Methods for Improving Regression Analysis for Skewed Continuous or Counted Responses

Abdelmonem A. Afifi,[1] Jenny B. Kotlerman,[2] Susan L. Ettner,[1,3] and Marie Cowan[2]

[1]School of Public Health, [2]School of Nursing, [3]School of Medicine, University of California, Los Angeles, California 90095-1772; email: afifi@ucla.edu, jkotlerm@ucla.edu, settner@mednet.ucla.edu, mcowan@sonnet.ucla.edu

Annu. Rev. Public Health 2007. 28:95–111

The *Annual Review of Public Health* is online at http://publhealth.annualreviews.org

This article's doi: 10.1146/annurev.publhealth.28.082206.094100

0163-7525/07/0421-0095$20.00

First published online as a Review in Advance on November 17, 2006

Key Words

smear factor, multiple imputation, bootstrap, attrition weight, zero inflation, two-part models

Abstract

Standard inference procedures for regression analysis make assumptions that are rarely satisfied in practice. Adjustments must be made to insure the validity of statistical inference. These adjustments, known for many years, are used routinely by some health researchers but not by others. We review some of these methods and give an example of their use in a health services study for a continuous and a count outcome. For the continuous outcome, we describe retransformation using the smear factor, accounting for missing cases via multiple imputation and attrition weights and improving results with bootstrap methods. For the count outcome, we describe zero inflated Poisson and negative binomial models and the two-part model to account for overabundance of zero values. Recent advances in computing and software development have produced user-friendly computer programs that enable the data analyst to improve prediction and inference based on regression analysis.

INTRODUCTION

Over the past few decades, researchers have developed several methods to improve the interpretation and generalizability of regression analyses. Although these methods were developed mostly by statisticians, they were used mainly in the econometric literature and were not adopted as widely in the public health and other health sciences research. Here, we review some methods that can be useful when analyzing either continuous or count outcomes.

For continuous outcomes, the most commonly used statistical technique is multiple regression analysis. The least squares procedure gives minimum-variance unbiased estimators of the regression coefficients when the variance of the error term is assumed to be the same regardless of the values of the covariates (homoscedasticity). Furthermore, the usual testing and confidence interval procedures make the additional assumption that the errors are normally distributed. Standard textbooks give comprehensive expositions of the subject [Afifi et al. (1), Chatterjee et al. (11), Kutner et al. (30)]. Because these assumptions are rarely satisfied in practice, certain adjustments must be made to improve the validity of the results. We describe some of these adjustments, including retransforming outcome variables using the smear factor, accounting for missing cases using multiple imputation and attrition weights and improving statistical inference with bootstrap methods.

For count outcomes, common methods of analysis are Poisson and negative-binomial regressions. However, in many situations the probability of a zero occurrence exceeds that assumed by the model. We describe some methods that aim to overcome this problem, namely zero-inflated and two-part models.

Example

To illustrate the analyses described in this paper, we give an example of their use in the Multidisciplinary, Doctor, Nurse Practitioner (MDNP) study [Cowan et al. (12), Ettner et al. (20)]. The aim of the study was to ascertain the efficacy of a multidisciplinary team of hospitalists/attending physicians and advanced nurse practitioners on management of care of general medicine patients during hospitalization and for four months after discharge. In this study, the patients were hospitalized in a large academic medical center. The design was a prospective, two-group, quasi-experimental, pre-post test with repeated measures at 30 days and 4 months post hospital discharge. A total of 1207 patients were enrolled, including 626 in the control (C) unit and 581 in the experimental (E) unit. Of these patients, 885 had a complete account of the cost after follow-up: 460 in unit C and 425 in unit E. The primary analyses in the study were concerned with how health care costs and utilization for the experimental unit compare with those of the control unit. In this chapter, for the sake of illustration, we compare cost and utilization for two other groups: severely and nonseverely ill patients. A thorough review of methods for analyzing health care costs and utilization is provided by Diehr et al. (15).

CONTINUOUS OUTCOME

We use multiple linear regression analysis to test whether the long-term cost (COST) for patients who were severely ill was different from patients who were mildly or not severely ill. The variables considered in our analysis are listed in **Table 1**. The majority of the covariates are dichotomous, but for ease of interpretation, we transform the continuous variables Age and Pre-Admit Cost into binary variables and break preadmission severity of illness (SEVERITY) into two categories (severe and not severe). The variable UNIT represents whether a patient was in the C unit (Intervention = 0) or the E unit (Intervention = 1). It is important to adjust the actual interview time in the regression analysis because many patients were difficult to locate and were unavailable at the time that

Table 1 Variables used in cost analysis in the MDNP study

Variable	Description	*N*
Cost of health care for 4 months post discharge (COST)	A continuous variable used as the outcome variable	885
Cost of health care for 3 months before admit into the study (PRE-COST)	1 = cost was ≥$15,000; 0 = cost was <$15,000	1192
Gender (MALE)	1 = male; 0 = female	1207
Caucasian (WHITE)	1 = Caucasian; 0 = not Caucasian	1207
African American (BLACK)	1 = African American; 0 = not African American	1207
Asian (ASIAN)	1 = Asian; 0 = not Asian	1207
Other race (OTHER)	1 = other race; 0 = not other race	1207
Latino ethnicity (LATINO)	1 = Latino; 0 = not Latino	1207
Pre-admit severity of illness (SEVERITY)	1 = severe; 0 = not severe	1194
Intervention (UNIT)	1 = intervention group; 0 = control group	1207
Age (AGE)	1 = 65 or older; 0 = under 65	1207
Number of days from enrollment to completion of the follow-up (DAYS)	A continuous variable	909

their follow-up was due. The variable DAYS is defined as the number of days from discharge to the last follow-up and accounts for differences in the amount of services received. We thus adjust the variable COST by multiplying it by the variable DAYS and dividing by 120 (days in a four-month period). Because the cost reflects the number of days since discharge, inflating or deflating the value makes the results directly comparable.

Log Transformation

The means and standard deviations for COST among nonsevere patients are $12,579 ± $16,716 and among the severe patients are $30,784 ± $40,996. The box plots are exhibited in **Figure 1**. Examination of both box plots shows that the distribution of the outcome variable, COST, is clearly skewed with a long tail toward larger values. To produce a more symmetric distribution, we transform COST into the log scale, obtaining LOGCOST. Because the initial stay at the hospital is included in COST, producing all positive values, the problem of taking the log of zero has been avoided. If zeros were present, other transformations, such as the square root or log (COST + a constant) could be considered [Cantoni & Ronchetti (9), Manning & Mullahy (40)]. A common choice of the latter constant is the lowest nonzero value observed in the sample. Alternatively, the two-part regression model could be used as we illustrate in our discussion of count outcomes later in this paper [Buntin & Zaslavsky (7), Duan et al. (18), Diehr et al. (15), Lachenbruch (31, 32), Manning et al. (39), Mullahy (46), Seshamani & Gray (59), Xie et al. (64)]. The box plots of LOGCOST are shown in **Figure 2**, which indicates that taking the log makes the distribution closer to being symmetric, thus better approximating a normal distribution. The results of the regression of LOGCOST on the variables described in **Table 1** are shown in **Table 2**, and the normal probability plot for the standardized residuals of that regression is seen in **Figure 3**. The box plots of LOGCOST and the normal probability plot of the residuals indicate a reasonably good fit to the data. Some of the independent variables are not statistically significant and are possible candidates for elimination from the regression equation.

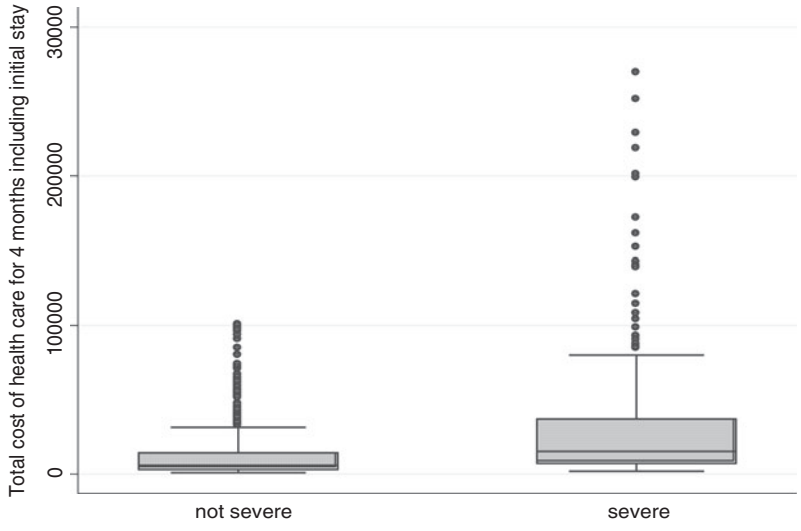

Figure 1

Box plots of COST divided into two severity groups.

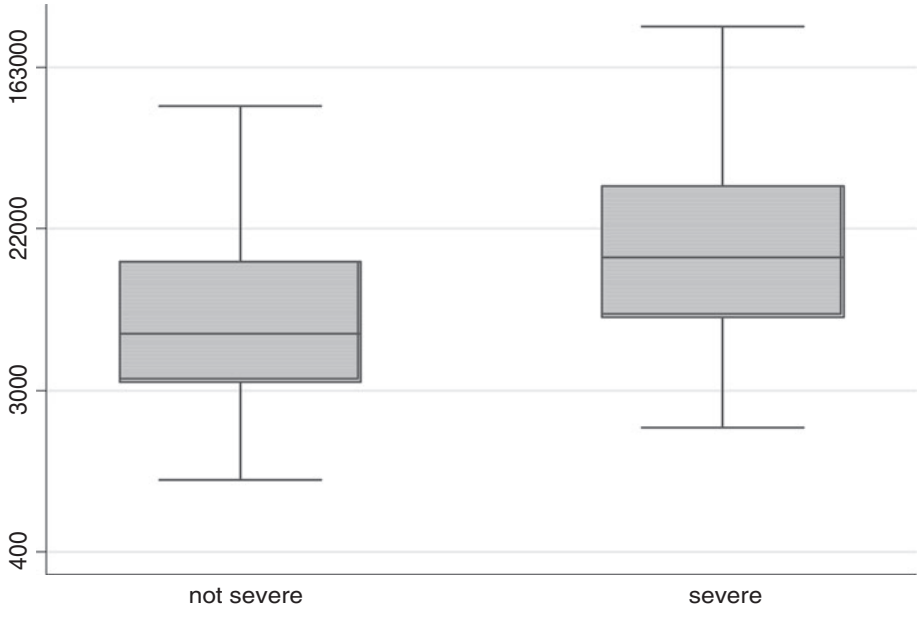

Figure 2

Box plots of LOGCOST for severe versus nonsevere patients.

Using Multiple Imputation to Adjust for Missing Values

A major problem facing most investigators is that of missing data. Almost invariably, a portion of the subjects provide incomplete data. One solution to the problem is to analyze only the complete observations. However, this method sacrifices information that could be incorporated into the analysis and may

Table 2 Regression of LOGCOST on baseline and demographic variables before and after applying multiple imputation

	Without multiple imputation			Using multiple imputation		
LOGCOST	Estimate	Standard error	P-value	Estimate	Standard error	P-value
SEVERITY	0.66	0.07	<0.0001	0.65	0.07	<0.0001
Intercept	9.07	0.30	<0.0001	8.76	0.07	<0.0001
UNIT	−0.16	0.07	0.01	−0.17	0.07	0.01
MALE	0.08	0.07	0.22	0.08	0.07	0.21
LATINO	−0.07	0.09	0.46	−0.06	0.09	0.51
BLACK	0.13	0.09	0.14	0.11	0.09	0.23
ASIAN	−0.22	0.14	0.12	−0.21	0.14	0.15
OTHER	−0.32	0.12	0.007	−0.33	0.12	0.007
PRE-COST	0.80	0.08	<0.0001	0.80	0.08	<0.0001
AGE	0.12	0.07	0.09	0.14	0.07	0.08
DAYS	−0.003	0.002	0.29	−0.002	0.002	0.31

introduce nontrivial bias. A large literature on the subject has appeared over the past 50 years, reviews of which can be found in Allison (4), Little & Rubin (36) and Schafer (57). The current wisdom in the statistical literature is that some form of an imputation method is preferable to the analyses of only complete cases. For example, Oostenbrink & Al (47) conclude, on the basis of a simulation study, that multiple imputation is the preferred method for the analysis of cost data under certain conditions. A number of publications provide comprehensive descriptions of methods for handling missing data, including multiple imputation. Some recommended references are by Kneipp & McIntosh (29), McCleary (41), Raghunathan (48), Rubin (54), Rubin & Schenker (55), and Schafer (58). Yang et al. (65) extend these methods for variable selection in regression analysis.

The multiple imputation process involves two steps:

- Simulate the posterior distribution of the regression coefficients; and
- use the simulated parameter values to estimate each missing value from its regression on the observed variables for the same subject, adding to this estimate a random error selected from the distribution of residuals obtained in the data set.

This process is repeated several times, resulting in a number of datasets, each with a different set of imputed values. Each dataset is analyzed separately, producing a set of estimates that are then combined into a final model. We used in our analysis the method of multiple imputation for the covariates (see **Table 2**).

In this example, the changes in the estimated regression coefficients and their

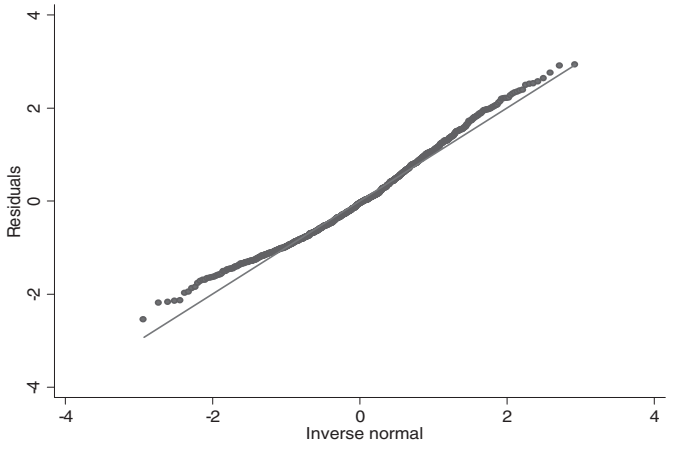

Figure 3

Probability plot of the standardized residuals for LOGCOST.

standard errors were not very large. For example, without adjusting for the missing covariates, the regression coefficient for the severity of illness is 0.66, whereas the combined estimate for the same variable after multiple imputation is 0.65. We also note that no variable changed from being statistically significant to being nonsignificant, or vice-versa. One procedure for gauging the effect of adjusting for the missing values is to take the ratios of the absolute values of the regression coefficients with and without multiple imputation. Another way to quantify the effect of missing information is called "fraction of missing information," which is a standard part of PROC MIANALYZE output in the SAS. It measures the fraction of information about each coefficient lost because of missing data in the total set [Allison (4)]. In our example, the fraction of missing information ranges from 0.0006 to 0.050; most are less than 0.01. These numbers confirm that the effect of missing values in this particular sample is negligible. However, when this fraction is large, e.g., greater than 50% for several variables, one may consider increasing the number of MI datasets to 10 or more [Allison (4)]. In that case, doing so can increase the efficiency of the estimation procedure from 91% to 95% for 5 versus 10 MI datasets, relative to an infinite number of MI datasets [Rubin (54)]. Although the resulting reduction in standard error is small, it may still be worth doing when the cost of computation is negligible.

Accounting for Loss to Follow-Up Using Attrition Weights

The method of multiple imputation applies best to observations obtained from subjects who remained in the sample, i.e., subjects for whom we have measurements on the outcome variables as well as some follow-up covariates. A different type of missing data occurs when the subject is lost to follow-up. In such cases we have only baseline data on those subjects, but not the follow-up outcome variables or other data. Such subjects are often eliminated from the sample altogether, thus potentially producing nonresponse or attrition bias.

Because more than 300 patients did not complete the follow-up interviews, we must correct for this potential attrition bias. For example, Lin et al. (34) discuss the use of survival analysis methods to estimate medical costs from incomplete follow-up data. A more common method to perform the adjustment is to use attrition weights [Foster & Bickman (22), Horowitz & Manski (25), Horvitz & Thompson (26), Little (35), McGuigan et al. (44), Robins & Rotnitzky (50), Wooldridge (63)]. The rationale behind this method is that subjects having certain characteristics may be more likely to respond than others. The adjustment procedure attempts to develop an equation that predicts the likelihood of whether the subject in question will actually complete the response. This probability can be viewed as the probability of nonattrition. It can be interpreted in a manner similar to the propensity score, which is the probability of being assigned to a treatment in a clinical trial conditional on the observed covariates. A large literature exists on the use of propensity scores to adjust for differential treatment assignments in clinical trials. Readers interested in that topic are referred to the following articles: D'Agostino (13), Rosenbaum (51), Rosenbaum & Rubin (52, 53), and Rubin & Thomas (56).

Computation of the attrition weights proceeds as follows. Using a dummy variable (1 = completed follow-up, 0 = did not complete follow-up) and modeling it with PROC LOGISTIC in SAS with an OUTPUT statement, or the LOGIT command in the STATA software with a PREDICT command afterward, we create a new variable in the dataset representing the predicted probability of completing the follow-up. The inverse of this predicted probability is the attrition weight. For example, if the subjects with a given profile of characteristics all respond, then their predicted probability of response will be 1 and their attrition weight will be 1/1 = 1. A subject with this profile will

therefore be representing only him- or herself. However, if the predicted probability is 1/2 for a subject of a given profile, then such a subject with a completed response will account for him- or herself and one missing person. Because the attrition weight is 1 divided by 1/2, that is 2, then in some sense that person will be counted twice. Similarly, for other probabilities of response, the inverse probability will determine how many times the person will be counted. This weight is used in conjunction with the log-transformed regression to reduce the bias. As a result, the confidence intervals in our example become larger because we are relying on some values more than others to account for the attrition bias. Doing so uses less of the available information and increases the variance, which is the price for reducing the bias. The weighted least squares process is familiar to many investigators when the weights are derived by other means. For example, if over- or undersampling is carried out at the data collection stage, then appropriate weights to adjust for this disproportionate sampling are incorporated into the least squares analysis [Kish (28)]. One alternative to using attrition weights is to use the Heckman model, which is based on obtaining an efficient regression estimator using maximum likelihood methods to estimate simultaneous equations [Heckman (24)]. Heckman's method, however, is highly dependent on the joint distributional assumptions, unlike single-equation estimation, which is more robust to such assumptions.

Retransforming the Outcome Variable—Using the Smear Factor

To answer the question, "What is the average cost for a patient?" we need to obtain the predicted values from each imputed regression equation. Because the predicted values are not automatically included in the regression output, we need to compute them with extra steps. For example, this step can be performed through the OUTPUT statement in PROC REG in SAS, or through the PREDICT command in STATA. Unfortunately the predicted value from the regression is on the log scale and should be retransformed into the original scale for ease of interpretation. When the normality assumption is made, the retransformed predicted value has the form $\hat{Z} = e^{\hat{Y}}$, where $\hat{Y}$ is the predicted value of LOGCOST. This value by itself is also not a correct retransformation because it is a biased estimate of the arithmetic mean. This bias follows from the fact that if a variable U is normally distributed with mean μ and variance σ^2, then e^U does not have a mean of e^{μ}. Rather, its mean is $e^{\mu+0.5\sigma^2}$. To adjust for this bias, we first compute a quantity called the smear factor, which is based on the distribution of the residuals [Duan (17)]. This factor is a number that is added to $\hat{Y}$ in the equation $\hat{Z} = e^{\hat{Y}}$ and makes the retransformed value less biased. A number of different smearing methods might be used for computation. The choice depends on the answers to the following two questions: First, are the residuals assumed to be normally distributed? And second, are the residuals homoscedastic, i.e., do the errors have the same variance for all subjects? The residual histogram (**Figure 4**) shows that the distribution of the residuals is not very well approximated by a normal distribution. However, the residual plots (**Figure 5**) indicate that the errors fall in an approximately uniform band over the range of the fitted values, except perhaps for predicted values below 8.5. The assumption of homoscedasticity is therefore plausible.

If it were reasonable to assume that the residuals are normal and homoscedastic, we can use the simplest form of retransformation, whereby $0.5\sigma^2$ is added to the predicted value before the retransformation. The quantity σ^2 would then be estimated as the mean squared error value from the regression of LOGCOST. In our case, the retransformed predicted value would be $\hat{Z} = e^{(\hat{Y}+0.5\hat{\sigma}^2)}$. For homoscedastic but nonnormal residuals, researchers recommend a more involved smear factor based on the empirical residual distribution [Ai & Norton (3), Duan (17), Manning

(38), Mullahy (46)]. Cantoni & Ronchetti (9), Manning & Mullahy (40), and Rascati et al. (49) expand on this concept and discuss other aspects of the analysis of transformed outcome variables. In practice, analysts generally use the empirical residual method because the cost of making the log-normal assumption can be very large when it is wrong.

The retransformation is also more complicated in the heteroscedastic case. In the case of a categorical covariate, different smear factors must be used for different groups. For example, if the two severity groups had different error variances, then two separate smear factors should be used, one for each group. These have to be calculated in the same way as in the homoscedastic approach.

We applied to our example both normal-homoscedastic and empirical residual methods and obtained very similar results. We defer presenting these results until we introduce the next two concepts.

Estimating the Difference in Cost Associated with Severity

In calculating the cost difference for patients with severe and nonsevere illnesses, we must estimate the regression-adjusted difference between the costs for these two groups of patients. When the outcome is not transformed, the estimated effect of the severity is simply the computed value of the regression coefficient of SEVERITY in the equation. But, because our outcome variable is obtained through a nonlinear transformation, we need to compute the expected effect of severity in a more elaborate way [see, for example, Barber & Thompson (6)].

We compute the predicted value of cost, $\hat{Z}_1(x_i)$ for a subject whose covariate profile is x_i and whose severity level = 1. Similarly, the predicted value for a nonsevere subject, with the same covariate profile x_i but a severity level = 0, is denoted by $\hat{Z}_0(x_i)$. The predicted effect of severity for an individual with covariate profile x_i is $\hat{Z}_1(x_i) - \hat{Z}_0(x_i)$, and the average effect of severity over the entire sample is $\frac{1}{N}\sum_{i=1}^{n}[\hat{Z}_1(x_i) - \hat{Z}_0(x_i)]$. We estimate these expected values through the following process:

1. First replace SEVERITY = 1 for all subjects, then compute $\hat{Y}_1$ for every subject in the sample, i.e., pretend that every person is from the severe group, where Y represents LOGCOST.
2. Replace SEVERITY = 0 for all subjects, then compute $\hat{Y}_0$ for every subject in the sample, i.e., pretend that every person is from the nonsevere group.
3. Retransform $\hat{Y}_1$ and $\hat{Y}_0$ by exponentiating them after adding a smear factor: $\hat{Z}_1 = e^{(\hat{Y}_1 + 0.5\hat{\sigma}^2)}$ and $\hat{Z}_0 = e^{(\hat{Y}_0 + 0.5\hat{\sigma}^2)}$.
4. Compute DIFF = $\hat{Z}_1 - \hat{Z}_0$.
5. Compute the average of DIFF over the whole sample.

The resulting value in step 5 is the estimated effect of severity.

Improving Statistical Tests with the Bootstrap

If the average value for the newly created variable DIFF is positive, one can conclude that patients with severe illness have higher costs on average. In our data, this average difference was $11,628. Although this difference is rather large, it does not have to be statistically significant. That is, checking for significance must account for the possibility that the error distribution is nonnormal. An efficient and effective way to adjust for nonnormality is bootstrapping [Barber & Thomson (5), Davison & Hinkley (14), Efron & Tibshirani (19) and Wehrens et al. (61)]. In this procedure, we repeatedly select individuals from the sample with replacement until we obtain a sample with the same size as the original sample. In this random drawing, some of the individuals may appear one or more times, while others not at all. The regression equation and the average difference that we seek are computed from the selected sample. This process is repeated many times (we used 1000), and a new set of estimates, the sampling distribution, is

created for the variable DIFF. We use this sampling distribution to compute a confidence interval for the true cost difference attributable to the severity. That is, we find a lower and an upper value such that, between them, they contain 95% of the means generated through the above process [Carpenter & Bithell (10)]. We note that the confidence interval can be obtained in several ways, and the choice of the number of replicates depends on which method is selected. The normal approximation method uses the percentiles of the normal distribution to compute the confidence limits based on the estimated mean and variance of the sampling distribution. The empirical, or percentile, method uses the upper and lower percentiles from the sampling distribution directly. However, both of these methods incur a certain amount of bias, which can be estimated from the sample. The so-called bias-corrected empirical method adjusts for this bias and is therefore usually the method of choice. In our case, we used the bias-corrected empirical method and chose 1000 replicates to obtain a precise confidence interval. If the normal approximation method is used, 500 replicates are usually sufficient [Efron & Tibshirani (19)]. The bootstrap procedure requires some programming, and the reader must consult the appropriate software manuals for packages such as SAS and STATA. In our example, we used STATA and obtained an average cost difference of $11,628 and a confidence interval of ($8696, $15,454). Because this 95% confidence interval does not include zero, we conclude that costs for the two severity levels are significantly different from each other at the 5% significance level.

COUNT OUTCOME

COST is computed from many components of health care utilization, such as hospital stays, outpatient visits, formal and informal home care, and so forth. For this part of the paper, we consider one of these components, the number of nights spent in a hospital after discharge from the initial stay (called NIGHTS). It is the sum of the nights spent in all hospital visits during the given time period. The mean and standard deviation for NIGHTS among nonsevere patients are 4.21 ± 11.70 and among severe patients are 8.66 ± 18.73. The unadjusted difference between the number of nights spent in the hospital for severe versus nonsevere patients is 4.45 nights. This is a count outcome and therefore does not satisfy the usual assumption of normality. Some analysts still use multiple linear regression, treating the distribution as approximately normal. A more rigorous approach is to treat the variable as discrete.

Poisson Regression

Several options are available for performing regression with count outcomes. One is to model it as having a Poisson distribution, that is, the probability that the variable has a value y is given by $e^{-\lambda}\lambda^{y}/y!$ for any integer value y greater than or equal to zero. This distribution is appropriate especially if the events being counted are rare, relative to the period of observation [Fisher & van Belle (21)]. The shape of the distribution varies greatly with the size of the mean. When the mean (λ) is large, the distribution is slightly skewed but is quite close to a normal distribution. When the mean is small, the most common count is zero. For example, if $\lambda = 0.5$, then the probabilities of obtaining a count of zero is 0.61, a count of one is 0.33, a count of two is 0.12, a count of three is 0.03, and counts of four or more have a very low probability. Hence, the distribution is highly skewed with the lowest value, zero, being the most apt to occur. Using a Poisson-distributed variable as the outcome leads to the concept of Poisson regression [Agresti (2), Hutchinson & Holtman (27), Long (37), and McCullagh (42)]. To perform the computations, we can use one of the several Poisson-regression programs available in software packages [Afifi et al. (1)]. Such programs rely on the concept of generalized

GLM: generalized linear models

ML: maximum likelihood

linear models (GLM) [Dobson (16), McCullagh & Nelder (43)].

The GLM has three components. The first is a random component consisting of the outcome variable Y, the number of events in a fixed time period. The second is the set of covariates. For a given set of covariates $X_1, \ldots, X_P$, we assume that the distribution of Y is Poisson with mean μ. The third component is a link function $g(\mu)$ that relates μ to the linear predictor, i.e., $g(\mu) = \alpha + \beta_1 X_1 + \cdots + \beta_P X_P$. When the outcome variable has a Poisson distribution, the link function is usually taken to be $\ln(\mu)$. Then a generalized linear model (or regression) program is used, specifying either the Poisson distribution or the link function. Most GLM programs use the maximum likelihood (ML) method to estimate the parameters of the model.

To perform a Poisson regression, one can use a program written specifically for that purpose (such as STATA poisson) or a GLM program (such as SAS GENMOD). If a GLM program is used, the analyst specifies simply that the distribution is Poisson and the link function is the log. The program then computes the ML estimates using iterative procedures and performs approximate tests and confidence intervals on the basis of the asymptotic distribution of the estimates. Some software, such as LogXact 5 from Cytel Software Corporation, allows the user to compute a Poisson regression and the corresponding inference using exact methods. This is useful when sample sizes are small because the approximate inference is recommended only for large samples.

Whichever program is used, the output typically includes the parameter estimates, say A, B_1, B_2, …, B_P and their standard errors, as well as possibly some tests of hypotheses or confidence intervals. For example, we can use an approximate normal test statistic $z = B_i/\mathrm{SE}(B_i)$ to test the hypothesis that $\beta_i = 0$. The 95% confidence interval for β_i is $B_i \pm 1.96\ \mathrm{SE}(B_i)$. The Wald and likelihood ratio tests commonly used for logistic regression are appropriate here as well. Also, similar to logistic regression, the exponentiated coefficients can be given useful interpretations. If X is dichotomous, taking on values 0 or 1, and β is its coefficient in the regression equation, then e^{β} is the ratio of μ_1 to μ_0, the means of Y corresponding to $X = 1$ and $X = 0$, respectively. If X is continuous, then e^{β} is the ratio of μ at $(X + 1)$ to μ at X. As in logistic regression, if other independent variables are in the equation, then these means are adjusted to them. In many Poisson regression situations, the period of observation is not the same for all individuals. In our example we are interested in the number of nights of hospitalization incurred by a patient during the four-month postintervention period. If we observe a patient for t months, then the number of nights, Y, spent in the hospital during that period has mean $\lambda = (\mu t)$, where μ is the mean number of hospital nights per month of observation. To apply Poisson regression to patients observed at different lengths of time t, we use the model

$$\ln(\lambda/t) = \alpha + \beta_1 X_1 + \cdots + \beta_p X_p,$$

or, equivalently,

$$\ln(\lambda) = \ln(t) + \alpha + \beta_1 X_1 + \cdots + \beta_p X_p.$$

The quantity $\ln(t)$ is called the offset in STATA and must be supplied as input to the program. In this case, the exponentiated coefficient is the ratio of the rates of occurrence per unit time. Thus, for Poisson regressions, exponentiated coefficients are rate ratios as contrasted with the odds ratios corresponding to logistic regression. For more information on Poisson regression and examples of its use see Cameron & Trivedi (8) or Greene (23). Winkelmann (62) provides a thorough theoretical background.

Zero Inflated Poisson and Negative Binomial Models

Because many of the patients did not spend any nights in the hospital, there are many zeros in the data (see **Figure 6**). A total of

548 patients (61%) had zero nights during the given period, whereas only 351 patients had spent one or more nights in the hospital. This proportion of zeros is higher than would be expected for a Poisson variable. One solution to this problem is to use zero inflated Poisson regression (called ZIP in STATA software). It is designed to apply to nonnegative count data with a large proportion of zeros [Lambert (33) and Mullahy (45)]. The idea of this model is that the variable follows a branching process: With probability P_0, the value of the variable is zero, and with probability (1-P_0), it acquires any integer value greater than or equal to zero according to the Poisson model. In the ZIP model, we first compute the probability p and then estimate the parameters of the regression of the Poisson part on the covariates. Logistic regression is used to obtain the probability of being either zero or Poisson distributed (the inflate portion of the output), whereas the GLM is used as described above to perform the Poisson regression.

The Poisson distribution forces the mean and the variance of the variable to be equal. For many of the count variables, the variance is much larger than the mean and a different model is necessary to account for this inequality. The most common is the negative binomial model, which essentially assumes that the variable is a mixture of several Poisson variables. That is, the subjects are assumed to have a common mean, plus a quantity that is specific to the subject, which is a random effect. When the entire subject population is evaluated together, the distribution of the outcome variable is negative binomial. The mathematical details of this model can be found in Agresti (2) or Cameron & Trivedi (8).

Several software programs exist for computing the regression of a negative binomial outcome variable, utilizing a GLM process similar to the one used for the Poisson regression. The proportion of zeros may also be more than expected under a negative binomial model, in which case a zero inflated negative binomial (ZINB in STATA software) can be used. The idea behind ZINB is quite similar to ZIP. In either analysis, we can perform a test, known as the Vuong test, of whether including zero inflation is necessary [Vuong (60)]. That is, the Vuong test compares Poisson to ZIP or negative binomial to ZINB. Furthermore, we can perform a likelihood ratio test of whether ZINB is more appropriate than ZIP for a given data set.

In our example, examination of the data showed that ZINB is more appropriate than ZIP and including zero inflation improves the analysis. We use the STATA program ZINB to obtain the regression of NIGHTS as a function of the covariates. In this analysis we have 883 complete observations. The results of running the ZINB program include two parts; one gives the negative binomial regression part of the output and the other gives the results of the logistic regression for excess zeroes, called "inflate" (**Table 3**). From **Table 3**, the coefficient for the variable SEVERITY is 0.45. To interpret this number, we exponentiate it and obtain a value of 1.57. That is, the rate of the number of nights of hospitalization for patients with severe illness is approximately 1.57 times the rate for the other patients, among those who can have nonzero nights. This difference is statistically significant ($z = 2.81$, $p = 0.005$). We can obtain a 95% confidence interval for the coefficient as $B_i \pm 1.96\ \mathrm{SE}(B_i) = 0.45 \pm 1.96 * 0.16 =$ (0.14, 0.76). These confidence limits, when exponentiated, result in a confidence interval of (1.15, 2.14) for the rate ratio.

The preadmission cost of health care (PRECOST) is highly statistically significant whereby patients with high preadmission cost have a rate of nights spent in the hospital ~1.7 times that of the other patients. The Vuong test showed that the NBREG model would be statistically significantly worse than the ZINB model, that is, the observed number of zeros is significantly higher than the number expected under the negative binomial model ($p < 0.0001$). In addition, the likelihood ratio test showed that ZINB fits significantly better than ZIP ($p < 0.0001$), confirming our own inspection of the data.

Table 3 Zero-inflated negative binomial regression of NIGHTS

	Negative binomial portion		Inflate portion	
Total nights	Coefficient (SE)	P-value	Coefficient (SE)	P-value
SEVERITY	0.45 (0.16)	0.005	−0.19 (0.21)	0.24
UNIT	−0.11 (0.15)	0.50	0.23 (0.19)	0.24
BLACK	−0.05 (0.18)	0.80	−0.85 (0.29)	0.003
ASIAN	−0.60 (0.43)	0.16	0.57 (0.44)	0.20
OTHER	−1.08 (0.29)	<0.0001	−0.07 (0.41)	0.86
PRE-COST	0.53 (0.16)	0.001	−1.65 (0.31)	<0.0001
AGE	0.27 (0.17)	0.12	−0.29 (0.21)	0.16
MALE	−0.11 (0.15)	0.49	−0.19 (0.19)	0.34
LATINO	−0.27 (0.20)	0.19	0.21 (0.27)	0.45
CONSTANT	−2.66 (0.18)	<0.0001	0.63 (0.22)	0.003

The STATA ZINB program produces additional output under the title "inflate" (see **Table 3**). This part of the output describes the logistic regression analysis for predicting whether the outcome variable must be equal to zero or whether it follows the negative binomial distribution. This output is of interest in and of itself because it differentiates between two types of patients: those who had zero nights, without the possibility of a positive number of nights, and those who followed the negative binomial distribution of the number of nights, including the possibility of zero nights. Specifically, it identifies the factors associated with whether a patient had only the possibility of zero nights or followed the negative binomial distribution of nights. For example, the coefficient for SEVERITY is –0.19, meaning that those with severe illness have a lower likelihood of having only zero hospital nights. In other words, more severely ill patients are more likely to follow the negative binomial distribution, and hence are more likely to have additional hospitalization than are those with lower severity. Because this is a logistic regression, we can interpret the exponentiated coefficient as an odds ratio. Specifically, the odds of a more severe patient having only zero hospital nights are $e^{-0.19} = 0.83$. The inverse of this quantity is $1/0.83 = 1.21$, indicating that the odds of a more severely ill patient following the negative binomial distribution are 1.21 times those of a less severely ill patient. However, we note from the output that this association is not statistically significant.

Estimating the Number of Hospital Nights

The output of ZINB can be used to estimate the predicted number of nights of hospitalization for any patient with a given profile of covariates. Because patients were followed for a different period of time (DAYS), we adjusted these estimates to correspond to exactly four months for each patient, using the offset option in STATA. As in the case of continuous outcome regression analysis, we can compute an average estimated difference for patients with severe and nonsevere illness, as well as a confidence interval for that difference. For a given patient, the unconditional estimated number of nights is $[1\text{-}P_0]$*(the estimated number of nights from the negative binomial regression), where P_0 is the probability of excess zeros as computed from the logistic regression. Furthermore, we can also use the bootstrap methodology as we have demonstrated in the continuous outcomes example. We applied this bootstrapping method to our example and concluded that, over the four-month period, patients with severe illness spent an average of 3.04 nights more than did patients with a less severe illness. The 95% confidence interval for the average difference

Table 4 Two-part and ZINB models for NIGHTS

	Two-part model		Zero inflated negative binomial model	
	MEAN (SE)	CONFIDENCE INTERVAL	MEAN (SE)	CONFIDENCE INTERVAL
Relative risk[a]	1.28 (0.15)	(1.02, 1.62)	1.73 (0.39)	(1.25, 2.43)
Difference[b]	0.76 (0.31)	(0.26, 1.49)	3.16 (1.06)	(1.36, 5.51)

[a]the probability of nonzero NIGHTS (two-part model), or the probability of following the negative binomial distribution (ZINB model), for severely ill patients relative to nonseverely ill patients.
[b]the average unconditional estimate of NIGHTS for severely ill patients minus that for nonseverely ill patients.

is (1.2, 5.45), confirming that the difference in the length of stay is significantly higher for patients with severe illness. These results incorporate both parts of the ZINB model, that is, the negative binomial and the excess zero parts.

Two-Part Model

Another method that may be used for this type of data is the two-part model (TPM). The TPM first uses logistic regression to predict the probability of whether the patient had any nights in the hospital (this is a multiple logistic regression with all the independent variables used in the regression to control for variations in the data). Then it calculates the conditional expectation of the number of nights for the subsample of only those patients who had at least one night's stay. Next, the unconditional expectation is calculated by multiplying the predicted probability of having any hospital nights by the conditional expectation just described. In the second part of the TPM, that is the regression of nonzero values, a number of different models can be used including either count or continuous methods. Multiple regression analysis is used frequently, in which case it is common practice to transform the outcome variable into a logarithmic form. The TPM is very useful for data that have a large number of zeros, as well as a few very large values. From the graph of the total number of nights spent in the hospital post–initial admission (see **Figure 6**), it is clear that our data fit this profile. Implementing the two-part model requires performing a series of steps detailed in Appendix A.

The numbers produced in this process can be used to estimate the average difference in the expected number of nights, as well as a corresponding confidence interval. Analysts recommend applying the bootstrap procedure to obtain these results. Applying this whole process to our data produces the results shown in **Table 4**. Thus we estimate the average difference to be 0.76 nights for severe versus nonsevere patients, with a 95% confidence interval of (0.26, 1.49) nights. Although these results are different from those produced by the ZINB method, both methods agree that there is a significant difference in the number of nights for severe versus nonsevere patients, at the 5% level.

TPM: two-part model

We also applied the bootstrap procedure to the relative risk estimates obtained in step 5. The computations resulted in an average relative risk for all patients of 1.28, that is, a patient with severe illness is 28% more likely to spend one or more nights in the hospital than is a patient with nonsevere illness. The confidence interval for this relative risk is (1.02, 11.62), showing a barely significant result at the 5% level. The results for the two-part model are also listed in **Table 4**. For the sake of comparison we performed a similar calculation of the risk ratios produced by the ZINB model. The average relative risk for all patients was 1.73, and the corresponding confidence interval was (1.25, 2.43) (see **Table 4**). Again these results differ from the ones just reported for the TPM, although, as we noted earlier, they both present a statistically significant higher risk of spending some nights in the hospital for those with severe illness. In practice, the choice between these two models

depends on which one better fits the data. Examination of residuals and other model checks is recommended as an aid in making this choice.

SUMMARY

We present several ideas frequently used in regression analyses. Some of these, such as multiple imputation and Poisson regression, are well known in public health and biomedical statistical applications. Other techniques, such as bootstrapping, TPM, and zero-inflated models, are also gaining popularity. We attempt to present an intuitive explanation of these techniques. We encourage our colleagues in all health fields to consider these ideas whenever they analyze either continuous or count types of data. These methods provide viable extensions of standard generalized linear models and help adjust for violations of the underlying model assumptions.

SUMMARY POINTS

1. Several adjustments to regression analysis can improve the generalizability and validity of the results.
2. Multiple imputation is a practical and useful way to adjust for missing values.
3. Attrition weights based on propensity scores help adjust for the potential bias resulting from differential loss to follow-up.
4. When transforming the regression outcome, using the smear factor enhances the validity of retransformed values.
5. Bootstrapping produces confidence intervals that are more robust to violations of regression assumptions than are standard methods.
6. Zero inflated models for Poisson and negative binomial regression are practical options to account for overabundance of zero outcomes.
7. Two-part models present another option to account for zeros among either continuous or count outcomes.

APPENDIX A: STEPS FOR IMPLEMENTING THE TWO-PART REGRESSION MODEL TO ESTIMATE THE NUMBER OF HOSPITAL NIGHTS

1. Create a dummy variable with 0 = zero hospital nights and 1 = one or more hospital nights.
2. Use multiple logistic regression on the dummy variable created in step 1. If desired, use attrition weights in this step.
3. Reset the value of severity to be 1 (severe) for every subject and, from the model in step 2, compute the predicted probability that the subject has one or more night's stay.
4. Reset the value of severity to be 0 (nonsevere) for every subject and, from the model in step 2, compute the predicted probability that the subject has one or more night's stay.
5. Divide the probability from step 3 by that from step 4, obtaining the relative risk of having severe illness. This quantity estimates the patient's probability of spending any nights in the hospital if the patient were severely ill, divided by the same patient's probability of spending nights in the hospital if the condition were not severe.
6. Run an attrition-weighted multiple linear regression with the log of the number of nights as the dependent variable, restricting the data to those patients who had at least one

hospital night's stay (remember, you can take the log because now all the patients in the subsample have 1 or more as their count).

The above steps produce one number for each subject. They afford us the opportunity to compute the expected number of nights for a patient if the patient were either high or low severity. The steps involved in this part of the computation are as follows:

7. Force the value of severity to be 1 (severe) for every subject and, from the model in step 6, compute the predicted log of the number of nights.
8. Force the value of severity to be 0 (nonsevere) for every subject and, from the model in step 6, compute the predicted log of the number of nights.
9. Retransform the estimates from steps 7 and 8 from the logarithmic to the original form, adding the smear factor.
10. Multiply the predicted probability from step 3 by the conditional expected number of nights from step 7, after adding the smear factor as in step 9. The result is the unconditional expected number of nights, assuming that each patient has severe illness.
11. Multiply the predicted probability from step 4 by the conditional expected number of nights from step 8, after adding the smear factor as in step 9. The result is the unconditional expected number of nights, assuming that each patient has nonsevere illness.
12. Compute the unconditional difference in the expected number of nights for each patient by subtracting the quantity computed in step 11 from that computed in step 10.

LITERATURE CITED

1. Afifi AA, Clark VA, May SJ. 2004. *Computer-Aided Multivariate analysis*. New York: Chapman & Hall. 4th ed.
2. Agresti A. 2002. *Categorical Data Analysis*. New York: Wiley. 2nd ed.
3. Ai C, Norton E. 2000. Standard errors for the retransformation problem with heteroskedasticity. *J. Health Econ.* 19:697–718
4. Allison PD. 2001. *Missing Data*. Thousand Oaks, CA: Sage
5. Barber JA, Thomson SG. 2000. Analysis of cost data in randomized trials: an application of the non-parametric bootstrap. *Stat. Med.* 19:3219–36
6. Barber J, Thompson S. 2004. Multiple regression of cost data: use of generalised linear models. *J. Health Serv. Res. Policy* 9:197–204
7. Buntin MB, Zaslavsky AM. 2004. Too much ado about two-part models and transformation? Comparing methods of modeling Medicare expenditures. *J. Health Econ.* 23:525–42
8. Cameron AC, Trivedi PK. 1998. *Regression Analysis of Count Data*. Cambridge, UK: Cambridge Univ. Press
9. Cantoni E, Ronchetti E. 2006. A robust approach for skewed and heavy-tailed outcomes in the analysis of health care expenditures. *J. Health Econ.* 25:198–213
10. Carpenter J, Bithell J. 2000. Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians. *Stat. Med.* 19:1141–64
11. Chatterjee S, Hadi AS, Price B. 1999. *Regression Analysis by Example*. New York: Wiley. 3rd ed.
12. Cowan MJ, Shapiro MF, Hays RD, Afifi AA, Vazirani S, Ettner S. 2006. The effect of a multidisciplinary, hospitalist/physician and advanced practice nurse collaboration on hospital costs. *J. Nursing Adm.* 36:79–85
13. D'Agostino RB. 1998. Tutorial in biostatistics. Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Stat. Med.* 17:2265–81

14. Davison AC, Hinkley DV. 1997. *Bootstrap Methods and their Application*. Cambridge, UK: Cambridge Univ. Press
15. Diehr P, Yanez D, Ash A, Hornbrook M, Lin D. 1999. Methods for analyzing health care utilization and costs. *Annu. Rev. Public Health* 20:125–44
16. Dobson AJ. 2002. *An Introduction to Generalized Linear Models*. New York: Chapman & Hall
17. Duan N. 1983. Smearing estimate: a non-parametric retransformation method. *J. Am. Stat. Assoc.* 78:605–10
18. Duan N, Manning WG, Morris CN, Newhouse JP. 1983. A comparison of alternative models for the demand for medical care. *J. Bus. Econ. Stat.* 1:115–26
19. Efron B, Tibshirani RJ. 1993. *An Introduction to the Bootstrap*. London: Chapman & Hall
20. Ettner SL, Kotlerman J, Afifi AA, Vazirani S, Hays RD, et al. 2006. A multi-disciplinary doctor-nurse practitioner model: an alternative approach to reducing the costs of patient care? *Med. Dec. Making* 26:9–17
21. Fisher LD, van Belle G. 1993. *Biostatistics: A Methodology for the Health Sciences*. New York: Wiley
22. Foster EM, Bickman L. 1996. An evaluator's guide to detecting attrition problems. *Eval. Rev.* 20:695–723
23. Greene W. 2002. *Econometric Analysis*. New York: Prentice Hall. 5th ed.
24. Heckman J. 1976. The common structure of statistical models of truncation, sample selection, and limited dependent variables and a simple estimator for such models. *Ann. Econ. Soc. Measur.* 5:475–92
25. Horowitz JL, Manski CF. 1998. Censoring of outcomes and regressors due to survey nonresponse: identification and estimation using weights and imputations? *J. Econometr.* 84:37–58
26. Horvitz DG, Thompson DJ. 1952. A generalization of sampling without replacement from a finite universe. *J. Am. Stat. Assoc.* 47:663–85
27. Hutchinson MK, Holtman MC. 2005. Analysis of count data using poisson regression. *Res. Nurs. Health*. 28:408–18
28. Kish L. 1965. *Survey Sampling*. New York: Wiley
29. Kneipp SM, McIntosh M. 2001. Handling missing data in nursing research with multiple imputation. *Nurs. Res.* 50:384–89
30. Kutner MH, Nachtsheim CJ, Neter J. 2003. *Applied Regression Models*. New York: McGraw-Hill
31. Lachenbruch PA. 2001. Comparisons of two-part models with competitors. *Stat. Med.* 20:1215–34
32. Lachenbruch PA. 2002. Analysis of data with excess zeros. *Stat. Methods Med. Res.* 11:297–302
33. Lambert D. 1992. Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics* 34:1–14
34. Lin DY, Feuer EJ, Etzioni R,Way Y. 1997. Estimating medical costs from incomplete follow-up data. *Biometrics* 53:419–34
35. Little RJ. 2004. To model or not to model? Competing modes of inference for finite population sampling. *J. Am. Stat. Assoc.* 99:546–56
36. Little R, Rubin D. 2002. *Statistical Analysis with Missing Data*. New York: Wiley. 2nd ed.
37. Long JS. 1997. *Regression Models for Categorical and Limited Dependent Variables*. Thousand Oaks, CA: Sage
38. Manning W. 1998. The logged dependent variable, heteroskedasticity, and the retransformation problem. *J. Health Econ.* 17:283–95

39. Manning WG, Basu A, Mullahy J. 2005. Generalized modeling approaches to risk adjustment of skewed outcomes data. *J. Health Econ*. 24:465–88
40. Manning WG, Mullahy J. 2001. Estimating log models: to transform or not to transform? *J. Health Econ*. 20:461–94
41. McCleary L. 2002. Using multiple imputation for analysis of incomplete data in clinical research. *Nurs. Res*. 51:339–43
42. McCullagh P. 1980. Regression models for ordinal data. *J. R. Stat. Soc. B*. 42:109–42
43. McCullagh P, Nelder JA. 1989. *Generalized Linear Models*. New York: Chapman & Hall. 2nd ed.
44. McGuigan KA, Ellickson PL, Hays RD, Bell RM. 1997. Adjusting for attrition in school based samples. *Eval. Rev*. 21:554–67
45. Mullahy J. 1986. Specification and testing of some modified count data models. *J. Econometrics* 57:307–34
46. Mullahy J. 1998. Much ado about two: reconsidering retransformation and the two-part model in health econometrics. *J. Health Econ*. 17:247–81
47. Oostenbrink JB, Al MJ. 2005. The analysis of incomplete cost data due to dropout. *Health Econ*. 14:763–76
48. Raghunathan TE. 2004. What do we do with missing data? Some opinions for analysis of incomplete data. *Annu. Rev. Public Health* 25:99–117
49. Rascati KL, Smith MJ, Neilands T. 2001. Dealing with skewed data: an example using asthma-related costs of Medicaid clients. *Clin. Therapeutics* 23:481–98
50. Robins JM, Rotnitzky A. 1995. Semiparametric efficiency in multivariate regression models with missing data. *J. Am. Stat. Assoc*. 90:122–29
51. Rosenbaum PR. 2002. *Observational Studies*. New York: Springer-Verlag. 2nd ed.
52. Rosenbaum PR, Rubin DB. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70:41–55
53. Rosenbaum PR, Rubin DB. 1984. Reducing bias in observational studies using subclassification on the propensity score. *J. Am. Stat. Assoc*. 79:516–24
54. Rubin DB. 1987. *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley
55. Rubin DB, Schenker N. 1991. Multiple imputation in health-care databases: an overview and some applications. *Stat. Med*. 10:585–98
56. Rubin DB, Thomas N. 1996. Matching using estimated propensity scores: relating theory to practice. *Biometrics* 52:249–64
57. Schafer JL. 1997. *Analysis of Incomplete Multivariate Data*. New York: Chapman & Hall
58. Schafer JL. 1998. Multiple imputation: a primer. *Stat. Meth. Med. Res*. 8:3–15
59. Seshamani M, Gray AM. 2004. A longitudinal study of the effects of age and time to death on hospital costs. *J. Health Econ*. 23:217–35
60. Vuong Q. 1989. Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica* 57:307–33
61. Wehrens R, Putter H, Buydens LMC. 2000. The bootstrap: a tutorial. Chemometrics intelligen? *Lab. Sys*. 54:35–52
62. Winkelmann R. 2000. *Econometric Analysis of Count Data*. New York: Springer. 3rd. ed.
63. Wooldridge JM. 2002. Inverse probability weighted M-estimators for sample selection, attrition, and stratification. *Portoguese Econ. J*. 1:117–39
64. Xie H, McHugo G, Sengupta A, Clark R, Drake R. 2004. A method for analyzing longitudinal outcomes with many zeros. *Ment. Health Serv. Res*. 6:239–46
65. Yang X, Belin TR, Boscardin WJ. 2005. Imputation and variable selection in linear regression models with missing covariates. *Biometrics* 61:498–506

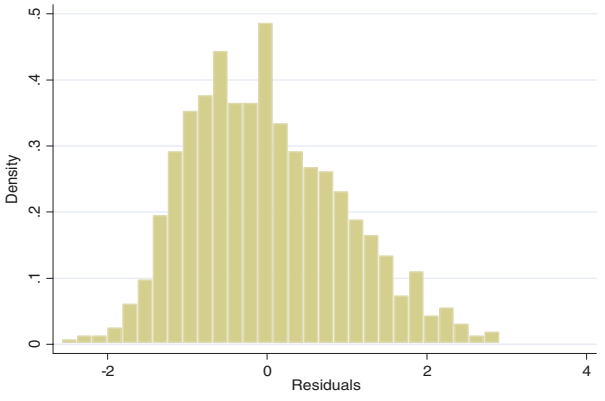

Figure 4

Histogram of residuals for LOGCOST after multiple imputation.

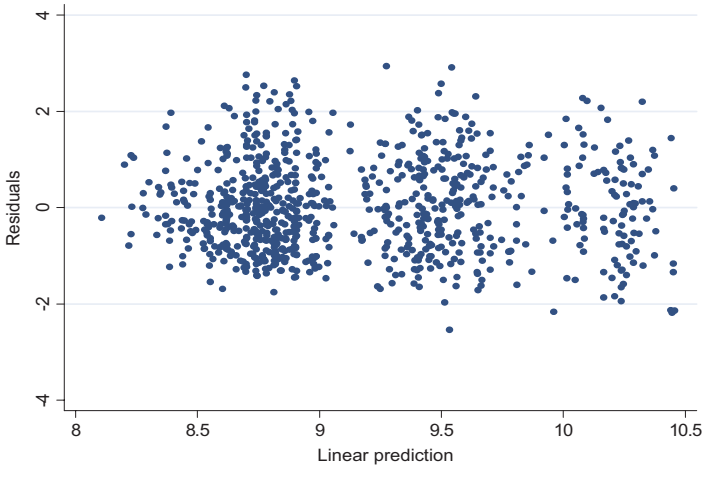

Figure 5

Scatter plot of the residuals for the LOGCOST after multiple imputation.

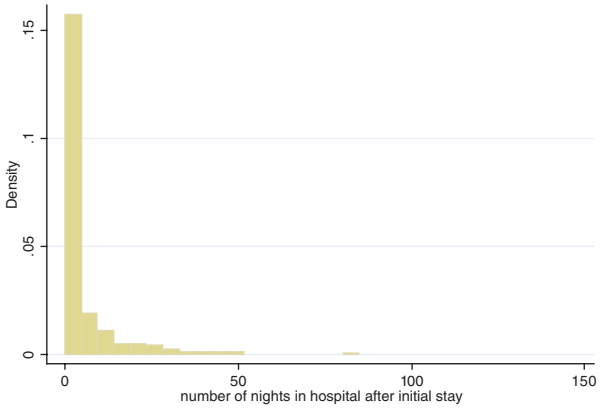

Figure 6

Number of hospital nights after the initial stay.

Annual Review of
Public Health

Volume 28, 2007

Contents

Symposium: Public Health Preparedness

Introduction: Preparedness as Part of Public Health
Nicole Lurie xiii

Assessing Public Health Emergency Preparedness: Concepts, Tools, and Challenges
Christopher Nelson, Nicole Lurie, and Jeffrey Wasserman 1

Quality Improvement in Public Health Emergency Preparedness
Michael Seid, Debra Lotstein, Valerie L. Williams, Christopher Nelson, Kristin J. Leuschner, Allison Diamant, Stefanie Stern, Jeffrey Wasserman, and Nicole Lurie 19

Risk Communication for Public Health Emergencies
Deborah C. Glik 33

First Responders: Mental Health Consequences of Natural and Human-Made Disasters for Public Health and Public Safety Workers
David M. Benedek, Carol Fullerton, and Robert J. Ursano 55

Epidemiology and Biostatistics

Network Analysis in Public Health: History, Methods, and Applications
Douglas A. Luke and Jenine K. Harris 69

Methods for Improving Regression Analysis for Skewed Continuous or Counted Responses
Abdelmonem A. Afifi, Jenny B. Kotlerman, Susan L. Ettner, and Marie Cowan 95

New Challenges for Telephone Survey Research in the Twenty-First Century
Angela M. Kempf and Patrick L. Remington 113

Seasonality of Infectious Diseases
David N. Fisman 127

Health Impact Assessment: A Tool to Help Policy Makers Understand Health Beyond Health Care
Brian L. Cole and Jonathan E. Fielding 393

Social Environment and Behavior

Physical Activity and Weight Management Across the Lifespan
Jennifer H. Goldberg and Abby C. King 145

The Hitchhiker's Guide to Tobacco Control: A Global Assessment of Harms, Remedies, and Controversies
Ronald M. Davis, Melanie Wakefield, Amanda Amos, and Prakash C. Gupta 171

Youth Violence Prevention Comes of Age: Research, Training, and Future Directions
Kara Williams, Lourdes Rivera, Robert Neighbours, and Vivian Reznik 195

Church-Based Health Promotion Interventions: Evidence and Lessons Learned
Marci Kramish Campbell, Marlyn Allicock Hudson, Ken Resnicow, Natasha Blakeney, Amy Paxton, and Monica Baskin 213

Risk Communication for Public Health Emergencies
Deborah C. Glik 33

Environmental and Occupational Health

The Epidemiology of Autism Spectrum Disorders
Craig J. Newschaffer, Lisa A. Croen, Julie Daniels, Ellen Giarelli, Judith K. Grether, Susan E. Levy, David S. Mandell, Lisa A. Miller, Jennifer Pinto-Martin, Judy Reaven, Ann M. Reynolds, Catherine E. Rice, Diana Schendel, and Gayle C. Windham 235

Beryllium: A Modern Industrial Hazard
Kathleen Kreiss, Gregory A. Day, and Christine R. Schuler 259

Adverse Late Effects of Childhood Cancer and Its Treatment on Health and Performance
Kirsten K. Ness and James G. Gurney 279

First Responders: Mental Health Consequences of Natural and Human-Made Disasters for Public Health and Public Safety Workers
David M. Benedek, Carol Fullerton, and Robert J. Ursano 55

Health Services

Managed Behavioral Health Care Carve-Outs: Past Performance and Future Prospects
Richard G. Frank and Rachel L. Garfield 303

Rationale and Public Health Implications of Changing CHD Risk Factor Definitions
Robert M. Kaplan and Michael Ong 321

Delivery of Health Services to Migrant and Seasonal Farmworkers
Thomas A. Arcury and Sara A. Quandt 345

Public Health Practice

Lessons from Cost-Effectiveness Research for United States Public Health Policy
Scott D. Grosse, Steven M. Teutsch, and Anne C. Haddix 365

Health Impact Assessment: A Tool to Help Policy Makers Understand Health Beyond Health Care
Brian L. Cole and Jonathan E. Fielding 393

How Can We Increase Translation of Research into Practice? Types of Evidence Needed
Russell E. Glasgow and Karen M. Emmons 413

Community Factors in the Development of Antibiotic Resistance
Elaine Larson 435

Assessing Public Health Emergency Preparedness: Concepts, Tools, and Challenges
Christopher Nelson, Nicole Lurie, and Jeffrey Wasserman 1

Quality Improvement in Public Health Emergency Preparedness
Michael Seid, Debra Lotstein, Valerie L. Williams, Christopher Nelson, Kristin J. Leuschner, Allison Diamant, Stefanie Stern, Jeffrey Wasserman, and Nicole Lurie 19

Indexes

Cumulative Index of Contributing Authors, Volumes 19–28 449

Cumulative Index of Chapter Titles, Volumes 19–28 454

Errata

An online log of corrections to *Annual Review of Public Health* chapters (if any, 1997 to the present) may be found at http://publhealth.annualreviews.org/