessive intake figures at levels of intake above the median (the four highest levels in Tables 7 and 8).

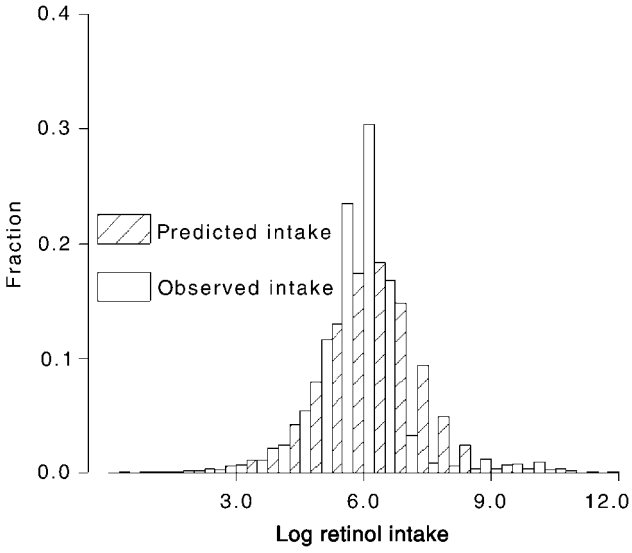

Fig. 7 Histogram of daily intakes predicted by Model III with logarithmic transformation compared with observed values

Discussion

Comparing the parameter values obtained by Model III with those obtained by Models I and II, we can see that the first two methods produce roughly equivalent results, but that the third gives results which are different in two ways:

1. the α coefficient, the intercept parameter, is sharply reduced from around 900 to around 500; and
2. there appears to be a significant effect of age in the first two models but not in the third.

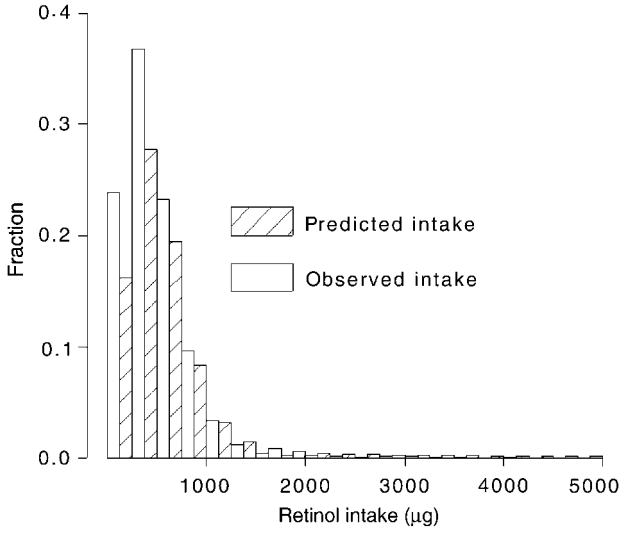

Fig. 6 Histogram of daily intakes predicted by Model III without transformation compared with observed values

Both of these phenomena can be explained by the fact that, in the third method, we are allowing each individual to have his/her own standard deviation. Although we are fitting a parametric distribution to the data (in this case a normal distribution), it is clear that for an outcome such as daily retinol intake any parametric distribution should be regarded as at best an approximation. We should expect many 'outliers', or data points that do not fit the data well. Because of the nature of the retinol intake data, particularly the obvious fact that it can only take positive values, we tend to expect such values to be higher than the predicted distribution rather than lower. These high outliers would tend, in the traditional model with a constant within-person standard deviation, to pull the intercept parameter α, which in some sense represents the overall mean, upwards. However, if we allow each

individual to have his or her own standard deviation, the outlier can be accommodated by giving the individual on whom that large daily intake was recorded a large value of their individual standard deviation, leaving the intercept parameter essentially unchanged. A similar argument explains why there appears to be a significant age effect in the first two models but not in the third. If we tabulate the average day-to-day standard deviation of subjects in each age group, we see that the standard deviation tends to increase with age (see Table 2). In the same way as described for the intercept parameter, this will tend to cause an apparent effect of intake increasing with age, unless we expressly model the different within-person standard deviation for each subject.

The implications of the above are first the already well-known phenomenon of a highly skewed distribution of retinol intakes, and second, the fact that the 'random random effects' model – which allows the within-person standard deviation to differ from person to person – is a potentially useful tool for modelling such irregularly distributed values.

One might ask why bother to model at all for the excessive intake probabilities when we can simply use the empirically observed properties. The answer is that we are not simply trying to describe what happened in this dataset: we are trying to describe, with a known range of uncertainty, the population excessive intake figures, taking account of important covariates such as age and sex. This estimation of population figures is of considerable importance for intakes of potentially hazardous chemicals in foods. This involves fitting an analysis of variance/regression model to the data, and the assumptions of the model about the variance structure can be seen from Tables 8 and 9 to be very influential. The advent of computer programs such as BUGS, which allow more complex hierarchical models, is potentially very useful in estimation of irregularly distributed dietary intakes.

Acknowledgements

We thank the Food Standards Agency for financial support. We also thank Caroline Taroudhin, Donna Yates, Cliff Gay and Paul Nunn for discussion and advice.

References

1 Spiegelhalter DJ, Thomas A, Best NG, Gilks WR. *BUGS: Bayesian Inference using Gibbs Sampling, Version 0.50*. Technical Report. Cambridge: Medical Research Council Biostatistics Unit, 1996.
2 Nelson M, Bingham SA. Assessment of food consumption and nutrient intake. In: Margetts BM, Nelson M, eds. *Design Concepts in Nutritional Epidemiology*, 2nd ed. Oxford: Oxford University Press, 1997; 123–69.
3 Bingham S, Gill C, Welch A, Day K, Cassidy A, Khaw KT, *et al*. Comparison of dietary assessment methods in nutritional epidemiology, weighed records v. 24 h recalls, food frequency questionnaires and estimated diet records. *British Journal of Nutrition* 1994; **72**: 619–43.
4 Gregory J, Forster K, Tyler H, Wiseman M. *The Dietary and Nutritional Survey of British Adults*. London: HM Stationery Office, 1990.
5 StataCorp. *Stata Statistical Software: Release 7.0*. College Station, TX: Stata Corporation, 2001.
6 Price G. *An investigation into the determination of energy intake and its variability in British adults*. Master's thesis, London School of Hygiene and Tropical Medicine, 2000.
7 Gilks WR, Richardson S, Spiegelhalter DJ, eds. *Markov Chain Monte Carlo in Practice*. London: Chapman & Hall, 1996.