

CONVENTIONAL ANALYSIS OF GROUPED DATA IS FLAWED

Nicholas Beaumont

*Working Paper 23/05
May 2005*

DEPARTMENT OF MANAGEMENT
WORKING PAPER SERIES
ISSN 1327-5216



Abstract

Social science research data is often collected in grouped form, respondents being asked to nominate ranges (\$30,000-\$39,999 or 30-39 years) instead of precise values. Likert scales reduce a continuous range of responses to five or seven discrete values. The distribution of a sample obtained by grouping continuous data differs from the distribution of the population, the former is a histogram, the latter usually approximately normal. This difference means that the statistical methods habitually used by social scientists can generate wrong inferences about the population from which samples have been drawn.

Statistics derived from samples of grouped data give biased and inefficient estimates of some population parameters, for example, the variance of a grouped sample overestimates the variance of the population and widths of confidence intervals calculated without allowing for the effects of grouping will be overestimated.

In this paper we: give a mathematical derivation of the corrections needed to counteract the distortions caused by grouping; apply the findings to bivariate regression analysis; show that, in multivariate regression, the effects of ignoring grouping are more difficult to calculate, are exacerbated by independent variables' collinearity, and that failure to correct may give very biased and inefficient estimates of population parameters; present the mathematical derivations (in appendices); give an example of the recommended corrections; describe simulations that verify the theory; reiterate the implications for social science researchers; and suggest additional research.

This paper is a work in progress. Material in the paper cannot be used without permission of the author.

CONVENTIONAL ANALYSIS OF GROUPED DATA IS FLAWED

Social science research data is often collected in grouped form, respondents and informants being asked to nominate ranges (\$30,000-\$39,999 or 30-39 years), instead of providing precise values of their incomes or ages. Likert scales reduce a continuous range of responses to five or seven discrete values. The distribution of a sample obtained by grouping continuous data differs from the distribution of the population, the former is a histogram, the latter is usually assumed to be normal. This difference means that the statistical methods habitually used by social scientists can lead to wrong inferences about the population from which samples have been drawn.

This paper shows that some statistics derived from grouped samples give biased and inefficient estimates of population parameters, for example, the variance of a grouped sample overestimates the variance of the population. Because variances appear in the denominators of expressions for bivariate regression and correlation coefficients, this implies that (without correction) estimates of population regression and correlation coefficients will usually be closer to zero than they ought to be and confidence intervals will be estimated to be wider than really are, implying that statistical tests will be biased and inefficient.

It is surprising that knowledge of appropriate corrections to univariate estimators, first derived by Sheppard (1898), seems to have been overlooked, perhaps because few researchers do statistical analysis by hand (grouping was a necessary labour-saving technique) and, as far as is known, Sheppard's correction is not incorporated in any statistical packages.

In this paper we: Extend Sheppard's corrections to the multivariate case, incidentally demonstrating that corrections to means and covariances required by grouping are negligibly small; give a nonmathematical argument demonstrating the need for correction; calculate the effects of grouping in bivariate regression analysis, noting that appropriate corrections are simple to calculate; show that, in multivariate regression, the effects of ignoring grouping are more difficult to calculate (requiring matrix algebra), are exacerbated by independent variables' collinearity, and that failure to correct may give very biased and inefficient estimates of population parameters; present the mathematical derivations (in appendices); describe simulations that verify the theory; give an example; reiterate the implications for social science researchers; and suggest additional research.

The problem of analysing grouped data has been neglected. Sheppard's corrections to univariate statistics are re-derived and discussed by Aitken (1957) who shows that, if any univariate two tailed probability distribution function is grouped into intervals of constant width h and all measurements falling into a group are represented by the group's midpoint; then (i) The adjustment which should be made to estimates of the population mean is negligible and (ii) The adjustment that should be made to estimates of the population variance obtained from the grouped data to compensate for the bias caused by grouping is $-h^2/12 + o(h^4)$ where $o(h^4)$ means "to the order of". Grouping has a negligible effect on estimating means but biases estimates of variances upwards.

The biases can be appreciable. If two normal distributions are sampled using five point Likert scales, the variables' variances, regression coefficients, correlation coefficients and confidence intervals will be misestimated by about 12% (see *The Effect on Bivariate Regression Statistics*). For multivariate regression, biases and underestimates of confidence intervals are less easily predicted but may be very large when there is independent variable collinearity.

We recommend that Researchers:

- (1) Note that grouping has negligible effect on estimates of means and covariances; biases estimates of variance upwards; biases confidence intervals' widths upwards; and usually biases towards zero estimates of regression coefficients and coefficients of determination.

- (2) Note that generally, ignoring grouping makes statistical tests less efficient and biases estimates of some population parameters (usually towards zero). Lessened efficiency and bias increase the probability of type I and type II errors respectively.
- (3) Use the suggested adjustments when group widths are about the same as or greater than variables' standard deviations and thereby avoid bias (estimates of population parameters that are different from the best possible) and inefficiency (confidence intervals that are wider than necessary). Inefficiency is tantamount to throwing away painfully collected data. It is irresponsible, perhaps even unethical, to ask respondents or informants to spend time completing questionnaires if the data obtained are not fully exploited.

PAST WORK

Grouping is a form of censoring. Examples of censored or imperfectly known data are grouped data (considered in this article), data with significant experimental error, and measurements known to be greater or smaller than a certain value (for example, human heights recorded not exactly but as greater than 200cm) (Kendall & Stuart, 1979: sections 32.15-22).

In the social sciences exact data may not be available: many respondents may not know their exact income or their firm's sales for the last fiscal year and insistence on precise answers may provoke guesses or press ethical boundaries. Most researchers would prefer that respondents answer immediately, not defer answering until exact information is obtained. There is little point in obtaining exact data if there are appreciable sampling errors. It might be possible to identify some responding organisations if exact data is promulgated in its original form.

The statistical problem is to obtain the best possible estimates of population parameters and their estimation errors from censored data. Greene (1997: 437) describes "attenuation"; the fact that imperfect knowledge of the data (exemplified by censoring in general and grouping in particular) biasing estimates of population parameters towards zero.

If univariate data is collected in grouped form from a normal population the sample data will be distributed as a histogram. The statistical problem is to obtain unbiased and efficient estimates of population parameters from samples comprising observations of variables some or all of which are grouped. However, social scientists seem to have neglected the corrections to univariate distributions first proposed by Sheppard (1898), and rederived by Aitken (1957: 44) and Whittaker & Robinson (1967: 194-196). As far as is known, there has been no attempt to extend Sheppard's analysis to the multivariate case, in particular, to derive the effect of grouping on regression coefficients.

Authorities sometimes publish data in grouped form (to preserve confidentiality) but supply the means of each group. Greene (1997) gave appropriate modifications to regression techniques appropriate to this circumstance. Haitovsky (1968) discussed optimal exploitation of grouped independent variables when group means are available and Kakwani (1993) assessed the effect of grouping both dependent and independent variables on the coefficient of determination when group means were known. These techniques are not applicable to most data used in the social sciences.

Cohen (1983) studied the effect on correlation coefficients of dichotomizing either or both variables in bivariate regression analysis. He found large effects and noted the loss of power. Bollen & Barb (1981; 1983) studied, by simulation, the effect of "coarsely categorising" variables on the correlation coefficient, discovering that estimates of the correlation coefficient move towards zero as the number of categories decreases. Gepart (1983) studied the effect of grouping on regression analysis. Aggregating data for each variable into 12 groups of equal width had negligible effects on regression coefficients but aggregating it into 4 groups of equal width had significant and unpredictable effects on regression coefficients. The data were not drawn from two

tailed distributions, but truncated by zero and therefore do not validly test this article's theory. Gephart's results alone demonstrate that grouping can seriously affect estimates of population parameters. None of these studies attempted a theoretical explanation.

Stewart (1983) and Cameron (1987), studied the problem of estimating regression parameters when observations of the dependent variable were grouped. Caudill (1993) extended Stewart's method to the heteroscedastic case. Cameron (1987) studied, essentially by simulation, the effect of group "coarsening" on regression statistics. For the bivariate linear regression model $y = \beta_0 + \beta_1 x = 1 + 2x$ she used simulation to study the effect of different group widths on regression coefficients, and the regression error obtained by Ordinary Least Squares (OLS) techniques based on midpoints and a variant of Maximum Likelihood Estimation (MLE). For the regression error especially, the former method gave appreciably higher estimates than the latter. Further results are given for other distributions.

Past work and present practice demonstrate that:

- In some cases at least, the effect of grouping on estimates of population parameters can be severe (K. Bollen, 1989; J Cohen & Cohen, 1983; Gephart, 1983).
- The effects have been studied by simulations but most simulations have been studies of effects, not tests of theory.
- Most social scientists seem unaware of the effect of grouping on estimates of population parameters. As far as is known, no practicing researchers take account of the bias and inefficiency that may be caused by grouping manifest, for example, in use of Likert scales.

This paper extends the theory underlying Sheppard's correction to the multivariate case, in particular, to regression analysis and strongly recommends that researchers note that failing to take into account grouping's effects may exacerbate type I and type II errors.

THE EFFECT OF GROUPING ON ESTIMATES OF POPULATION PARAMETERS

Estimating the effects of grouping on estimates of population parameters such as regression coefficients entails obtaining the effect of grouping on distributions' moments and applying this knowledge to better estimate population parameters such as means, variances and regression coefficients. The first task requires heavy mathematics fortunately yielding simple formulae. The implications are therefore easy to understand. For univariate and bivariate statistics especially, the adjustments to statistical formulae are easy to apply.

The Effect of Grouping on Moments

We extend Sheppard's method to multivariate distributions. Grouping means that an interval level measurement is made by identifying the group into which an attribute of a data item falls and representing it (usually) by the midpoint of that group. We assume that all variables have interval measurement and each variable's domain is divided into groups of equal width (or not grouped at all, i.e. known exactly). The method entails finding the way in which grouping biases various moments about zero exemplified by forms underlying the calculation of means, covariances, and variances such as: $\sum_{i=1}^m x_i$, $\sum_{i=1}^m x_i y_i$ and $\sum_{i=1}^m x_i^2$ where m is the population size and x_i , y_i are interval measurements.

Expressions for the biases in moments caused by grouping are simple, in particular, grouping does not materially affect the "odd" moments used to calculate means, skewness, or covariances. Grouping does affect "even" moments; in particular, it overestimates the population variance and

kurtosis. Adjustments consequent on grouping to almost all commonly used statistical measures (variances, regression coefficients and coefficients of determination) and their estimation errors are easily calculated if the lower order moments are known. The mathematical derivations are given in Appendix A.

In the following paragraphs μ_r^G, μ_r^{NG} where r is a vector comprising non-negative integers, respectively denote the r -th grouped and non-grouped moments about zero, for example: μ_1^G, μ_1^{NG} respectively denote the first moments about zero using grouped and ungrouped data respectively (when divided by the sample size these give grouped and ungrouped estimates of the mean). Similarly μ_{11}^G, μ_2^{NG} are respectively the moment (based on grouped data) used to calculate covariance of two variables and the non-grouped second moment about zero of a single variable.

Special Cases

The application of formula (16) in Appendix A encapsulates the effect of grouping on the statistics most commonly used in the social sciences (for which the elements of r are 0, 1, or 2). The effects are described in the following sections.

Univariate distribution. If there is one variable (x) whose class width is h , equation (16) reduces to: $\mu_r^G = \mu_r^{NG} + \frac{1}{24}h^2 r(r-1)\mu_{r-2}^{NG} + \frac{1}{1520}h^4 r(r-1)(r-2)(r-3)\mu_{r-4}^{NG} + \dots$. Commonly used moments about the *mean* obtained from small values of r are ($\tilde{\mu}$ denotes a moment round the mean):

$$r = 0 \Rightarrow \tilde{\mu}_0^G = \tilde{\mu}_0^{NG}. \text{ Probabilities still sum to 1.}$$

$$r = 1 \Rightarrow \tilde{\mu}_1^G = \tilde{\mu}_1^{NG}. \text{ Grouping does not bias the mean.}$$

$$r = 2 \Rightarrow \tilde{\mu}_2^G = \tilde{\mu}_2^{NG} + \frac{1}{12}h^2. \text{ Grouping causes overestimation of the variance. } \textit{The variance of a sample obtained by grouping overestimates the variance of the population from which it was drawn.}$$

This bias can be corrected for by subtracting $h^2/12$ (Sheppard's correction).

$$r = 3 \Rightarrow \tilde{\mu}_3^G = \tilde{\mu}_3^{NG}. \text{ Grouping does not bias skewness.}$$

$$r = 4 \Rightarrow \tilde{\mu}_4^G = \tilde{\mu}_4^{NG} + \frac{1}{2}h^2\tilde{\mu}_2^{NG} + \frac{1}{80}h^4. \text{ Grouping causes overestimation of kurtosis.}$$

It is easy to show that a correction is necessary without using mathematics. Consider a normally distributed population with mean 0 and standard deviation 1 that has been sampled using the groupings ...[-2,-1], [-1,0], [0,1], [1,2]... The means of normal data within the intervals [0, 1] and [1, 2] are 0.3145 and 1.4377 respectively. It is obvious that the using the midpoints (0.5 and 1.5) to represent these groups creates bias.

When the midpoint is used to calculate the mean (or any odd moment about zero), the error in the interval [0,1] exactly cancels the like error in the interval [-1,0]. Because the normal distribution is symmetric, the cancellation is exact; for non-symmetric two-tailed distributions the Euler-MacLaurin theorem described in Appendix A can be used to show that cancellation over all intervals is almost exact.

However, when the variance (or kurtosis) is calculated using the groups' midpoints, the errors in intervals [-1,0] and [0,1] reinforce each other and create an appreciable error that should be corrected by using Sheppard's corrections. The analysis is robust and the corrections very accurate for any two-tailed probability distribution unless the group widths are very large.

The implications for univariate tests are obvious. If a t -test is used to test the hypothesis of no difference between means of two groups and the data is obtained from five point Likert scales assumed to cover the range -3σ to 3σ , then the class width (h) is 1.2σ . If Sheppard's correction is not used, the variance will be overestimated by 12% $\left(\frac{(1.2\sigma)^2}{12} = 0.12\sigma^2\right)$ and tests of null hypotheses will be weaker than they could and should be.

Bivariate distribution. For a bivariate distribution, the first two terms of equation (16) in Appendix A reduce to:

$$\mu_{rs}^{NG} = \mu_{rs}^G + \frac{1}{24}h_x^2r(r-1)\mu_{r-2,s} + \frac{1}{24}h_y^2s(s-1)\mu_{r,s-2} + \dots \quad r, s \geq 0. \quad (1)$$

Commonly used moments are obtained from small values of r and s .

Covariance: $r = s = 1 \Rightarrow \mu_{11}^{NG} = \mu_{11}^G$. The covariances are unaltered by grouping. Analogous to the effects described in the previous section, the effects of grouping in alternate quadrants cancel. The error caused by calculating covariance in the cell $[0 \leq x \leq 1, 1 \leq y \leq 2]$ is cancelled by the error in the cell $[0 \leq x \leq 1, -1 \leq y \leq -2]$. The cancellation is exact if the distribution (e. g. the bivariate normal distribution) is symmetric in x and y , otherwise it is almost exact. Moments such as μ_{20}^G and μ_{02}^G are biased as explained in the previous section. The extension to multivariate distributions is obvious. It would be possible to obtain the effect of grouping on higher order moments such as $\sum_{j=1}^n x_j^2 y_j^2$ by setting $r = s = 2$ in (1) but such moments are rarely used in the social sciences.

The Effect on Bivariate Regression Statistics

When biases in lower order moments are known, it is possible to calculate the biases grouping creates in commonly used bivariate statistics. In the following, x and y are the independent and dependent variables with group widths h_x and h_y respectively.

Bivariate Regression Coefficients. The regression coefficient of y on x can be expressed as $\beta_{yx} = \sigma_{xy} / \sigma_{xx}$ where σ_{xy} and σ_{xx} represent the covariance of x and y and the variance of x respectively. If the right hand side (RHS) is calculated using grouped data a biased estimator $\beta_{yx}^G = \mu_{xy}^G / \mu_{xx}^G$ will be obtained. An unbiased estimate of the population regression coefficient is $\mu_{xy}^G / (\mu_{xx}^G - h_x^2/12)$. Expressed in relative terms the bias is:

$$\delta\beta_{yx} / \beta_{yx} = \frac{h_x^2}{12(\sigma_{xx} + h_x^2/12)} \cong \frac{h_x^2}{12\sigma_{xx}}. \quad (2)$$

Thus, independent variable grouping causes the regression coefficient's magnitude to be underestimated but dependent variable grouping has no effect. If the independent variable's group width is equal to its standard deviation, the population regression coefficient's magnitude will be underestimated by about 8%.

Bivariate Coefficient of Determination. The coefficient of determination of two variables x and y is conveniently expressed as $r^2 = \sigma_{xy}^2 / (\sigma_{xx} \sigma_{yy})$. If r^2 is calculated using grouped data, a biased estimate $(r^2)^G = \mu_{xy}^2 / (\mu_{xx}^G \mu_{yy}^G)$ is obtained. An unbiased estimate is:

$$(r^2)^{NG} = \frac{\mu_{xy}^2}{(\sigma_{xx} - h_x^2/12)(\sigma_{yy} - h_y^2/12)} \quad (3)$$

$$\text{The relative error is } \delta r^2/r^2 \cong \frac{1}{12} \left[\frac{h_x^2}{\sigma_{xx}} + \frac{h_y^2}{\sigma_{yy}} \right]. \quad (4)$$

If each variable's group width is equal to its standard deviation, the coefficient of determination will be underestimated by about 17%.

These results exemplify "attenuation" (Greene, 1997: 437) and help explain the findings of Cohen (1983) and Bollen & Barb (1981; 1983). Imperfect knowledge biases regression and determination coefficients towards zero, making relationships appear weaker than they actually are and makes rejection of the null hypotheses more difficult. If the population data is two-tailed it is possible to correct the bias by applying equations (2) and (4).

Bivariate regression error. Grouping increases the standard error of the regression. The standard error can, after choosing a suitable origin, be expressed (Gujarati, 1988: 83) as:

$$\sigma^2 = E(e^2)/(m-2) = E(y - x\beta_{yx})^2/(m-2) = E(y(y - x\beta_{yx}))/m \text{ whence,}$$

$$(m-2)\sigma^2 = \sum_{i=1}^m y_i^2 - \beta_{yx} \sum_{i=1}^m x_i y_i \text{ or,} \quad (5)$$

$$\frac{(m-2)\sigma^2}{m} = \sigma_{yy} - \beta_{yx} \sigma_{yx},$$

where $E(\cdot)$ is the expectation operator, β_{yx} the regression coefficient of y on x and m is the sample size. If the group midpoints are used, a biased estimate of the regression error is obtained because β_{yx} and $\sum_{i=1}^m y_i^2$ are biased (but the cross terms $\sum_{i=1}^m x_i y_i$ are not). When grouping adjustments are made to the second order moments we have:

$$\text{If } m \text{ is large then } \sigma^2 \cong \sigma_{yy} + h_y^2/12 - \beta_{yx} \sigma_{xy} + \beta_{yx}^2 h_x^2/12 \sigma_{xx}. \quad (6)$$

The increase in the regression error resulting from grouping is:

$$\delta(\sigma^2) \approx \frac{1}{12} (h_y^2 + \beta_{yx}^2 h_x^2 / \sigma_{xx}^2). \quad (7)$$

Equation (7) shows that if grouping is ignored, the regression error will be over-estimated; we have earlier seen that, for univariate regression, the regression coefficient's magnitude will be underestimated. The squared error of estimate of the regression coefficient (SEE) is $\sigma^2/m\sigma_{xx}$ where m is the sample size. We know that grouping biases the numerator upward by $\frac{1}{12}(h_y^2 + \beta_{yx}^2 h_x^2)$ and the denominator upward by $mh_x^2/12$. It can be shown that the correction required by grouping is approximately:

$$\frac{1}{12m\sigma_{xx}} \left[h_y^2 + h_x^2 \beta_{yx}^2 - \frac{\sigma_{yy}^2 h_x^2}{\sigma_{xx}} \right] = \frac{1}{12m\sigma_{xx}} \left[h_y^2 + 2h_x^2 \beta_{yx}^2 - \frac{\sigma_{yy}^2 h_x^2}{\sigma_{xx}} \right] \quad (8)$$

$$= \frac{1}{12m\sigma_{xx}} \left[h_y^2 + h_x^2 \beta_{yx}^2 \right].$$

This term is non-negative implying that, if we naively use statistics obtained from grouped data to test the hypothesis that the regression coefficient is different from zero then, because the test value will be biased downward and the SEE biased upward, the t statistic obtained will be biased downward. It follows that *statistical tests for rejecting the null hypotheses $\beta_{yx} = 0$ will be weaker than they would be if the effects of grouping were acknowledged.* A rough guide is that effects of grouping should be considered when the group widths are of the same order as variables' standard deviations.

MULTIPLE REGRESSION WITH GROUPING

The effect of grouping on multiple regression can be obtained using matrix algebra (see Appendix B, equations (19) (20) (21) and (22)). We report the interaction of grouping and collinearity, give an example of the effect of grouping on regression analysis and note mathematical difficulties. In a separate section we report on simulations.

The Interaction of Grouping and Collinearity

Matrix equation (20) gives estimates of regression coefficients that are superior to those based on naïve analysis of grouped data. If there is no correlation amongst the independent variables (the variance-covariance matrix Z is diagonal) the adjustment to the regression coefficients would be the same as in equation (2). The grouping adjustment reduces the absolute values of the diagonal elements of matrix Z , thereby reducing diagonal dominance, and may exacerbate the effect of collinearity.

We illustrate this with the covariance matrix given in Table 1. In this example the group widths and standard deviations of all variables are 1. The covariance (denoted by t) of the independent variables x and z is changed from 0.0 to 0.9 in steps of 0.1. The differences between the estimates of the x regression coefficient if the effect of grouping is acknowledged or not acknowledged are given in Table 2. Using very large samples (10,000 observations) has minimized the effect of sampling error.

Insert Tables 1 and 2 about here

Other Methods

Lindley (1950) uses Maximum Likelihood methods to find the adjustment consequent on grouping for any univariate distribution. The method can be extended to multivariate distributions (Yoel Haitovsky, 1973: 73; Lindley, 1950). This method has the advantage of not requiring the assumption of that the population is two-tailed. For some distributions, for example exponentially distributed univariate data, grouping biases both the mean and variance. A disadvantage of Lindley's method is that, for normal distributions at least, the method requires intractable mathematical manipulations.

An Example of the Effect of Grouping on Regression

This example of the effect of grouping on regression uses data reported in AMC (1994) that describes a survey of Australian manufacturing sites. The assumption underlying the survey was that "Best Practice" would yield superior outcomes. The regression example, which uses variables reported in Australian Manufacturing Council (Australian Manufacturing Council, 1994), was not designed to test meaningful hypotheses but to show that ignoring grouping can materially affect estimates of population parameters and tests of hypotheses.

The dependent variable (denoted by PO1e) was *Total cost per unit of product* and was measured on a five point Likert scale anchored by "Relative to your major domestic and international competitors; our total cost per unit output is Much higher (1)/Much lower (5)".

The independent variables were elements of the site's planning process. All were measured on a five point Likert scale anchored by "strongly disagree" (1) and "strongly agree" (5). After cases were excluded list-wise, there were 864 valid observations.

The data needed comprises the following matrices whose meanings are given in Appendix B.

$$\mathbf{Z} = \begin{bmatrix} 864 & 3112 & 3153 & 3361 & 2912 & 3344 \\ 3112 & 12144 & 11806 & 12366 & 10975 & 12371 \\ 3153 & 11806 & 12473 & 12638 & 11069 & 12504 \\ 3361 & 12366 & 12638 & 13775 & 11675 & 13287 \\ 2912 & 10975 & 11069 & 11675 & 10968 & 11611 \\ 3344 & 12371 & 12504 & 13287 & 11611 & 13638 \end{bmatrix}$$

$$\mathbf{Xy} = [2425 \quad 8767 \quad 8867 \quad 9511 \quad 8240 \quad 9484]^T$$

$$\mathbf{H} = \text{Diagonal}\{0,1,1,1,1,1\}$$

Excel was used to calculate the matrices \mathbf{Z} and \mathbf{Xy} . The matrix manipulations were done in Mathematica (Wolfram, 1991) but could have been done in Excel. The lessened likelihood of type I and type II errors well justifies the extra time required for analysis.

Table 3 gives the individual variables' means, standard deviations, variances naively calculated, and variances with Sheppard's correction applied. Table 4 gives the bivariate correlation coefficients. Table 5 gives the values of best estimates of the regression coefficients, confidence intervals, t-ratios and the probability that a regression coefficient is different from zero. Perhaps because collinearity was not marked, the differences (except for the variable PL3) are quite mild. In this case, no estimate of a regression coefficient changed its categorization e.g. from significant to highly significant.

As explained in Appendix B, the effect on grouping on estimates of population parameters such as regression coefficients, their confidence intervals, and the coefficient of determination has to be obtained by using matrix algebra. The naively calculated standard error of regression was 0.867. The corrected value obtained from equation (9) was $0.867 + 0.094772 = 0.961772$.

 Insert Table 3, Table 4, and Table 5 about here

SIMULATIONS

The theory outlined was tested by simulations. The simulations all compared the statistical properties of a large sample of normally distributed measurements and measured how these properties were affected when the samples were segmented into groups of different widths. The simulations were in good agreement with theory.

Univariate Statistics

The theoretical results for univariate and bivariate normal distributions were tested by comparing them with the results of simulations. We considered the relationship $y = 2 + x + e$ where x and e were randomly drawn from the normal distribution with mean zero and standard deviation 1. Using a sample size of 10,000, means and variances of x , y , covariances, regression coefficients of y on x , the coefficient of determination, and regression errors were obtained when x and y were divided into groups of widths 0.25, 0.50, 0.75, 1.00 and 1.25. The univariate and bivariate results are summarised in Table 6 and Table 7 respectively.

Insert Table 6 and Table 7 about here

Table 6 shows an appreciable discrepancy between the theoretical and measured values of the variance that we are unable to explain (measured variance for zero group width is 1.054, not the expected 1.000). However, the *change* in variance is in good agreement with the theory. It is clear that, if allowance is not made for the bias caused by grouping, t -tests will be less discriminating than necessary.

Table 7 demonstrates that, for large samples at least, the results from simulations of bivariate statistics are in good agreement with theory. If allowance is not made for the bias caused by grouping, the standard error and the regression coefficient magnitude will be respectively over and underestimated and tests of the hypothesis that the regression coefficient or coefficient of determination is different from zero will be weaker than necessary.

Multivariate Statistics

We simulated the effect of grouping on multivariate regression. Ten thousand points comprising three independent variables (x , y and z) drawn from $N(0,1)$ and correlated according to Table 8 were generated. A dependent variable (w) was defined as the sum of x , y and z . Six regressions whose results are summarized in Table 9 were run. Each regression included all four variables either ungrouped or rounded to the nearest multiple of 0.25, 0.50, 0.75, 1.00 and 1.25 (the last roughly modeling a five point Likert scale applied to a population encompassing $6SD$). Grouping with widths that are comparable with variables' standard deviations (a) appreciably underestimate regression coefficients' magnitudes and coefficients of determination and (b) appreciably overestimate regression errors. In general, the direction and magnitude of the bias is not easy to predict, some regression coefficients may be biased away from zero.

Insert Table 8 and Table 9 about here

CONCLUSION

Grouped data, exemplified by Likert scales, are commonly used in the social sciences. The distribution of a sample collected using grouped data is different from the distribution of the population from which it was drawn. Estimates of population parameters obtained from grouped

samples may therefore be biased and/or inefficient. This paper has shown how researchers can modify statistical formulae, thereby compensating for the effect of grouping, and improving the power of statistical tests. Not using these modifications implies that estimates of population parameters are biased, that statistical tests are weaker than they need be, and the likelihood of type I and type II errors is increased. Not using these modifications is tantamount to discarding a fraction of expensively acquired data and wasting informants' time. The adjustments have trivial computational costs. The theory expounded has been confirmed by simulations comprising 10,000 observations.

We want to extend the analysis to other statistics commonly used in the social sciences for example, Cronbach's alpha and statistics underlying factor analysis. The proposed adjustments might be slightly improved by incorporating extra terms of the series used in Appendix A. There has been sharp, albeit dated, debate (Gaito, 1980) on the validity of applying interval statistics and tests to ordinal data (exemplified by data collected using Likert scales). Although it is theoretically wrong to apply parametric statistical techniques (those that rely on means and standard deviations) to ordinal data, doing so seems to produce hypothesis tests that are about as strong as non-parametric tests (e.g. the Mann-Whitney test of medians) that, strictly speaking, ought to be used on ordinal data. We believe that the considering grouped data might illuminate this issue. The effect of grouping on estimates of population parameters derived from small samples needs further consideration.

Table 1: A Varying Covariance Matrix

Variables	x	y	z
x	1	0	t
y	0	1	0
z	t	0	1

Table 2: Effect of Independent Variable Correlation on the Regression Error

Correlation coefficient	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Error in the regression coefficient	0.08	0.09	0.10	0.12	0.14	0.17	0.21	0.28	0.42	0.83

Table 3: Descriptive statistics

Variable	Mean	Std. Dev	Variance	
			Not corrected	Corrected
PO1e Total cost per unit of product	2.807	0.878	0.770	0.687
PL1 We have a mission statement which has been promulgated throughout the organization.	3.602	1.041	1.083	1.000
PL3 Our plans focus on the achievement of "best practice"	3.649	1.058	1.120	1.037
PL4 We always incorporate stakeholders requirements in plans	3.890	0.901	0.812	0.728
PL5 We have a written statement of strategy	3.370	1.156	1.337	1.253
PL6 Our site's manufacturing operations are aligned with strategy	3.870	0.898	0.806	0.723

Table 4: Covariance Matrix

	PL6	PL3	PL5	PL4	PL1
PL6	1.000				
PL3 o	-0.094	1.000			
PL5	-0.146	-0.154	1.000		
PL4	-0.224	-0.281	-0.171	1.000	
PL1	-0.203	-0.281	-0.278	-0.012	1.000

Table 5: Effects of grouping on regression coefficients and estimation errors

Variable	Constant	PL1	PL3	PL4	PL5	PL6
Results when the effect of grouping is ignored						
Standard error of estimate	0.16134	0.03499	0.03459	0.03894	0.03090	0.03852
Regression coefficient	2.17198	-0.01912	-0.05707	0.08489	0.02421	0.12920
T-ratio	13.46192	-0.54653	-1.64980	2.17979	0.78367	3.35388
Probability	0.0000	0.5848	0.0993	0.0295	0.4334	0.0008
Results when the effect of grouping is taken into account						
Standard error of estimate	0.1775	0.0404	0.0400	0.0457	0.0351	0.0450
Regression coefficient	2.09484	-0.02730	-0.07713	0.10236	0.02349	0.15873
T-ratio	12.49354	-0.69675	-1.99097	2.29325	0.69683	3.62103
Probability	0.0000	0.4996	0.0548	0.0250	0.4902	0.0004

Table 6: Effect of Grouping on Univariate Statistics

	0.0	0.25	0.5	0.75	1.0	1.25
Class Width	0.0	0.25	0.5	0.75	1.0	1.25
Mean x (theory)	0.0	0.0	0.0	0.0	0.0	0.0
Mean x (measured)	-0.004	-0.007	-0.005	-0.002	-0.003	0.004
Variance x (theory)	1.00	1.005	1.021	0.067	1.083	1.13
Variance x (measured)	1.054	1.058	1.077	1.096	1.146	1.174

Table 7: Effect of Grouping on Bivariate Statistics

	0.000	0.250	0.500	0.750	1.000	1.250
Class width	0.000	0.250	0.500	0.750	1.000	1.250
R Square (theory)	0.500	0.496	0.484	0.465	0.438	0.402
R Square (measured)	0.499	0.496	0.484	0.472	0.444	0.412
Standard Error Sq (theory)	1.000	1.005	1.021	1.046	1.080	1.123
Standard Error Sq (measured)	1.003	1.007	1.022	1.042	1.080	1.122
Intercept (theory)	3.000	3.000	3.000	3.000	3.000	3.000
Intercept (measured)	3.005	3.006	3.005	3.002	3.002	2.997
Beta (theory)	1.000	0.995	0.979	0.953	0.917	0.870
Beta (measured)	1.000	0.997	0.980	0.968	0.925	0.887

Table 8: Correlation Matrix for Multiple Regression Simulation

	x	y	z
x	1.000		
y	0.427308	1.000	
z	0.412166	0.275219	1.000

Table 9: Results of Simulations of Multiple Regression Analyses

Interval Width	0	0.25	0.5	0.75	1.00	1.25
Regression coefficient of x measured	1.0000	0.9969	0.9908	0.9798	0.9573	0.9348
Regression coefficient of x calculated	1.0000	0.9971	0.9885	0.9741	0.9539	0.9280
Regression coefficient of y measured	1.0000	0.9962	0.9907	0.9709	0.9533	0.9218
Regression coefficient of y calculated	1.0000	0.9972	0.9888	0.9748	0.9551	0.9299
Regression coefficient of z measured	1.0000	0.9982	0.9844	0.9731	0.9492	0.9411
Regression coefficient of z calculated	1.0000	0.9972	0.9886	0.9744	0.9544	0.9288
R sq measured	1.0000	0.9961	0.9851	0.9660	0.9427	0.9118
R sq calculated	1.0000	0.9967	0.9867	0.9700	0.9467	0.9167
Regression error squared measured	0.0000	0.0215	0.0818	0.1886	0.3199	0.4959
Regression error squared calculated	0.0000	0.0208	0.0833	0.1875	0.3333	0.5208

APPENDIX A DERIVATION OF SHEPPARD'S CORRECTION

This contains the underlying mathematics.

Notation

The following notation is used throughout.

m = the number of variables (denoted by x_1, x_2, \dots, x_m). It will sometimes be convenient to denote the co-ordinates of a point by the vector $\mathbf{x}_j = x_{1j}, x_{2j} \dots x_{mj}$

n = the number of groups into which measurements are resolved. The midpoint of the group indexed by j has co-ordinates $\mathbf{m}_j = m_{1j}, m_{2j} \dots m_{mj}$

The vector $\mathbf{h} = (h_1, h_2, \dots, h_m)$ comprises the class widths for each variable. The vector $\mathbf{r} = (r_1, r_2, \dots, r_m)$ (sometimes written $r(1), r(2), \dots, r(m)$) is a vector of integers used to express the r -th moment of the probability distribution about zero.

μ_r^G, μ_r^{NG} are respectively the r -th grouped and non-grouped moments about zero.

$\phi(\mathbf{x}) = \phi(x_1, x_2, \dots, x_m)$ is a probability density function.

$\phi_p(\mathbf{x})$ = The partial derivative of ϕ with respect of x_p evaluated at \mathbf{x} .

$\phi_{pq}(\mathbf{x})$ = The partial derivative of ϕ with respect of x_p and x_q evaluated at \mathbf{x} .

$\int_{\mathbf{u}}^{\mathbf{v}} \mathbf{f}(\mathbf{x}_j) d\mathbf{x}$ abbreviates $\int_{x_{1j}=u_1}^{v_1} \int_{x_{2j}=u_1}^{v_2} \dots \int_{x_{mj}=u_m}^{v_m} f(x_{1j}, x_{2j}, \dots, x_{mj}) dx_{mj} \dots dx_{2j} dx_{1j}$.

Derivation

The Euler-Maclaurin theorem (Aitken, 1957: 44) relates the integral of a function $\phi(x)$ over an interval (say $[a, b]$) and the sum of values of the function taken at points that are equally spaced (with class width h) over $[a, b]$. The difference between the integral and the sum is expressible as a function of odd derivatives of $\phi(x)$ taken at the boundaries. For two tailed statistical distributions, whose derivatives tend rapidly to zero as $x \rightarrow \pm\infty$, the approximation (9) is excellent.

$$\int_{-\infty}^{\infty} x^p \phi(x) dx \cong w \sum_{i=-\infty}^{\infty} m_i^p \phi(m_i) \quad (9)$$

where:

w is the width of the intervals into which the real line is divided,

m_i is a point belonging to the interval indexed by i ,

ϕ_s is the s – th derivative of ϕ .

Aitken(: 44), asserts that the approximation is excellent especially if $w \leq \sigma_x$, and gives the example $\int_{-\infty}^{+\infty} e^{-x^2/2} dx = 2.506628275$ and $\sum_{-\infty}^{+\infty} e^{-x^2/2} x = 0, \pm 1, \pm 2 \dots = 2.506628288$. The sum $\sum_{-\infty}^{+\infty} e^{-x^2/2} x = \pm 1/2, \pm 1 1/2, \pm 2 1/2 \dots = 2.506628261$.

It is necessary to extend the Euler-Maclaurin theorem to several variables. The appropriate generalisation (provable by mathematical induction) assuming that the error in (9) is negligible is:

$$\int_{\mathbf{x}=-\infty}^{\infty} \left(\prod_{i=1}^n x_j^{r(i)} \right) \phi(\mathbf{x}) d\mathbf{x} = \left(\prod_{i=1}^n h_i \right) \sum_{j=-\infty}^{\infty} \left(\prod_{i=1}^n m_j^{r(i)} \right) \phi(\mathbf{m}_j). \quad (10)$$

We apply Aitken's method to relate grouped and ungrouped statistics for a probability density function of m variables. Because the data is available only in grouped form, the experimenter maps every observation in a cell to that cell's midpoint (\mathbf{m}_j).

The probability of an observation being in the j -th group is:

$$p_j = \int_{\mathbf{m}_j-h/2}^{\mathbf{m}_j+h/2} \phi(\mathbf{x}) d\mathbf{x}. \quad (11)$$

The distribution's r -th moment around zero calculated on the basis of the groups' midpoints is:

$$\mu_r^G = \sum_{j=1}^n \left(\prod_{i=1}^m m_{ij}^{r(i)} \right) p_j. \quad (12)$$

If the data were exactly known (i.e. not grouped) the distribution's r -th moment around zero (the true or non-grouped moment) would be calculated as:

$$\mu_r^{NG} = \int_{-\infty}^{\infty} \left(\prod_{i=1}^m x_i^{r(i)} \right) \phi(\mathbf{x}) d\mathbf{x}.$$

Consider equation (11). By substituting $\mathbf{m}_j + \mathbf{x}'$ for \mathbf{x} and dropping the dash this becomes $p_j = \int_{-h/2}^{+h/2} \phi(\mathbf{m}_j + \mathbf{x}) d\mathbf{x}$. Taylor's theorem can be used to expand $\phi(\mathbf{m}_j + \mathbf{x})$ about \mathbf{m}_j , yielding:

$$\phi(\mathbf{m}_j + \mathbf{x}) = \phi(\mathbf{m}_j) + \sum_{i=1}^m x_i \phi_i(\mathbf{m}_j) + \frac{1}{2} \sum_{i=1}^m \sum_{k=1}^m x_i x_k \phi_{ik}(\mathbf{m}_j) + o(\mathbf{m}_j^3) \quad \text{so}$$

$$p_j = \int_{-h/2}^{+h/2} \left\{ \phi(\mathbf{m}_j) + \sum_{i=1}^m x_i \phi_i(\mathbf{m}_j) + \frac{1}{2} \sum_{i=1}^m \sum_{k=1}^m x_i x_k \phi_{ik}(\mathbf{m}_j) \right\} d\mathbf{x}$$

If third and higher order terms are dropped. Integrating by parts with respect to x_1, x_2, \dots, x_m between $-h_1/2$ and $h_1/2, -h_2/2$ and $h_2/2, \dots -h_m/2$ and $h_m/2$ we obtain

$$p_j = \left(\prod_{i=1}^m h_i \right) \left[\phi(\mathbf{m}_j) + \frac{1}{24} \sum_{i=1}^m h_i^2 \phi_{ii}(\mathbf{m}_j) + \sum_{i=1}^m o(h^4) \right]. \quad (13)$$

It is worthwhile interpreting (13). The factor $\prod_{i=1}^m h_i$ is the hyper-volume of each of the hyper-cuboid cells into which uniform grouping divides the domain of measurement. The term $\phi(\mathbf{m}_j)$ is the probability density at the centre of cell j . The term, $\left(\prod_{i=1}^m h_i\right)\phi(\mathbf{m}_j)$ the product of the cell volume and $\phi(\mathbf{m}_j)$ approximates the probability of an observation occurring in that cell. The term $\left(\prod_{i=1}^m h_i\right)\left[\frac{1}{24}\sum_{i=1}^m h_i^2\phi_{ii}(\mathbf{m}_j)\right]$ is a second order adjustment to that probability. It is easy to see (by symmetry) that the first order adjustment is zero.

Substituting (13) into (12) we obtain:

$$\begin{aligned}\mu_r^G &= \sum_{j=1}^n \left(\prod_{i=1}^m m_{ij}^{r(i)}\right) \left(\prod_{i=1}^m h_i\right) \left[\phi(\mathbf{m}_j) + \frac{1}{24}\sum_{i=1}^m h_i^2\phi_{ii}(\mathbf{m}_j) + \sum_{i=1}^m o(h^4)\right] \\ \mu_r^G &= \left(\prod_{i=1}^m h_i\right) \left[\sum_{j=1}^n \left(\prod_{i=1}^m m_{ij}^{r(i)}\right) \phi(\mathbf{m}_j) + \frac{1}{24}\sum_{j=1}^n \left(\prod_{i=1}^m m_{ij}^{r(i)}\right) \sum_{i=1}^m h_i^2\phi_{ii}(\mathbf{m}_j) + \sum_{j=1}^n \left(\prod_{i=1}^m m_{ij}^{r(i)}\right) \sum_{i=1}^m o(h^4)\right] \quad (14)\end{aligned}$$

Applying (10) in reverse to the first two terms of (14) and dropping the negligible third term we obtain:

$$\mu_r^{NG} = \int_{-\infty}^{\infty} \left(\prod_{i=1}^m x_i^{r(i)}\right) \phi(\mathbf{x}) d\mathbf{x} + \frac{1}{24}\sum_{i=1}^m h_i^2 \int_{-\infty}^{\infty} \left(\prod_{i=1}^m x_i^{r(i)}\right) \phi_{ii}(\mathbf{x}) d\mathbf{x} + \dots \quad (15)$$

The first term is μ_r^G , the ungrouped moment order r_i $i=1\dots m$. The integral in the second and subsequent terms can be integrated by parts assuming that ϕ and its derivatives vanish at $\pm\infty$ to give:

$$\begin{aligned}\mu_r^G &= \mu_r^{NG} + \frac{1}{24}\sum_{i=1}^m h_i^2 r_i (r_i - 1) \mu_{r_1 r_2 \dots (r_i - 2) \dots r_m}^{NG} + \frac{1}{80}\sum_{i=1}^m h_i^2 r_i (r_i - 1)(r_i - 2)(r_i - 3) \mu_{r_1 r_2 \dots (r_i - 4) \dots r_m}^{NG} + \dots \quad \text{defining} \\ \mu_r^{NG} &= 0 \text{ if } \forall r < 0. \quad (16)\end{aligned}$$

This demonstrates that grouped moments of order two or more (and hence variances) in any variable are biased by grouping but that means and covariances are not affected by grouping.

APPENDIX B POPULATION MULTIPLE REGRESSION WITH GROUPING

To explain the effect of grouping on regression analyses we use the notation and methods of (Gujarati, ch 9)

Notation

The following notation is required.

N = The number of observations

$\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_k)$ a $k+1$ vector of regression coefficients (β_0 is the constant term).

$\mathbf{y} = (Y_1, Y_2, \dots, Y_N)$ is a vector of observations of the dependent variable.

X_{ij} $i \in I = \{0, 1, \dots, k\}$, $j \in J = \{1, 2, \dots, N\}$ a $((k+1) \times N)$ matrix of observations of the independent variables. Note that $X_{0j} = 1$ (the constant term).

h_0 = group width of the dependent variable.

$H = \text{Diag}\{0, h_1^2, h_2^2, \dots, h_k^2\}$ a diagonal matrix comprising zero and the squares of the independent variables' group widths.

$\mathbf{u} = \{u_1, u_2, \dots, u_N\}$ an error term.

The population regression model is:

$$y_j = \sum_{i=0}^k \beta_i X_{ij} + u_j, j = \{1, 2, \dots, N\} \text{ or, in matrix form, } \mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}.$$

The purpose of regression analysis is to estimate $\boldsymbol{\beta}$ by minimising some function of \mathbf{u} , usually the sum of the squares of the errors $\mathbf{u}^T \mathbf{u}$, using some standard assumptions detailed in Gujarati (1988, section 9.2). The estimate of $\boldsymbol{\beta}$ that best minimises $\mathbf{u}^T \mathbf{u}$ is:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}. \quad (17)$$

The matrix $\mathbf{Z} = \mathbf{X}^T \mathbf{X}$ has shape $(k+1) \times (k+1)$, is symmetric and has the simple structure:

$$\mathbf{Z} = \begin{bmatrix} N & \sum X_{1i} & \sum X_{2i} & \dots & \sum X_{ki} \\ \sum X_{1i} & \sum X_{1i} X_{1i} & \sum X_{1i} X_{2i} & \dots & \sum X_{1i} X_{ki} \\ \sum X_{2i} & \sum X_{1i} X_{2i} & \sum X_{2i} X_{2i} & \dots & \sum X_{2i} X_{ki} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sum X_{ki} & \sum X_{ki} X_{1i} & \sum X_{ki} X_{2i} & \dots & \sum X_{ki} X_{ki} \end{bmatrix} \quad (18)$$

As shown in Appendix A, the only terms of \mathbf{Z} whose estimates are affected by grouping are those involving squares of variables, i.e. only those on the main diagonal (except for the element N). The adjustment for grouping entails replacing \mathbf{Z} by

$$\hat{\mathbf{Z}} = \mathbf{Z} - \mathbf{NH}/12, \quad (19)$$

An improved estimate of β is

$$\hat{\beta} = \hat{\mathbf{Z}}^{-1} \mathbf{X}^T \mathbf{y} \quad (20)$$

It can be shown that an improved estimate of the squared error of the estimate is

$$\hat{\sigma}^2 = \hat{\sigma}^2 + \frac{1}{12} \left(h_0^2 + \sum_{i=1}^k (\beta_i^{NG})^2 h_i^2 \right). \quad (21)$$

The variances of estimates of $\hat{\beta}(\mathbf{v})$ are given by the diagonal elements of $\sigma^2 \mathbf{Z}^{-1}$ (Gujarati, 1988: 258). An improved estimate of \mathbf{v} is given by the diagonal elements of:

$$\hat{\mathbf{v}} = \mathbf{v} + \frac{1}{12} \left[-N\sigma^2 \mathbf{H} \mathbf{Z}^{-1} + h_0^2 + \sum_{i=1}^k \beta_i^2 h_i^2 \right] \mathbf{Z}^{-1} \quad (22)$$

REFERENCES

- Aitken, A. C. (1957). *Statistical Mathematics* (8th ed.). London: Oliver and Boyd.
- AMC. (1994). *Leading the Way: A study of best manufacturing practices in Australia and New Zealand*. Melbourne: Australian Manufacturing Council.
- Australian Manufacturing Council. (1994). *Leading the way: a study of best manufacturing practices in Australia and New Zealand*. Melbourne: Australian Manufacturing Council.
- Bollen, K. (1989). *Structural Equation with Latent Variables*. New York: John Wiley.
- Bollen, K. A., & Barb, K. H. (1981). Pearson's R and Coarsely Categorized Measures. *American Sociological Review*, 46(2), 232-239.
- Bollen, K. A., & Barb, K. H. (1983). Collapsing Variables and Validity Coefficients (in Comments). *American Sociological Review*, 48(2), 286.
- Cameron, T. A. (1987). The Impact of Grouping Coarseness in Alternative Grouped-data Regression Models. *Journal of Econometrics*, 35, 37-57.
- Caudill, S. B., & Jackson, J. D. (1993). Heteroscedasticity and Grouped Data Regression. *Southern Economic Journal*, 60(1), 128-153.
- Cohen, J. (1983). The cost of dichotomization. *Applied Psychological Measurement*, 7(3), 249-253.
- Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences* (2nd ed.): Erlbaum.
- Gaito, J. (1980). Measurement scales and statistics: Resurgence of all old misconception. *Psychological Bulletin*, 87, 564-576.
- Gephart, R. P. J. (1983). Multiple r, the "parametric strategy" and measurement imprecision. *Sociological Perspectives*, 26(4), 473-500.
- Greene, W. H. (1997). *Econometric analysis* (3rd ed.). New Jersey: A. Simon & Schuster.
- Gujarati, D., N. (1988). *Basic Econometrics* (2nd ed.). Singapore: McGraw-Hill.
- Haitovsky, Y. (1968). Regression Analysis of Grouped Observations when the Cross Classifications are Unknown. *Review of Economic Studies*, 35(1), 77-89.
- Haitovsky, Y. (1973). *Regression estimation from grouped observations* (Vol. 33). London: Griffin.
- Kakwani, N. (1993). The Coefficient of Determination for a Regression Model Based on Group Data. *Oxford Bulletin of Economics and Statistics*, 52(2), 245-251.
- Kendall, M., & Stuart, A. (1979). *The advanced theory of statistics* (4th ed. Vol. 2). London: Charles Griffin & Co., Ltd.
- Lindley, D. V. (1950). Grouping corrections and maximum likelihood equations. 106-110.
- Sheppard, W. F. (1898). On the calculation of the most probable values of frequency constants for data arranged according to equidistant divisions of a scale. *Proc. London Math. Soc.*, 29, 353-380.
- Stewart, M. B. (1983). On Least Square Estimation when the Dependent Variable is Grouped. *Review of Economic Studies*, 50, 737-753.
- Whittaker, E. T., & Robinson, G. (1967). Sheppard's Corrections. In *The Calculus of Observations: A Treatise on Numerical Mathematics* (4 ed., pp. 194-196). New York: Dover.
- Wolfram, S. (1991). *Mathematica: A system for doing mathematics by computer* (2 ed.). Redwood City Ca: Addison-Wesley.

ACKNOWLEDGEMENTS

I am indebted to many colleagues in my department for their suggestions for improvements in this paper's expression. Professor Ralph Snyder and Dr Lee Gordon-Brown made many cogent comments on the underlying mathematics.