A MODEL FOR COARSE GROUPING WITH APPLICATION TO SURVEY DATA

Daniel F. Heitjan, U.C.L.A.
David M. Reboussin, The RAND Corporation
Daniel F. Heitjan, Dept. of Biomathematics, UCLA, Los Angeles, CA 90024-1766

Abstract

A model for data rounding is proposed. A set of data featuring both fine and coarse rounding -- and for which both true and rounded data values are available -- is presented and analyzed. A model for the dependence of rounding coarseness on variable values is described and fit to the data.

Key words: Non-sampling errors; Rounded data; Self-reported data.

1. Introduction

Although it is common practice to model data as though they were realizations of real-valued random variables, all data are fundamentally discrete. Because this data rounding -- grouping, as it is often called -- is a universal practice, an understanding of its consequences for inferences is essential.

In survey data, grouping arises in a variety of contexts. Self-reported variables describing ages and durations, for instance, are particularly prone to these kinds of errors. For example, if a group of ex-smokers is asked how long ago they last smoked, a recent quitter might respond "four months ago," while a less recent quitter may say "about five years ago." The exact durations in days or months may not be known (or knowable), but even if they are, the values are reported, and recorded, to less precision than is possible. Furthermore, as in this example, the level of precision can depend in a significant way on the time duration.

Age is another variable that is often coarsely grouped. For infants, ages are typically reported in weeks or months, while for slightly older children it is more common to either round or truncate to the hearest year or half year. For adolescents and adults one typically truncates age to the next lower year; for the very old one may even round to the nearest five or ten years. In both the age and smoking examples, a variety of grouping precisions may coexist in a single sample, and the coarseness can be correlated with the values of the variable under study. These phenomena, as common as they appear to be, have scarcely ever been the subjects of statistical modeling and inference.

Grouping is a source of problems in data analysis because virtually all methods of analysis suppose that data have been recorded exactly; if grouping violates the model in a significant way, inferences can be vitiated. The case of rounding to a single level of precision has been studied in great detail (see the reviews of Gjeddebaek 1968 and Haitovsky 1982). The main conclusion is that grouping must be fairly coarse to make any difference, and unless it is very coarse, the application of simple moment corrections can often remove most of the bias. More complex rounding patterns, like that for age reporting, have only recently been considered (see Heitjan 1985 and Heitjan and Rubin 1986).

Our purpose in this paper is to study more carefully the case where grouping coarseness is itself variable and depends upon the value of the characteristic under study. To this end we examine a set of data in which both true values (collected prospectively) and rounded values (collected retrospectively) are known for the same set of subjects. We analyze these data to determine the extent of rounding and its dependence on the true values. We then fit the statistical model of Heitjan (1985) to the retrospective data alone. We conclude that the model can be a valuable component in the analysis of survey data that are subject to coarse, uneven grouping.

2. How does Coarse Grouping Arise?

We will assume that coarsely grouped data on self-reported variables arise in stages. We assume that the respondent knows, or can determine, the true value to a high degree of precision. Each data reporting event then involves recalling the datum, deciding what units are relevant, rounding to a common level of precision in those units, and reporting it. In reporting an infant's age, therefore, months or weeks are relevant units, since the infant's stage of development changes rapidly from week to week and month to month. To identify two infants aged one and five months as "zero years old" would be to lump together two very different creatures. On the other hand, calling two children seven years old when their ages are really seven years one month and seven years five months is unlikely to lead to confusion.

This description of data rounding is an oversimplification in that many other potential sources of error exist. For example, individuals may make false statements, either consciously or unconsciously, and have trouble remembering the dates of events. In demographic surveys in the Third World, dates of birth and other events are often unavailable because they have neither been recorded nor remembered. Thus although rounding certainly takes place in these studies, it may be less important than other kinds of non-sampling errors. (See the data of Caldwell 1966 and Pison 1979, 1980 and the review volume of Ewbank 1981).

3. An Example: The INCAP Data on Amenorrhea

The RAND/INCAP Guatemala study (Corona, undated) gathered data on the reproductive lives and other characteristics of a sample of Guatemalan villagers. The study included both prospective and retrospective phases, in which investigators asked similar sets of questions. For a number of women who gave birth during the study both prospective and retrospective lengths of post-partum amenorrhea are available. Since the prospective values were collected at a series of interviews during the post-partum period, they are held to be more accurate than the retrospective data.

From the total set of dual amenorrhea lengths we selected a subset of 150 pregnancies in which the respondents appeared to have rounded amenorrhea length to the nearest month, the nearest three months or the nearest year. For example, if prospective amenorrhea length was reported to be eight months, the pregnancy was included in our subsample if retrospective amenorrhea was eight (nearest month), nine (nearest three months) or twelve (nearest year). For these data the rounding mechanism we have described is a plausible explanation.

Figures 1 and 2 contain histograms of retrospective and prospective amenorrhea length, respectively. Rounding is clearly present in the retrospective data, where peaks appear at twelve and twenty-four months and all integral multiples of three. The prospective histogram has no such peaks and is consistent with what one expects to see when the data are not coarsely grouped.

4. Analysis of the Data

We analyzed the data to determine what proportion of pregnancies at each true (prospective) amenorrhea length were i) rounded to the nearest month, ii) rounded to the nearest three months and iii) rounded to the nearest year. Although the doubly coded sample greatly facilitates this task, there are ambiguities even in this sort of data. For example, if a mother with a true amenorrhea length of eleven months has a retrospective amenorrhea length of twelve months, she is either a nearest-three-months rounder or a nearest-twelve-months rounder. Therefore to compute probabilities of each kind of rounding at each length requires some kind of smoothing or adjustments. A detailed description of our method follows.

For each prospective amenorrhea length Y in the sample (for these data Y = 1,..., 26), we computed i) the proportion of lengths rounded to the nearest month (PEXACT(Y)), ii) the proportion to the nearest three months (PR3(Y)) and iii) the proportion rounded to the nearest year (PR12(Y)). For lengths not divisible by three, we set PEXACT to be the number of exact rounders divided by the total number at that true length. For Y divisible by three, the average of PEXACT(Y-1) and PEXACT(Y+1) was substituted.

For Y not divisible by three and not adjacent to twelve or twenty-four, PR3(Y) was taken to be the number at that true Y who rounded to the nearest three months divided by the number at that Y. For the remaining Y values (Y divisible by three or Y=11, 13, 23, 25) an average of nearby, correctly estimated PR3 values was used. We carried out a similar scheme for computing adjusted PR12 estimates. To bound proportions away from zero and one, we added 1/2 to the numerator and 1 to the denominator in each proportion calculated.

Values of the probit of PEXACT, PR3 and PR12 are plotted as stars versus Y in Figures 3, 4 and 5, respectively. A lowess smooth (Becker and Chambers 1984) is the solid line, and the upper and lower solid curves are the middle curve plus and minus one standard error. (The SE bars are jagged because the proportions are based on different numbers of exactly observed amenorrhea lengths.) From these plots it appears that the proportion rounding to the nearest month is roughly 50% or above for low amenorrhea lengths, decreasing somewhat as length increases. PR3 starts around 50%, drops in the 10-15 range, and recovers somewhat thereafter. There is a clear trend in PR12, which increases steadily across the range of Y values. These data therefore support, although not wholeheartedly, the description of rounding behavior developed in the examples previously discussed.

5. A Model-Based Analysis

A convenient and parsimonious way to summarize the kind of rounding behavior we have discussed is to relate it to a statistical model. To this end we have used the model proposed in Heitjan (1985). Suppose that there are three (or in general more or fewer) possible grouping outcomes, that may be ordered from least coarse (nearest month) to most coarse (nearest year). Heitjan has proposed explaining the dependence of the probabilities of these rounding categories upon the variate under study by an ordered categories probit regression, similar to the model described in Ashford (1958). The joint distribution of amenorrhea length and rounding type would then be completed by specifying the marginal distribution of amenorrhea length.

More precisely, we assume that the value of an underlying continuous variate Z determines, for each unit, the coarseness of rounding. To relate Z to true amenorrhea length (denoted by Y), we assume that Z is normal with conditional mean $\alpha+\beta Y$ and conditional variance σ^2. Then a probit regression can be constructed by assuming that

$$\begin{aligned} Z > 1 &\rightarrow \text{gross rounding (nearest year);} \\ 0 < Z \leq 1 &\rightarrow \text{moderate rounding (nearest 3 mos.);} \\ Z \leq 0 &\rightarrow \text{fine rounding (nearest month).} \end{aligned}$$

This results in the following expressions relating the probability of rounding to amenorrhea lengths:

$$\Pr\,[\text{nearest year} \mid \alpha,\ \beta,\ Y] = 1 - \Phi[1 - (\alpha + \beta Y))/\sigma],$$

$$\Pr\,[\text{nearest 3 mos.} \mid \alpha,\ \beta,\ Y] = \Phi\ [(1 - (\alpha+\beta Y)/\sigma] - \Phi[-(\alpha+\beta Y)/\sigma],$$

$$\Pr\,[\text{nearest month} \mid \alpha,\ \beta Y] = \Phi\ [-\ (\alpha+\beta Y)/\sigma],$$

where Φ is the standard normal integral. Any cutpoints besides 0 and 1 would work as well, but would result in corresponding changes in parameter values.

This model has characteristics desirable in a description of rounding behavior. If the slope $\beta > 0$, then the probability of fine rounding decreases to zero -- and the probability of coarse rounding increases to unity -- as the true amenorrhea length increases. This is consistent with the behavior observed in reality. The probability of moderate rounding first increases and then decreases, with the size of the maximum probability depending upon σ. Any number of intermediate rounding types could be incorporated into such a model by introducing additional cutpoints and, perhaps, introducing a dependence of σ on Y.

The analysis in section 4 is a rough attempt to fit the probit model using nearly complete data on Y and Z. Since doubly-coded data are rarely available in practice, however, Heitjan (1985) has developed and described a maximum likelihood procedure for fitting the model using only retrospective data (henceforth Y_R). Suppose that the marginal density of Y is $f_Y(y|\theta)$ and the conditional of Z given Y is $f_{z|y}(z|y,\theta)$. Then if $X(Y_R)$ is the set of possible (y,z) values leading to retrospective amenorrhea Y_R, the contribution to the likelihood corresponding to Y_R is

$$L(\theta) = Pr(Y_R|\theta) = \int_{S(Y_R)} f_Y(y|\theta) f_{Z|Y}(z|y,\theta)\, dzdy.$$

In other words the likelihood is a product of terms, each consisting of the integral of the joint density $f_Y f_{Z|Y}$ over the set of values of Z and Y leading to the observed retrospective datum Y_R.

The following examples should clarify these somewhat abstract arguments. Suppose that Y_R = 10 months. Then prospective amenorrhea is presumably 10 as well, up to rounding error. Thus this amenorrhea is rounded to the nearest month, i.e., $Y \varepsilon (9.5, 10.5)$ and $Z \leq 0$, and so

$$X(Y_R = 10) = (9.5, 10.5) \times (-\infty, 0].$$

If Y_R = 9 months, a multiple of 3 but not of 12, then amenorrhea may have been reported exactly ($Y \varepsilon (8.5, 9.5)$, $Z \leq 0$) or it may have been rounded to the nearest three months ($Y \varepsilon (7.5, 10.5)$, $Z \varepsilon (0, 1]$). Therefore

$$X(Y_R = 9) = (7.5, 10.5) \times (0,1] \cup (8.5, 9.5) \times -\infty, 0].$$

Finally, if Y_R = 12 (a multiple of 3 and 12), by similar arguments

$$X(Y_R = 12) = (6,18) \times (1,\infty) \cup (10.5, 13.5) \times (0,1] \cup (11.5, 12.5) \times (-\infty, 0].$$

Heitjan (1985) describes a Newton-Raphson algorithm for maximizing this likelihood under the assumption that Y is normal.

We present the results of fitting the probit model to the amenorrhea data in Table 1. We assumed that Y^λ was normal, resulting in a bivariate normal model for (Y^λ, Z), and fit the model for a range of λ values, eventually settling on $\lambda = 1$. A comparison of the estimated mean and variance of Y from the assumption of no rounding shows that the naive variance is an overestimate, although the mean of Y is hardly sensitive to the grouping. This behavior is typical of estimation strategies that account for rounding in the data.

The regression coefficients in Table 1 show that the slope estimate is small in magnitude but large compared to its SE. Thus, as the earlier analysis suggested, the propensity to round coarsely increases with Y. Plots of the MLE rounding propensity curves appear as the dashed lines in Figures 3, 4 and 5. Except for PR12, the model-based lines predict a more marked dependence than the previous analysis suggested, although the two sets of curves are, for most of the range, separated by less than a standard error. It is possible that the smoothed curves from the simpler analysis of section 4 are overly conservative, since those proportions were weighted artificially toward 0.5 by "starting" the numerators and denominators. These plots may also suggest that the moderate rounding category is superfluous in explaining these data. In any case, the result for PR12 is quite encouraging.

6. Conclusions

In this paper we have focused upon a single aspect of data misreporting: the grouping or coarse rounding of data. We have demonstrated how grouping may arise in survey data and have shown how one can explicitly model this behavior. Our efforts have been aided by the existence of a set of data including grouped and ungrouped versions of the same variable.

We believe that our analysis has the potential for application to many kinds of data, particularly self-reported ages, durations and the like. The kind of grouping we have described, however, may be only one of many kinds of misreporting errors afflicting a given data set. In the INCAP data, only 150 of over 800 original amenorrhea lengths appeared to be rounded as we have described. Several misreporting mechanisms were at work, although rounding was clearly one of them.

In any event, the variable-precision rounding that we have described is the rule in many kinds of data. We believe that models like that presented here can be valuable tools for summarizing rounding and its relation to other variables under study. We anticipate further applications and refinements of the method.

Acknowledgements

The authors wish to thank Julie Da Vanzo and Donald Rubin for their assistance. Dr. Heitjan's work was supported by USPHS grant CA-16042.

References

ASHFORD, J. R. (1958). "An approach to the analysis of data for semiquantal responses in biological assay." Biometrics 15, 573-581.

BECKER, R. A. and CHAMBERS, J.M. (1984). S: An Interactive Environment for Data Analysis and Graphics. Belmont, CA: Wadsworth.

CALDWELL, J C. (1966). "Study of age misstatement among young children in Ghana." Demography 3, 477-490.

CORONA, HENRY L. (undated). INCAP-RAND Guatemala Survey: Codebook and User's Manual. RAND Corporation Mimeo.

EWBANK, D.C. (1981). Age Misreporting and Age-Selective Undernumeration: Sources, Patterns, and Consequences for Demographic Analysis. Washington, D.C.: National Academy Press.

GJEDDEBAEK, N. F. (1968). "Statistical analysis: Grouped observations." In International Encyclopedia of the Social Sciences 15, edited by D. K. Sills. New York: Macmillan and the Free Press.

HAITOVSKY, J. (1982). "Grouped data." In Encyclopedia of the Statistical Sciences 3, edited by N. L. Johnson and S. Kotz. New York: John Wiley.

HEITJAN, D. F. (1985). "Analysis of a Set of Coarsely Grouped Data." Ph.D. dissertation, The University of Chicago.

HEITJAN, D. F. and RUBIN, D. B. (1986). "Inference from coarse data using multiple imputation." In Proceedings of the 18th Symposium on the Interface, edited by T. J. Boardman. Washington, DC: American Statistical Association.

PISON, G. (1979). "Age déclaré et âge réel: Une mesure des erreurs sur l'âge en l'absence d'état civil." Population 34, 637-648.

PISON, G. (1980). "Calculer l'âge sans le demander: methode d'estimation de l'âge et structure par âge des Pauls Bande (Sénégal Oriental). Population 35, 861-892.

Table 1.

Estimated Parameters in the Probit Regression Model.

	Naive (Ignoring Rounding)		ML Estimated	
Parameter	Est.	SE	Est.	SE
Mean(Y)	12.34	0.47	12.35	0.18
Var(Y)	33.74	3.90	32.43	3.95
α	_____	____	-0.46	0.31
β	____	____	0.052	0.021
σ^2	_____	____	0.72	0.26

Figure 1. Histogram of Retrospective Amenorrhea Length

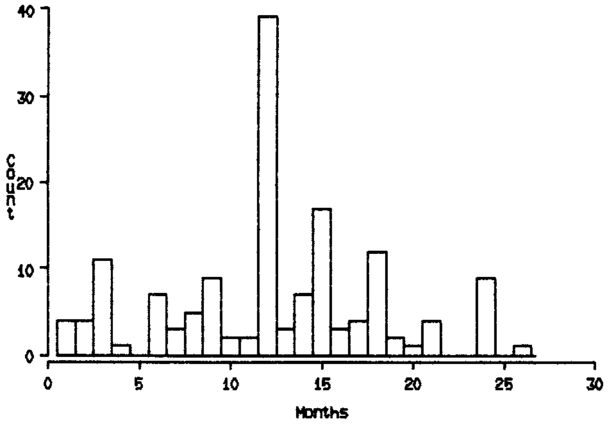


Figure 2. Histogram of Prospective Amenorrhea Length

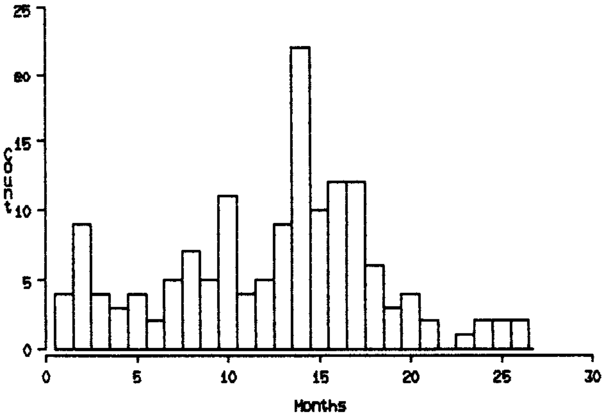


Figure 3. Proportion Not Rounding.

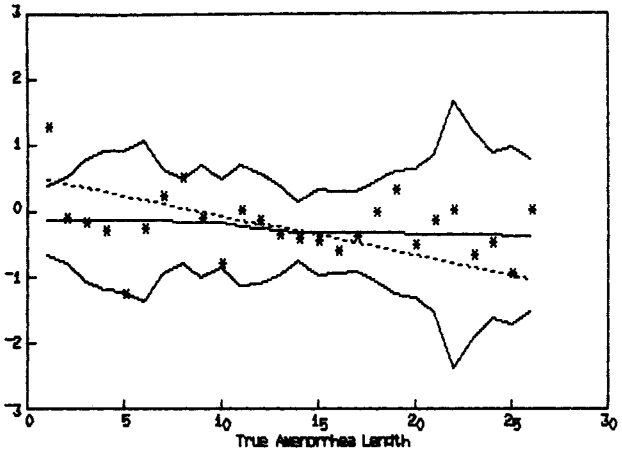


Figure 4. Proportion Rounding to Nearest 3 Mos.

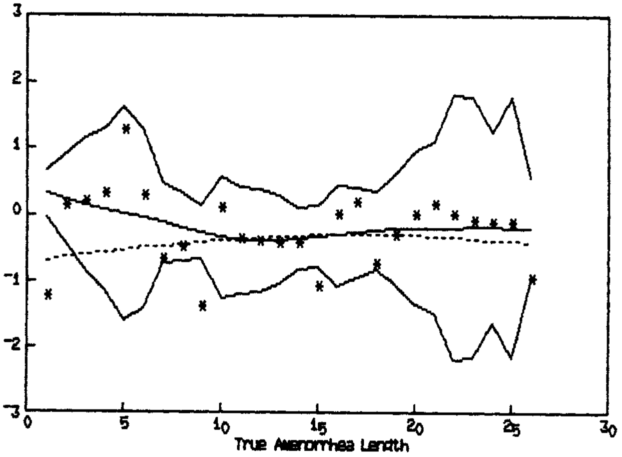
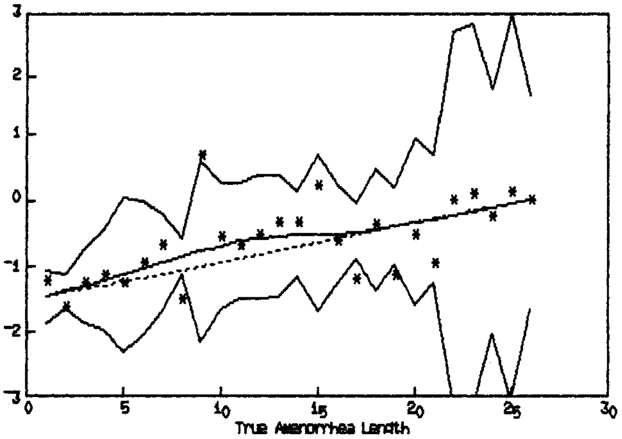


Figure 5. Proportion Rounding to Nearest Year.


True Amenorrhea Length