ELSEVIER

Journal of Clinical Epidemiology 53 (2000) 1036–1043

Journal of Clinical Epidemiology

Analyzing data with clumping at zero
An example demonstration

Bei-Hung Chang[a,b,*], Stuart Pocock[a,c]

[a]*New England Research Institutes, Watertown, MA, USA*
[b]*Center for Health Quality, Outcomes, and Economic Research, Bedford VAMC, 200 Springs Road, Bedford, MA 01730, USA*
[c]*Medical Statistics Unit, London School of Hygiene and Tropical Medicine, London, UK*

Received 27 August 1998; received in revised form 28 January 2000; accepted 31 January 2000

Abstract

This article demonstrates the use of two approaches to analyzing the relationship of multiple covariates to an outcome which has a high proportion of zero values. One approach is to categorize the continuous outcome (including the zero category) and then fit a proportional odds model. Another approach is to use logistic regression to model the probability of a zero response and ordinary least squares linear regression to model the non-zero continuous responses. The use of these two approaches was demonstrated using outcomes data on hours of care received from the Springfield Elder Project. A crude linear model including both zero and non-zero values was also used for comparison. We conclude that the choice of approaches for analysis depends on the data. If the proportional odds assumption is valid, then it appears to be the method of choice; otherwise, the combination of logistic regression and a linear model is preferable.

Keywords: Zero values; Proportional odds model; Regression analysis

1. Introduction

One difficulty statisticians often encounter is in analyzing outcomes that have a substantial proportion of zero values. An example from studies of disabled elderly is the hours of care elders received. Measures for various types of care often have a substantial proportion of zero values, while the remaining values are continuous which sometimes appear lognormally distributed. Because the distribution of the outcome measures cannot be made close to any commonly defined distribution (e.g., normal or lognormal), it creates difficulties in applying an ordinary least squares linear regression model to analyze these outcomes.

In this example, the occurrence of zero hours of care is not due to truncation or self-selection. Instead, zero values are actual outcome values and it is important to account for these zero values explicitly in the analysis. Methods such as the Tobin model [1] and the self-selection model [2,3] which assume truncation or censoring in the outcome measure are therefore not appropriate for this type of data. The Tobin model assumes that the outcome measures follow a censored normal distribution. The self-selection model also assumes that the outcome measure is censored and estimates what the relationship between some covariates and the outcome measure would have been if the outcome measure were not censored.

More recently, several approaches have been proposed to analyzing data with this feature. These approaches analyze zero values explicitly and do not require the assumption of censoring on the outcome measure. One approach used by Saei et al. [4] is to apply threshold models to outcome variables that have a large number of zero values and the remaining values are positive and continuous. They recommend recoding the continuous response into an ordinal scale by grouping the positive values into intervals, and then apply threshold models to relate the ordinal outcome variable to covariates. Green [5] showed that even with a very coarse grouping in the continuous response, the estimates of the regression coefficients of the covariates and their standard errors are close to those for the ungrouped data. Therefore, to avoid estimating a large number of nuisance parameters, which is required when continuous response is used, a grouped response variable is used for threshold models. One crucial assumption of this approach is that the relationship between the cumulative probabilities of the ordinal categories and the covariates should be the same for each category of the outcome variable. This proportional odds assumption can be verified by statistical tests (e.g., score test and Wald test) or graphical methods. The score test and a graphical

* Corresponding author. Tel: 781-275-7500, ext 6007; fax: 781-687-3106.

E-mail address: bhchang@bu.edu (B.-H. Chang)

0895-4356/00/$ – see front matter
PII: S0895-4356(00)00223-7

method developed by Harrell and colleagues [6] are described in the methods section. A simulation study conducted by Peterson and Harrell [7] showed that both the score and Wald tests give erroneous or suspicious results when the cross-tabulation table for the outcome variable by a covariate contains empty cells or when data are sparse in the cells of the cross-tabulation table. The score test results might also be invalid when the number of observations at one of the level of the outcome variable is small relative to the total sample size. Based on these results, to have a valid test for the proportional odds assumption it is desirable to group the continuous variable especially when the number of unique values is large.

Another approach is proposed by Lachenbruch with the notion that the covariates which predict whether the outcome measure is zero might be different from the covariates which predict the level of the outcome measure given the outcome is non-zero. A two-degree of freedom test was developed by Lachenbruch [8] in 1976 for a two-sample testing problem. One degree of freedom is for testing the equality of proportion of zeros between two groups. The other degree of freedom is for testing the equality of non-zero observations. Lachenbruch [9] extended the two-sample testing problem (one independent variable model) to a multivariate model. In the multivariate setting, logistic regression or probit regression models can be used to model the probability of a zero response, while ordinary least squares regression can be used to model the non-zero responses.

In this article, we demonstrate the use of these two approaches using the example of the Springfield Elder Project (SEP) [10]. We also demonstrate that it may be possible to use a crude least squares regression to the overall outcome with both zero and non-zero values included.

2. Example

The Springfield Elder Project is a research study concerned with older persons in three target populations (Puerto Rican, African American, and White), their needs for long-term care assistance, and the sources and amounts of that help. One aim of the study is to identify the correlates of amount of services disabled elders received from their caregivers who are either relatives or friends of the elders. Functionally disabled elders received services in various types of daily activities. Hours of services in personal care and housekeeping were chosen as the example outcome measures for this article. The independent variables investigated included elders' ethnicity (Puerto Rican, African American, and White), elders' disability level (1–13: higher values indicate higher disability level), elders' socioeconomic status (0–100: higher values indicate higher status), caregiver gender, caregiver employment (yes/no), and coresidence of caregiver with elder (yes/no). Two indicator variables were used for ethnicity with White as the reference group. Data were collected by telephone interview from 409 caregivers of the disabled elders.

Among 392 elders who had non-missing data in hours of personal care, 185 (47%) received no (zero hours of) service, while the other 53% received between less than one hour to 165 hours in a month. Of the 377 elders with non-missing data in hours of housekeeping, 35 (9%) received zero hours, while the others received between less than one hour to 240 hours in a month. The distributions of both outcomes appear to be skewed and are not close to any commonly defined distribution such as longnormal or Poisson (Fig. 1). When the outcomes are transformed as log (hours+1), the non-zero hours of housekeeping appear close to a normal distribution, but there is still a spike in the zero hour. To apply the threshold model, the personal care hours (P) were grouped into four categories: $P = 0$, $0 < P < 3$, $3 \leqslant P < 15$, and $P \geqslant 15$. The counts in each category were 185, 71, 76, and 60, respectively. The housekeeping hours (h) were also grouped into four categories: $h = 0$, $0 < h < 15$, $15 \leqslant h < 30$, and $h \geqslant 30$. The counts in each category were 35, 143, 83, and 116, respectively. These groupings were chosen so that the counts in each of the non-zero categories were roughly equal.

3. Statistical methods

The first approach used here follows the idea of Saei *et al.* [4] who used a threshold model to analyze outcomes of a methadone randomized controlled trial. A threshold model is developed for categorical data which are measured on an ordinal scale. As discussed by McCullagh and Nelder [11], the threshold model may be derived from the notion of an unobserved continuous random variable Z, such that $Z + X\beta$ has the cumulative distribution function G. If the unobserved variable lies in the interval $\theta_{j-1} < Z \leqslant \theta_j$ then response $Y = j$ is observed, where θ_j are threshold parameters, $j=1,...,k$ levels. Thus the threshold model is derived as

$$\begin{aligned} \Pr(Y \leqslant j|X) &= \Pr(Z \leqslant \theta_j|X) \\ &= \Pr(Z + X\beta \leqslant \theta_j + X\beta) \\ &= G(\theta_j + X\beta), \quad j = 1,2,...,k-1. \end{aligned} \tag{1}$$

Examples of G are the cumulative standard normal distribution, the cumulative standard logistic distribution, and the cumulative extreme-value function (also called the Gompertz distribution). When G is the cumulative standard logistic distribution [12], Eq. (1) can be expressed as

$$\Pr(Y \leqslant j|X) = \frac{\exp(\theta_j + X\beta)}{1 + \exp(\theta_j + X\beta)}.$$

This can be written as the familiar proportional odds model (also called the ordinal logistic regression model) in the generalized linear model format,

$$\text{Logit}\,(Y \leqslant j|X) = \theta_j + X\beta, \quad j = 1,2,....,k-1. \tag{2}$$

The name proportional odds is given because the ratio of the odds of the event $Y \leqslant j$ at two values of X is independent of the choice of category (j). In other words, $k - 1$ parallel re-

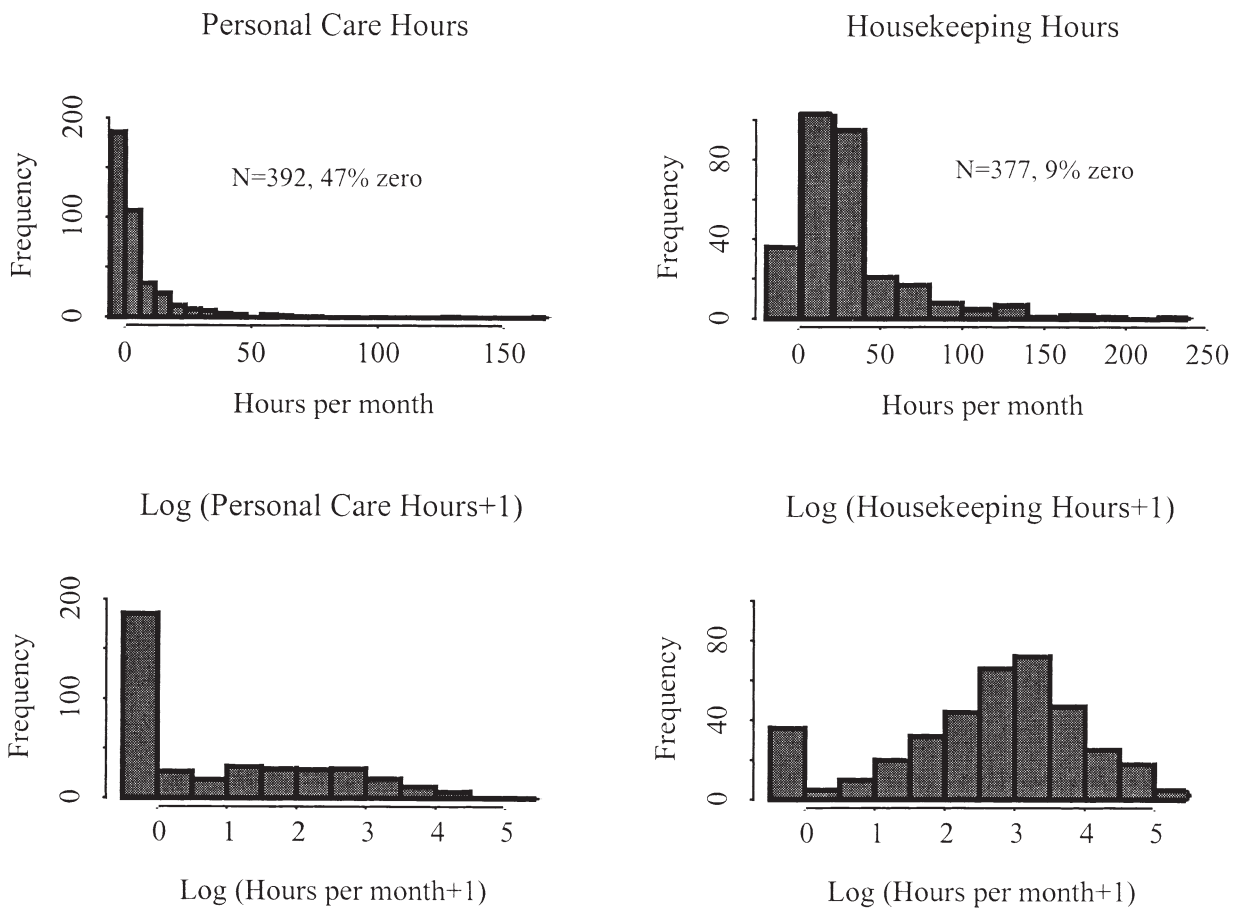

Fig. 1. Distributon of outcome measures.

gression lines are assumed for k categories of the response variable. To test this parallel lines assumption, suppose there is only one covariate X in the model and consider the model with non-parallel lines:

$$\text{Logit}\,(Y \leqslant j|X) = \theta_j + X\beta_j, \quad j = 1,2,.....k-1.$$

Under the parallel assumption, $\beta_1 = \beta_2 = ... = \beta_{k-1}$. Let $\hat{ß}$ be the MLE of the slope parameter under the parallel lines assumption. The score statistic that is evaluated at $\hat{ß}$ for all k-l slope parameters has an asymptotic χ^2 distribution with k-2 degrees of freedom. This score statistic can be used to test the parallel lines assumption. Similarly, a score test can be derived when the model has more than one (say q) covariates. In this case, the score statistic has an asymptotic χ^2 distribution with q(k-2) degrees of freedom. The procedure PROC LOGISTIC in the SAS computer package fits the proportional odds model and tests for the global proportional odds assumption [13] using this score test.

Another graphical approach for checking the proportional odds assumption has been developed by Harrell and colleagues [6] using a residual plot for each covariate. In this approach, the score residual for cut-off point j in Y is calculated as the component U_{im} of the score statistic (first derivative of the log likelihood function).

$$U_{im} = X_{im} * ([Y_i \geqslant j] - P_{ij})$$

for subject i and covariate X_m,

where $[Y_i \geqslant j]$ is an indicator variable with value 1 if $Y_i \geqslant j$, otherwise 0, and $\hat{P}_{ij}$ is the predicted probability of $Y_i \geqslant j$ estimates from the proportional odds model. Then the mean score residual ($\bar{U}_{.m}$) and its 95% confidence interval are calculated for each covariate. An S-plus function written by Harrell and colleagues [11] plots these means score residuals and their 95% confidence intervals versus the cut-off point j in Y (j=2,..K). If the proportional odds assumption holds, mean score residuals would be close to zero and without a trend over increasing j for each covariate.

As noted, the proportional odds model is used for categorical outcomes measured on an ordinal scale. The corresponding model for continuous outcomes which has the similar assumption as for proportion odds and has format similar to Eq. (2) is the linear model with log link function, (i.e. the model):

$$\text{Log}\,(Y|X) = \beta_0 + X\beta.$$

Therefore, if the derived ordinal outcomes fit the proportional odds model, it is likely that the original continuous outcomes, from which the ordinal scales are derived, might fit the above linear model.

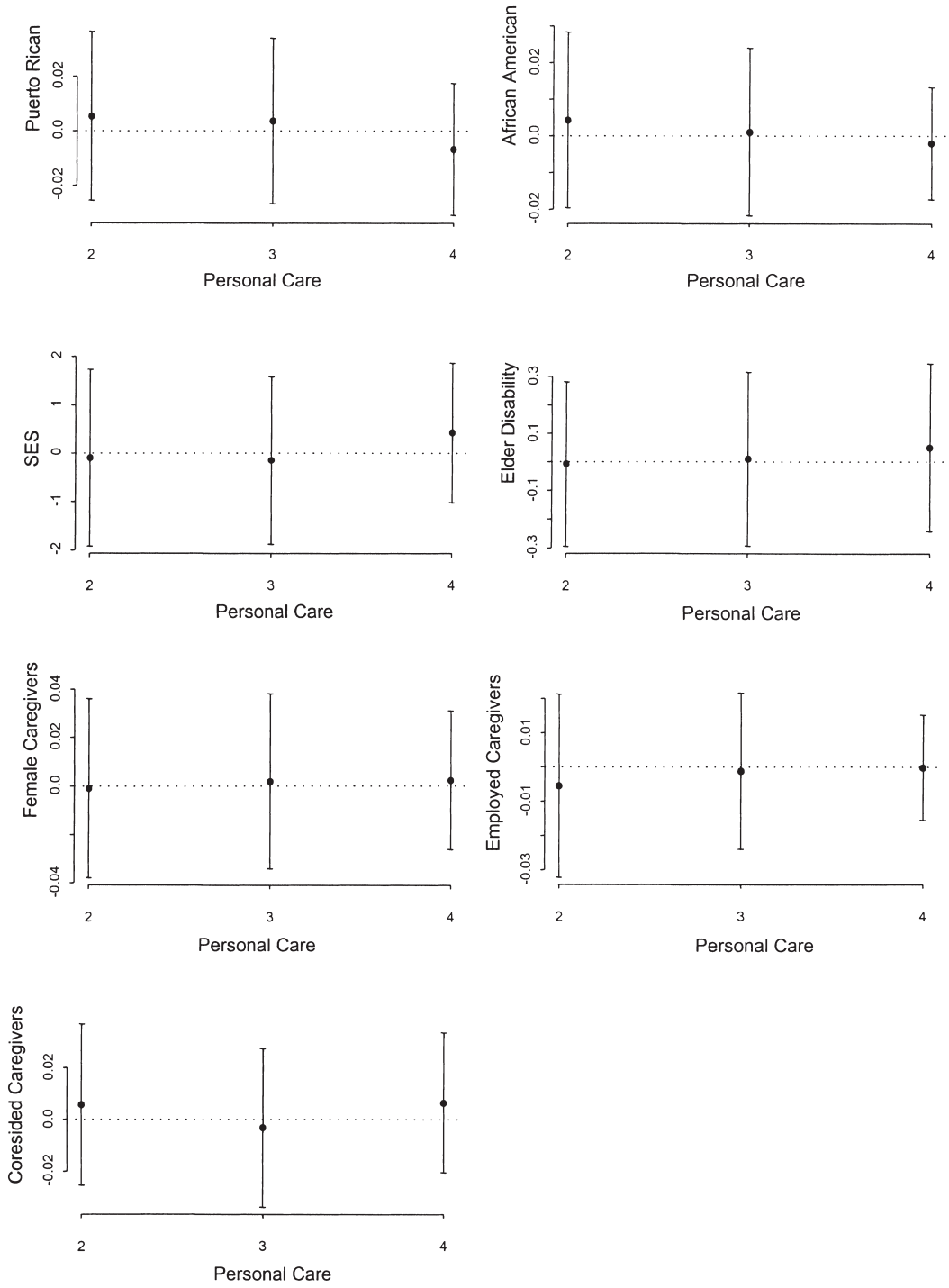

Fig. 2. Mean score residual of proportional odds model with 95% confidence interval for each cut-off point of personal care hours.

If the proportional odds assumption is not valid and the ordinal dependent variable includes a zero category, then it is possible that the predictors of zero response are different from those that predict non-zero levels. In this case, two separate models would be preferable. One model can be used to predict zero versus non-zero responses. Logistic or probit regression models are commonly used for this purpose. The other model is applied to the non-zero values using an ordinary least squares linear regression model which might require a suitable transformation of the non-zero values. For the SEP data, the non-zero hours appear close to a lognormal distribution, so ordinary least squares regression was applied to the log transformed outcome.

4. Results

The proportional odds model was first fitted to both outcomes: personal care hours and housekeeping hours with

Table 1
Results for personal care hours

a Proportional odds model

Variables	β	SE (β)	P-value	Odds ratio (e^{β})
Intercept1	−5.130	0.532	0.0001	–
Intercept2	−3.780	0.501	0.0001	–
Intercept3	−2.817	0.484	0.0001	–
Puerto Rican	0.268	0.308	0.3837	1.308
African American	−0.319	0.307	0.2994	0.727
Socioeconomic status	0.008	0.005	0.1246	1.008
Elder disability	0.306	0.033	0.0001	1.358
Female caregivers	0.709	0.241	0.0033	2.031
Employed caregivers	−0.333	0.223	0.1359	0.717
Coresided caregivers	0.523	0.216	0.0152	1.688

b Ordinary least squares regression of log(hours+1): included both zero and non-zero hours

Variables	β	SE (β)	P-value	Mean ratio (e^{β})
Intercept	−0.656	0.245	0.0077	–
Puerto Rican	0.051	0.169	0.7632	1.052
African American	−0.261	0.163	0.1110	0.770
Socioeconomic status	0.005	0.003	0.0782	1.005
Elder disability	0.185	0.016	0.0001	1.203
Female caregivers	0.416	0.128	0.0012	1.516
Employed caregivers	−0.215	0.121	0.0767	0.807
Coresided caregivers	0.304	0.118	0.0102	1.355

four categories each, as defined above (0, 0–3, 3–15, ⩾15 for personal care: 0, 0–15, 15–30, ⩾30 for housekeeping). The categories were arranged in the order such that large values of the regression coefficient (β) indicate increased probability of being in a larger hours category. The score test for the global proportional odds assumption is nonsignificant $\chi^2 = 9.63$ with 14 *df*, P = 0.79) for personal care hours which supports the proportional odds assumption. The proportional odds assumption for each covariate is further supported by the graphic approach developed by Harrell et al. The score residual plot indicates that the residuals are not significantly different from zero at each cut-off point j (=2,3,4) of personal care hours and no trend with j was observed for any of the independent variables (see Fig. 2). The results (Table 1, part a) show that higher levels of elder disability, having a female caregiver and coresidence of caregiver are all significantly associated with more hours of personal care. The regression coefficients estimated from the proportional odds model indicate that the odds ratio for the personal care time of 15 hours or greater is 1.36 for each one-point increase in the elder's disability level. The odds ratio (OR) for the personal care time of 15 hours or greater is 2.03 for female caregivers versus male caregivers and the odds ratio is 1.67 for coresided versus non-coresided caregivers. The odds ratios based on the other two cut-off points (i.e., greater than 3 hours or greater than 0 hours) are the same as those based on the cut-off point of 15 hours, because the proportional odds model was applied.

As mentioned in the statistical methods section, because the proportional odds assumption appears appropriate for the derived ordinal personal care hours, we expect that a log linear model may be used for the continuous personal care hours. So, the ordinary least squares (OLS) regression model was then fitted to the personal care hours with log (hours+1) as the dependent variable. The constant one was arbitrarily chosen to be added to the observed hours values for log transformation due to the existence of zero hours. This is an easy and commonly used solution for transforming data with zero value. While a sophisticated and complex method for choosing an optimal constant to be added is available [14], for the purpose of this article, we chose the simpler approach. This OLS regression gives the same three significant covariates (with similar P-values) as those using a proportional odds model (Table 1, part b). However the interpretation of the regression coefficients is different when using an OLS regression model compared to using a proportional odds model. The exponential of the regression coefficient estimated from the OLS regression can be interpreted as the ratio of mean outcome (MR) of two groups,

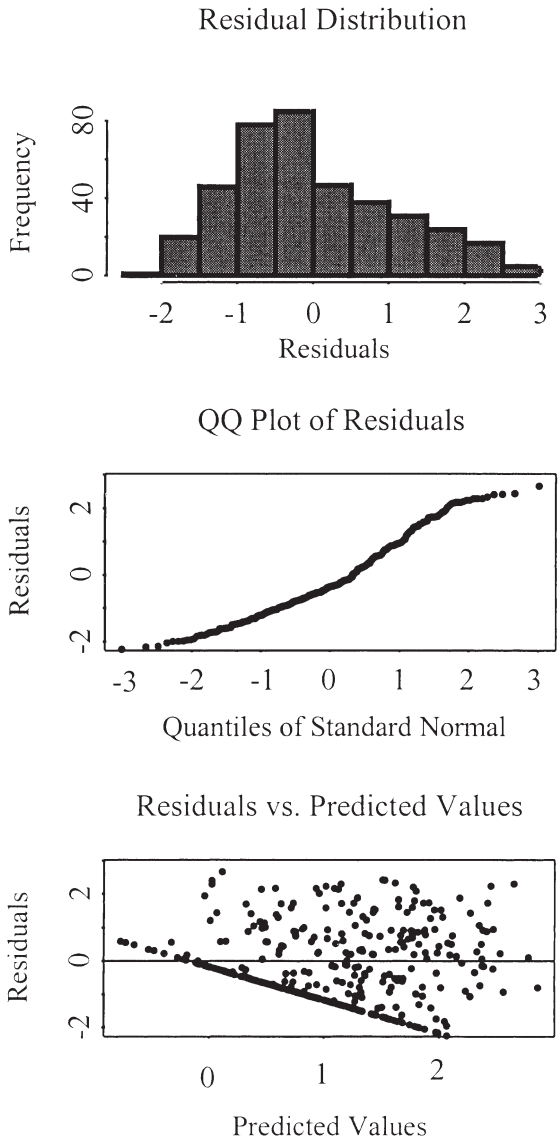

Fig. 3. Residuals of OLS log linear model for personal care hours (all data included in the model).

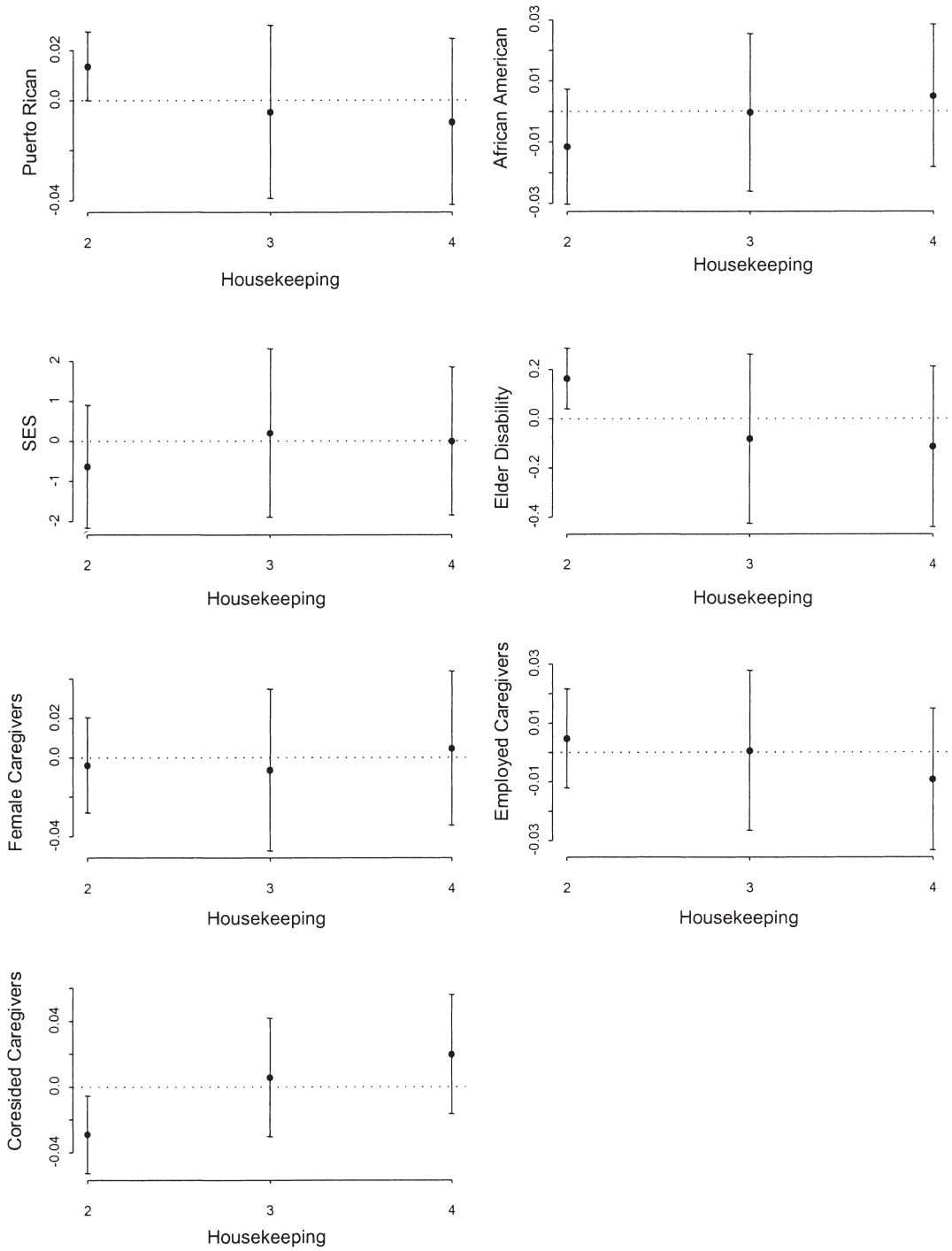

Fig. 4. Mean score residual of proportional odds model with 95% confidence interval for each cut-off point of housekeeping hours.

the "mean" in this instance is close to the geometric mean, being antilog [mean log (hours+1)]. The ratio of mean hours of personal care is 1.20 for each one-point increase in the elder's disability level, 1.51 for female versus male caregiver, and 1.36 for coresided versus non-coresided caregivers. One worthwhile observation is that despite the fact that the distribution of log (personal care hours+1) has a large count at zero such that the distribution is not close to a normal distribution (Fig. 1), the residuals appear much closer to a normal distribution (Fig. 3). However, the plot of residuals versus predicted values has a boundary line corresponding to the zero outcome because residuals are equal to the negative values of predicted values.

Unlike personal care hours, the housekeeping hours outcome does not fit the assumption of the proportional odds model. The score test ($\chi^2 = 51.65$, $df = 14$, and P-value =

Table 2
Results for housekeeping hours

a Logistic regression

Variables	β	SE (β)	P-value	Odds ratio (e^{β})
Intercept	0.752	0.790	0.3412	–
Puerto Rican	1.116	0.596	0.0611	3.053
African American	0.073	0.455	0.8720	1.076
Socioeconomic status	−0.010	0.009	0.2937	0.990
Elder Disability	0.372	0.081	0.0001	1.450
Female caregivers	−0.124	0.431	0.7738	0.883
Employed caregivers	0.469	0.428	0.2732	1.598
Coresided caregivers	−0.470	0.437	0.2825	0.625

b OLS Linear Regression of log(hours+1): only include non-zero hours

Variables	β	SE (β)	P-value	Mean ratio (e^{β})
Intercept	1.957	0.250	0.0001	–
Puerto Rican	0.238	0.169	0.1606	1.270
African American	0.224	0.166	0.1796	1.251
Socioeconomic status	0.006	0.003	0.8337	1.006
Elder Disability	0.029	0.016	0.0788	1.029
Female caregivers	0.317	0.126	0.0122	1.373
Employed caregivers	−0.015	0.118	0.8993	0.985
Coresided caregivers	0.579	0.113	0.0001	1.784

0.0001) indicates that the global proportional odds assumption is not valid. The graphic approach identified two covariates, elder disability and coresided caregivers, which showed evidence of non-proportional odds (Fig. 4). For these two variables, a trend was observed in score residuals against the cut-off point j in housekeeping hours. In particular, the score residuals differ significantly from zero at the cut-off point of 2 for both variables, this cut-off dividing the housekeeping hours into zero hours versus non-zero hours. Therefore, the residual plots suggest that the covariates which predict non-receiving care (zero hour) could be different from covariates which predict level of non-zero hours. The results of subsequent analysis using a logistic regression and a log linear regression support this argument. The results of the logistic regression indicate that elders' disability level predicts whether housekeeping care was provided (OR = 1.45; Table 2, part a), while the results of the log linear regression indicate that gender of caregivers (MR = 1.37) and coresidence between elders and caregivers (MR = 1.78) predict the number of hours given that care is provided (Table 2, part b). These results imply that the elder characteristics, namely disability level, determines whether the care is given. Once the care is given, the caregivers characteristics determine the amount of care that was given. These results also agree with the results of the score residual plots which indicate that elder disability and coresided caregivers do not fit the proportional odds assumption when both zero and non-zero hours are included in the model.

5. Discussion

Although two approaches, proportional odds model and the combination of logistic regression and OLS regression, have been previously recommended for analyzing outcomes with a large number of zero values, we have demonstrated that each approach is appropriate only under certain conditions. The proportional odds model is appropriate only when the proportional odds assumption is valid, and failing this, the combination of logistic regression and OLS regression model is preferable. We have also demonstrated that when the proportional odds assumption is valid for the ordinal outcome, the log linear model for the corresponding continuous outcome should give very similar results. Therefore, it is suggested that the analysts consider using the basic OLS regression model when the observed outcome variables deviate from a normal distribution, as long as a careful examination of residuals shows their distribution to be satisfactory.

Whereas the proportional odds model has the advantage of being able to include both zero and non-zero values in one model, it has potential drawbacks when applied to continuous outcomes. Before applying a proportional odds model, the proportional odds assumption needs to be tested for validity. To avoid an invalid test, it is desirable to group the continuous outcome variable so that the sample size in each level of the outcome variable is sufficiently large. However, the validity of the proportional odds assumption may depend on the cut-off points chosen for the grouping. Another difficulty is the choice of number of groups. To investigate how the estimated regression coefficients and their standard errors vary when different numbers of groups are used, we ran

Table 3
Estimates of regression coefficients and their standard errors in the proportional odds model using various cut-off points for personal care hours

	β	β		SE (β)	SE (β)	
Variables	4 categories	5 categories	6 categories	4 categories	5 categories	6 categories
Puerto Rican	0.268	0.200	0.207	0.308	0.305	0.305
African American	−0.319	−0.378	−0.389	0.307	0.305	0.305
Socioeconomic status	0.008	0.008	0.008	0.005	0.005	0.005
Elder disability	0.306	0.313	0.318	0.033	0.033	0.033
Female caregivers	0.709	0.775	0.775	0.241	0.241	0.241
Employed caregivers	−0.333	−0.382	−0.372	0.223	0.222	0.222
Coresided caregivers	0.523	0.520	0.520	0.216	0.214	0.214

4 categories = 0, 0< to <3, 3⩽ to <15, ⩾15 hours; sample size=185, 71, 76, 60.
5 categories = 0, 0< to ⩽2, 2< to ⩽5, 5< to ⩽15, >15 hours; sample size = 185, 52, 48, 53, 54.
6 categories = 0, 0< to ⩽2, 2< to ⩽4, 4,< to ⩽8, 8< to ⩽15, >15 hours; sample size = 185, 52, 30, 36, 35, 54.

two additional proportional odds models using 5 and 6 groups in the personal care hours. Groupings were again chosen so that zero had its own group and other groups had roughly equal counts. The results of these models with 4, 5, and 6 groups respectively are compared in Table 3. Comparing the 4-group and 5-group analyses, the estimated regression coefficients are quite similar: five coefficients agree to the first decimal place, one to the second decimal place and one to the third place. The regression coefficient estimates are even more similar when 5 and 6 groups are compared. The estimated standard errors of the regression coefficients are almost identical in all three models. The P-values of the score test for the proportional odds assumption are 0.47 ($df = 21$) and 0.23 ($df = 28$), respectively, for the 5 and 6 groups models, supporting the proportional odds assumption in both cases. Overall, the results and conclusion from the proportional odds models using 4, 5, and 6 groups in the personal care hours in the SEP data were quite similar.

We have used one example to demonstrate the use of the proportional odds model and the combination of logistic regression and OLS linear models in analyzing outcomes with a substantial proportion of zero values. One conclusion we make from this demonstration is that the choice of approaches is data dependent. If the proportional odds assumption is valid, then it appears the method of choice; otherwise, the combination of a logistic regression and a linear model appears preferable. However one needs to make an appropriate choice of cut-off points when applying a proportional odds model to a continuous outcome. Although it is desirable to group the continuous variable, and broadly similar answers are likely to be obtained with different groupings, an optimal choice of the number of groups and the cut-off points is yet to be determined. Future study is needed for solving this problem.

Acknowledgments

This study was supported by the National Institute on Aging, Grant No. AG 11171. The authors are grateful for valuable comments from Dr. Peter Lachenbruch, Dr. Sharon Tennstedt, and Dr. Jennifer Anderson.

References

[1] Tobin J. Estimation of relationship for limited dependent variables. Econometrica 1958;26:24–36.
[2] Heckman J. Shadow prices, market wages, and labor supply. Econometrica 1974;42:679–94.
[3] Heckman J. The common structure of statistical models of truncation, sample selection, and limited dependent variables, and a sample estimator for such models. Annals Economic Social Measurement 1976; 5:475–592.
[4] Saei A, Ward J, McGilchrist CA. Threshold models in a methadone programme evaluation. Stat Med 1996;15:2253–60.
[5] Green PJ. Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives. J R Stat Soc B 1984;46:149–92.
[6] Harrell FE, Margolis PA, Grove S, Mason KE, Mulholland EK, Lehmann D, Muhe L, Gatchalian S, Eichenwald HF. Development of clinical prediction model for an ordinal outcome: the World Health Organization multicentre study of clinical signs and etiological agents of pneumonia, sepsis, and meningitis in young infants. Stat Med 1998;17:909–44.
[7] Peterson B, Harrell FE. Partial proportional odds models for ordinal response variables. Appl Stat 1990;39:205–17.
[8] Lachenbruch PA. Analysis of data with clumping at zero. Biometric J 1976;18:351–6.
[9] Lachenbruch PA. Utility of logistic regression in epidemiologic studies of the elderly. In: Wallace RB, Woolson RF, editors. The Epidemiologic Study of the Elderly. New York, Oxford: Oxford University Press, 1992. pp. 371–81.
[10] Tennstedt S, Chang B. The relative contribution of ethnicity vs. socioeconomic status in explaining differences in disability and recipient of care. J Geronotol Soc Sci 1998;53B:S61–70.
[11] McCullagh P, Nelder JA. Models for polytomous data. In: Generalized Linear Models, 2nd edition. Chapman and Hall, 1989. pp. 149–55.
[12] McCullagh P. Regression models for ordinal data. J R Stat Soc B 1980;42:109–42.
[13] SAS Institute Inc. SAS/STAT User's Guide, Version 6, 4th edition, vol. 2, Cary, NC: SAS Institute Inc., 1989.
[14] Berry D. Logarithmic transformations in ANOVA. Biometrics 1987; 43:439–56.