Available online at www.sciencedirect.com

ELSEVIER

Journal of Health Economics 23 (2004) 525–542

JOURNAL OF HEALTH ECONOMICS

www.elsevier.com/locate/econbase

Too much ado about two-part models and transformation? Comparing methods of modeling Medicare expenditures

Melinda Beeuwkes Buntin[a,*], Alan M. Zaslavsky[b]

[a] RAND Health, Washington Office, 1200 South Hayes Street, Arlington, VA 22202-5050, USA
[b] Department of Health Care Policy, Harvard Medical School, USA

Received 1 April 2003; accepted 1 October 2003

Abstract

Many methods for modeling skewed health care cost and use data have been suggested in the literature. This paper compares the performance of eight alternative estimators, including OLS and GLM estimators and one- and two-part models, in predicting Medicare costs. It finds that four of the alternatives produce very similar results in practice. It then suggests an efficient method for researchers to use when selecting estimators of health care costs.
© 2004 Elsevier B.V. All rights reserved.

JEL classification: 110; Cl; C5

Keywords: Health care cost modeling

1. Introduction

The health economist or health services researcher modeling health care costs or use faces a daunting literature about alternative estimators. The econometric challenges posed by these data include restricted range (non-negative observations), a “spike” of zero values, and skewness (a heavy right-hand tail). These properties make ordinary least squares estimation biased and inefficient. Popular alternatives to OLS include two-part models (Manning et al., 1981; Duan et al., 1983; Duan et al., 1984), which model the probability of nonzero costs separately from their level conditional on nonzero costs. The dependent variable is commonly log-transformed before OLS estimation to accommodate skewness. Predictions

* Correspondng author. Tel.: +1-703-413-1100x5581; fax: +1-703-413-8111.
E-mail address: buntin@rand.org (M.B. Buntin).

0167-6296/$ – see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.jhealeco.2003.10.005

from these models must be retransformed to obtain estimates on the original scale, and these retransformations can be sensitive to model misspecification. In particular, heteroscedasticity can bias estimates drawn from the frequently used two-part logged dependent variable models even when smeared retransformation factors are used (Duan, 1983; Manning, 1998). More recently, generalized linear models (GLMs) have been proposed to facilitate inferences about predictors of expected costs (Mullahy, 1998). Manning and Mullahy (2001) compared the performance one-part transformed models and GLMs using simulated data representing various violations of model assumptions, and suggested some model selection criteria.

With these new approaches, more than a half dozen alternative estimators are available to the conscientious researcher. Depending on the characteristics of the data and the research questions, each of them could be the “best” estimator under certain circumstances.

In this paper, we model Medicare payments for elderly Medicare beneficiaries, comparing eight alternative estimators. Our results suggest that researchers modeling health care costs and use might first fit one-part GLMs, proceeding to two-part GLMs or OLS models with transformed dependent variables if warranted.

2. Objectives

Our objective was to model expected health care costs incurred by Medicare beneficiaries given their demographic characteristics and responses to survey questions about their health status and conditions. Other objectives might have differently shaped our choices. If interpretation of regression coefficients were of interest, then the interpretability of the scale on which they are modeled would assume greater importance. If it were important to model the full predictive distribution of the costs, then GLM quasilikelihood models would be inadequate without a more completely specified probability model. Likewise, if the probability of nonzero cost or use were of interest, the two-part models would be particularly appropriate because they explicitly model that probability.

3. Data and model

Data were drawn from the 1996 Medicare Current Beneficiary Survey (MCBS), which combines survey responses and administrative data for a random sample of Medicare beneficiaries (Adler, 1994). We restricted our analysis to aged, non-institutionalized, non-HMO Medicare beneficiaries who were eligible for both Medicare hospital insurance (Part A) and outpatient insurance (Part B), both of which were included in our definition of Medicare reimbursements. HMO members and those ineligible for Part A or Part B were excluded because complete reimbursement data were not available for them. The non-aged and the institutionalized were not examined because their cost distributions differ from those of the community-dwelling elderly population. The final sample drawn from the MCBS had 10,134 individuals.

The purpose of the analysis described here was to predict costs so they could be compared for Medicare + Choice risk plans operating in overlapping areas, using data on health status and conditions of their members, and thereby to quantify risk selection (Zaslavsky and Beeuwkes Buntin, 2002). For this reason, we modeled the ratio of total Medicare costs to the average annual per-capita cost (AAPCC) by county, age and sex, which is the basis for reimbursement to the plans, rather than the dollar amount of costs. This removed much of the variation due to geographic location, allowing us to focus on relative costs affected by individual characteristics. However, in a loglinear model formulation this is empirically identical to modeling total expenditures with AAPCC as a predictor to represent regional variation: in that model, the estimated coefficient of AAPCC is almost exactly 1 and the coefficients of individual characteristics are almost the same as in the ratio model.

Some beneficiaries (8.6%) had no Medicare-reimbursed health care costs, and about half had expenses under US$ 1000. The cost distribution had a long right tail: a dozen sample beneficiaries had costs over US$ 100,000. Mean annual costs per beneficiary were approximately US$ 4000. The coefficient of variation in this sample of Medicare beneficiaries is similar to that reported for seniors in other surveys, such as the MEPS sample from the same year. It is lower than that of younger populations who have lower mean costs; however, the method for choosing a cost model that we describe below should work well for most medical cost distributions.

MCBS questions that were expected to be predictive of costs were chosen as covariates. Covariates were also limited to those that could be matched to the questions on the Consumer Assessment of Health Plans Study (CAHPS®) survey since the ultimate use of the model was predicting costs of CAHPS respondents (Zaslavsky and Buntin, 2002). These variables were age (by five year intervals), gender, age interacted with gender, whether the respondent needed help with instrumental activities of daily living (IADLs) or with activities of daily living (ADLs), whether the respondent had a history of heart disease, cancer, stroke, diabetes, or chronic obstructive pulmonary disease (COPD), whether the respondent needed special medical equipment, and the respondent's self-rated health status on a five-point scale from excellent to poor. These covariates have been shown to be associated with significant differences in the use of health care services. They were fixed across alternate estimation strategies.

4. Potential methods: what estimators should be considered?

Given our research objectives, none of the estimators suggested in the health econometrics literature could be rejected out of hand. Each of the estimators is described briefly below; see also the review by Jones (2000).

4.1. Two-part models: modeling the probability of any use of care

Two-part models are often used to model cost data that include many zero observations. The first part of the model predicts the probability of any use, specified as a probit.

$$\mathrm{Prob}(y_i > 0) = \Phi(x'\beta) \tag{1}$$

where Φ represents the standard normal cumulative distribution function, or logit

$$\mathrm{Prob}(y_i > 0) = \frac{\mathrm{e}^{\chi'\beta}}{1+\mathrm{e}^{\chi'\beta}}. \tag{2}$$

The logistic distribution has heavier tails than the normal but the two models produce very similar results in practice over a wide range of values (Greene, 1993).

The second part of the model predicts costs conditional on nonzero costs; alternative specifications of this part of the model are discussed below. To obtain unconditional predicted costs, the probabilities of use from the first part are multiplied by expected levels from the second part of the model:

$$E(y_i|x_i) = \mathrm{Prob}(y_i > 0|x_i)\, E(y_i|x_i, y_i > 0). \tag{3}$$

4.2. *Modeling expenditure and use levels: transformed OLS models*

The second part of a two-part model is often an OLS model with a transformed outcome variable. The log transformation "pulls in" the upper tail of the distribution, as does the square root transformation to a lesser extent. For example, Ettner et al. (1998) found that models using the square root transformation fit their data on psychiatric services better than those using the log transformation because the distribution of their psychiatric data was less skewed than that of total health care costs.

When dependent variables are transformed, predictions must be retransformed back to the original scale to draw useful conclusions about the original variables: as noted by Manning (1998), "Congress does not appropriate log dollars." The expected value of a conditionally lognormal variable is

$$E(y|y>0; x) = \exp\left(x'\beta + \tfrac{1}{2}\sigma^2\right). \tag{4}$$

If the error term is not normally distributed then the smearing estimator developed by Duan (1983) consistently estimates the expectation provided the errors are independent and identically distributed. The smearing factor is the average exponentiated residual from the OLS regression:

$$S = \frac{1}{n}\sum_{i=1}^{n}\exp(e_i), \qquad \text{where } e_i = \log y_i - x_i'\hat{\beta}. \tag{5}$$

The exponentiated predictions are then multiplied by the smearing factor to predict expected values on the raw (unlogged) scale.

$$E(y|x, y>0) = S\exp(x'\beta). \tag{6}$$

A single smearing factor is often employed, but this can be problematic. If S depends on x because the distributional form or scale of the residuals is related to x, the predictions will be biased. If the error term is heteroscedastic, the retransformation should model that heteroscedasticity (Manning, 1998). For example, when the RAND Health Insurance Experiment researchers originally employed the smearing factor they minimized heteroscedasticity by creating a separate smearing factor for each insurance plan and type of care (Newhouse, 1993).

Heteroscedasticity (error variance dependent on $x\beta$) might cause a model with a single smearing factor to under- or overpredict in a particular range of predicted costs. In this study we evaluated use of a single smearing factor and use of a second smearing factor for the top decile of predicted costs. This was motivated by the consideration that lack of model calibration has the most serious consequences at the high end of the distribution of costs.

There is a similar retransformation for predicting the mean on the original scale from a square-root transformed OLS regression, although in this case the adjustment is additive rather than multiplicative. For a homoscedastic regression model,

$$E(y|y > 0; x) = (x'\beta)^2 + \sigma^2,$$

regardless of the distributional form of the residuals. If the square-root scale residuals are heteroscedastic, then separate estimates of σ^2 can be calculated for ranges of predicted values, the equivalent of calculating separate smearing factors in the log-transformed model.

4.3. Quasi-likelihood generalized linear models (GLMs)

GLMs directly model both the mean and variance functions on the original scale of y (Nelder and Wedderburn, 1972; McCullagh and Nelder, 1989). The mean function, $E(y|x)$ is represented as $\mu(x'\beta)$, where μ is the inverse link between the expectation of the observed raw-scale y and the linear predictor $x'\beta$. As discussed below, the log is commonly chosen as the link function in health expenditure applications; then μ is the exponential function. A commonly used family of variance functions includes the power functions of the form

$$v(x) = \kappa(\mu(x'\beta))^{\lambda}. \tag{7}$$

Where $\lambda = 0$ the variance is constant. Where $\lambda = 1$ the variance is proportional to the mean (sometimes called a “Poisson-like” model because the variance function for the Poisson distribution takes this form); and where $\lambda = 2$, the variance is proportional to the mean squared (sometimes called a “gamma-like” model because a gamma scale family, or indeed any scale family, has a variance function of this form). However, the GLM residuals need not take these distributional forms and indeed there is no compelling reason why λ should even be an integer. In quasi-likelihood models the relationship between mean and variance are specified, but not the full distribution of the residuals. Thus they may also be characterized as nonlinear least squares (NLS) models (weighted inversely proportional to the variance function). These models have been implemented in several standard statistical packages.[1]

Generalized linear modeling has recently received considerable attention in the health economics literature (Mullahy, 1998; Blough et al., 1999; Manning and Mullahy, 2001). It is attractive because the link function directly characterizes how the expectation on the original scale is related to the predictors. In health care cost modeling, the log-link relationship is usually chosen, $\ln(E(y)) = x'\beta$, or $E(y|x) = \exp(x'\beta)$. With this link, an effect on the linear

[1] A simple model with outcome y and one predictor x is implemented in Stata using the command glm y x, family (z) link(log) where z is gauss for $\lambda = 0$, poisson for $\lambda = 1$, and gamma for $\lambda = 2$. In S-Plus, a typical model would be fit by glm(data, formula = $y \sim x$, family = quasi(link = log, variance = constant)) where constant can be replaced by mu or muˆ2 for $\lambda = 1$ or 2, respectively. In SAS, the corresponding language is PROC GENMOD; MODEL Y = X/LINK=LOG DIST = z; where z is NORMAL, POISSON or GAMMA as in Stata.

predictor can be interpreted directly as a multiplicative effect on total costs. The results in the GLM model can thus be interpreted directly without retransforming the results from the log to the original scale. In contrast, in log-transformed OLS models $E(\ln y|x)) = x'\beta$, which does not directly translate into any statement about $E(y|x)$. The GLM model can be estimated on the entire sample, since zeroes in the data pose no problem for fitting such models, or it may be used as the second part of a two-part model.

In addition, there are general grounds on which the GLM models might be preferable to modeling the log-transformed dependent variable, when modeling the mean structure is ultimately of interest. The mean of logged data can be very sensitive to small changes in the distribution of values in the left tail, even if that tail accounts for a very small portion of the total. For example, the mean on the logged scale is equally affected by changing some observations from US$ 1 to 2 (a trivial change by almost any standard) as by changing the same number of observations from US$ 1000 to 2000 (a substantial change). In principle, the smearing estimator could correct for the change in the tail distribution, but if the distributional form of the observations in the left tail varies across the range of predicted values (as is likely to be the case), it would be necessary to estimate smearing factors that might vary continuously across the range of fitted values. Estimation of a separate smearing factor for the top decile is a simple, somewhat adhoc fix that can improve model fit where it matters most—at the top of the range of predicted costs. Nonetheless, whatever method is used to incorporate variation in the smearing factors, the model predictions are then particularly difficult to interpret since they involve both the transformed linear prediction and the functions that determine the smearing factor. Selection of a link function relating the linear predictor to the mean costs is more interpretable.

Another advantage of the GLM approach is that it separates the specification of the mean function (i.e. the link function μ) from that of the variance function κ. If the mean function (link and linear predictor) are correctly specified, the choice of the variance function is essentially a question of efficiency. The estimation procedure weights the observations inversely proportional to their variances, and the standard errors of the coefficient estimates will be smallest if the variance function is specified (approximately) correctly. In the worst case, misspecification of the variance function (typically, assuming homoscedasticity when actually residual variance grows with the predicted mean) might cause the estimation algorithm to fail altogether to converge, because the procedure is overly sensitive to the outliers that might appear at the high end of such a distribution.

If the mean function is misspecified, as will almost always be the case to at least some degree, the model does not fit the data equally well across the entire range of predicted values. Then optimal fit in one part of the range implies at least slightly worse fit in another part. In that case, the variance function affects the relative weighting of goodness of fit in different parts of the range. A model fitted with $\lambda = 0$ (constant variance) will fit better at the upper end of the range, compared to one with $\lambda = 2$ (variance proportional to mean squared) which gives less weight to error at the high end. Thus with a misspecified link or linear predictor, the variance function affects both the efficiency of estimation and the criterion of fit (driven by scientific rather than purely statistical considerations), and the choice of variance function might reflect a compromise between the two.

A potential disadvantage of the GLM approach is that GLM estimation can be less efficient (i.e. less precise for a given sample size) than OLS, a natural consequence of its weaker model assumptions. In particular, if the variance function is misspecified or if the log scale residuals have high kurtosis, then the GLM estimates may be imprecise (Manning and Mullahy, 2001).

5. Model selection procedures

The models described in Section 4 are the alternatives from which health economists customarily choose when modeling health care cost or use data. The first step in evaluating these models is to examine the distribution of the data to be modeled in more detail and conduct the various diagnostic tests proposed in the literature.

As previously described, the distribution of costs in the MCBS sample was skewed (asymmetrically distributed) with many zeros. Would transforming the positive part of the distribution by taking the natural log or the square root of the cost ratio variable approximate a normal distribution? A histogram of the log-transformed nonzero data appeared much more nearly symmetric than that of the raw or square-root transformed data. Since the distribution is not perfectly lognormal, however, if the logged dependent variable OLS method is employed, the lognormal retransformation might not be consistent, while the smearing retransformation (5) might be more reliable. Plotting of the data with various transformations is thus a useful preliminary step, but ultimately the distribution of the residuals from the transformed model is critical because the model assumptions concern the distribution of the residuals, not of the data. In this case, however, the residuals from the second part of the logged OLS model are also approximately normal.

Given this, the two-part log-transformed OLS model might be a good estimator. However, as noted earlier, if the variance or distributional form of the errors is related to the predictions, then retransformed estimates of the expectation could be biased.

We next consider the GLMs. In our MCBS data the kurtosis of the log scale residuals from the OLS fit is 3.05, only slightly exceeding 3, the value for the normal distribution. This suggests that with an appropriate choice of variance function, the GLM might be reasonably efficient. To help make this choice, Manning and Mullahy (2001) suggest the Park test to estimate the relationship between the mean and the variance (Park, 1966). This procedure regresses the squared residuals from a provisional model (GLM or logtransformed OLS) on the predictions ($\hat{y}$) from the same model, both log transformed:

$$\ln((y_i - \hat{y}_i)^2) = \lambda_0 + \lambda_1 \ln(\hat{y}_i) + v_i. \tag{8}$$

The coefficient λ_1 corresponds to λ in (7), indicating which GLM variance function is appropriate. For the MCBS data, this procedure estimated $\lambda = 1.85$ (S.E. $= 0.03$), with little sensitivity to the choice of provisional model. This suggests that the constant variance model ($\lambda = 0$) is not a good candidate and the best model falls between the variance proportional to mean ($\lambda = 1$) model and the standard deviation proportional to mean ($\lambda = 2$) model. Manning and Mullahy also noted that it can be hard to tell which GLM model to use from fitting the alternative models to the data since all models will overfit the data when the data are skewed or kurtotic. Robust variance estimators (for standard errors of parameters) or cross-validation (for comparisons of predictive accuracy) can better compare models.

Overall, the results of these tests did not point to a single superior model for predicting Medicare costs. The nonzero data seem to be distributed roughly but not perfectly lognormally, so the log-transformed two-part OLS models could work well. While retransformed means from these models are biased by heteroscedasticity, the homoscedastic OLS model is the most precise if its assumptions hold (Manning and Mullahy, 2001). The variance proportional to mean and variance proportional to mean squared GLM models are possibilities, as suggested by the results of the Park test.

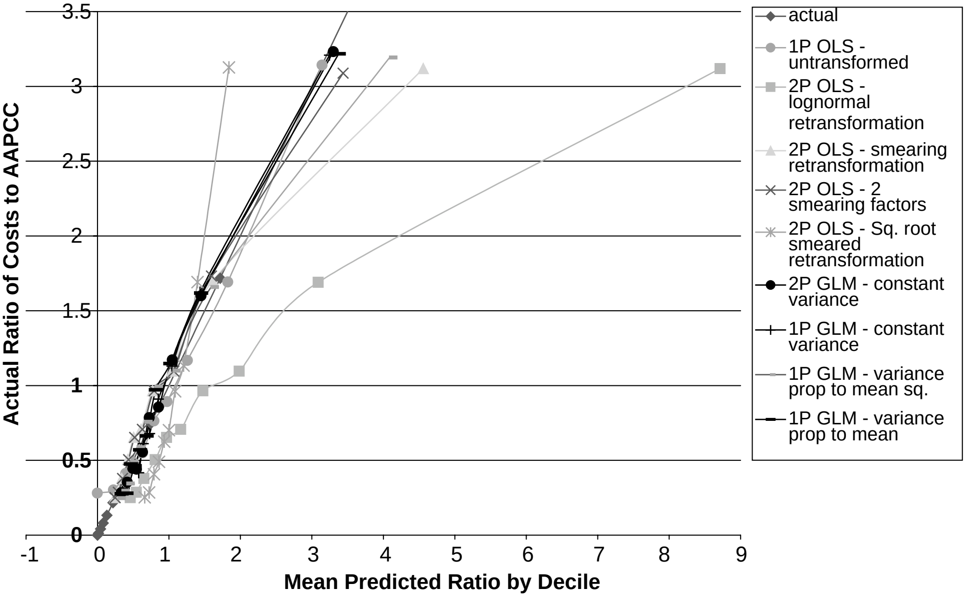
For illustrative purposes, we also estimated the square-root transformed model, the two-part OLS logged dependent variable model with the homoscedastic retransformation, and the constant variance GLM model. The constant variance GLM model was estimated in both one- and two-part versions, illustrating the fact that the GLM can be used in either setting (Mullahy, 1998). All of our GLM models use a log link function. In addition, for comparison, we fit a standard one-part OLS model without a transformation of the dependent variable.

6. Fitting models to the entire sample

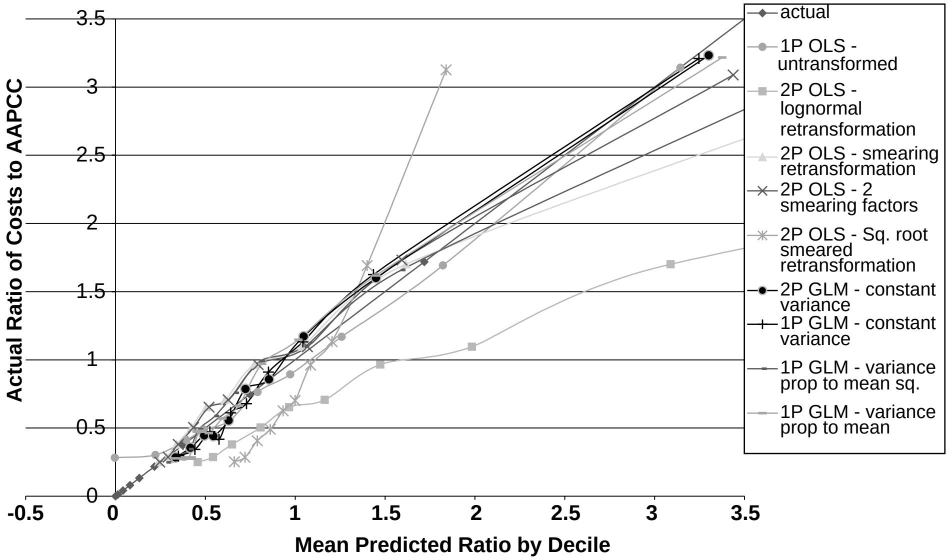
Nine alternative models were estimated on the MCBS data, including the standard OLS model. The models' predictions were then examined to see how well they predicted costs. We also investigated how well calibrated they were against sample means for groups of observations in each range of predicted costs and for other analytically relevant subgroups of the sample.

Schemes 1 and 2 illustrate the calibration of the alternative estimators, that is, how well predicted means agree with sample means, across the range of predicted values. Scheme 1 plots the mean costs predicted by each of the alternative estimators, by decile of predicted costs, against the actual mean spending for the individuals in that decile. This provides a graphical depiction of the degree to which these models can estimate actual costs across the span of the data. The actual decile means form a 45° line: the points on the other estimator lines above or below this 45° line indicate where the other estimators over-predict (fall below the line) or under-predict (rise above the line).

It is notable that all of the estimators do about equally well in detecting the high-cost group. The mean actual cost ratio for the cases predicted to be in the top decile is between 3.08 and 3.22 for each estimator. (The mean cost ratio for the true top decile is much higher, about 6.4, but it is not surprising that our limited set of variables could not perfectly identify this group.) The estimators differ, however, in how well their predictions match the true mean cost ratios. The square-root transformed OLS model drastically underpredicts in the top decile and overpredicts in the lower deciles, suggesting that this transformation does not sufficiently "pull in" the upper end of the distribution to linearize the relationships. All of the log-transformed models, on the other hand, overpredict in the top decile. The best linearizing transformation might be intermediate between the square-root and log transformations, for example a 0.2 power transformation, however, we did not further explore such nonstandard alternative transformations (Basu and Rathouz, in preperation). Four of the log-transformed models are about equally well calibrated at the top decile: the two-part model with log-transformed OLS and two smearing factors, the two-part and one-part constant variance



Scheme 1. Mean predictions of alternative estimators plotted against actual cost ratios, by decile.



Scheme 2. Detail, lower end of the distribution, mean predictions of alternative estimators plotted against actual cost ratios, by decile.

GLM models, and the one-part variance proportional to mean GLM model. Each of these models has features that cause it to fit well at the upper end of the distribution: the separate smearing factor in the top decile, or the variance function that gives equal (or almost equal) weight to errors across the range of predicted costs. The GLM with variance proportional to mean squared downweights errors at the high end and therefore fits less well there, illustrating the sensitivity of the fit to the choice of variance function. The two-part OLS models with a single smearing factor or the lognormal retransformation drastically overpredict in the top decile (in the latter case by a factor of almost three); these models are most dependent on the accuracy of model assumptions that apparently do not hold for these data.

Scheme 2 expands the same presentation for the lower end of the distribution. The mean predictions of the four estimators that fit well in the top decile also track the observed means well in the other deciles: they cluster around the actual distribution at the lower end and follow it with a slight degree of over-prediction through the middle deciles. Thus these estimators are about equally well calibrated. The square root and lognormal (retransformed) two-part OLS models overpredict at the lower end of the distribution. The untransformed one-part OLS model underpredicts at the low end of the distribution, in fact it actually predicts negative mean costs in the lowest decile. Conversely, it overpredicts in the seventh through ninth deciles of predicted costs.

Table 1 assesses the calibration of predictions from the eight models by comparing their predictions of the sample mean to the actual sample mean, for the entire sample and some analytically relevant subgroups. Of the eight models, the GLM model with variance proportional to mean predicts the sample mean the best in five of the eight cases shown. It predicts particularly well for groups split on whether or not they have functional limitations. The two-part constant variance GLM model is close behind. The one-part constant variance and two-part logged OLS with two smearing factors also fare well, not surprisingly, predicting the mean costs of beneficiaries with chronic conditions particularly well. The lognormal and square root transformations perform poorly— especially for sicker groups (i.e. those with chronic conditions, functional limitations, and in poor health). These comparisons do, however, show the strength of simple OLS models: by construction the OLS regression on the untransformed data perfectly predicts the overall mean and the means of the subgroups defined by variables included in the models. Likewise, the one-part GLM variance proportional to mean estimator will have a zero in-sample mean residual if a constant term is included.

Table 2, columns (1) and (2) show the mean square error and absolute prediction error for each of the models in predicting the cost ratios across the entire sample. The two constant variance GLM models have the lowest mean squared error because they give more weight to accuracy of the large predicted values, and correspond to the unweighted MSE criterion used here. The variance proportional to mean GLM model and the two-part logged OLS with two smearing factors have increasingly large prediction errors (in that order.) The four other models do considerably worse. In terms of absolute prediction error, the model with the two smearing factors performs the best, perhaps because the second factor improves fit in the top decile where the errors are largest, and all of the GLM models do well, especially the variance proportional to mean model. The standard OLS model does fairly well in terms of squared error, but less well in terms of absolute error.

Table 1
Predictions of the mean cost ratio for the entire sample and relevant subgroups, from alternative estimators

Estimator	Mean of sample	Beneficiaries with chronic conditions	Beneficiaries without chronic conditions	Beneficiaries with ADL limitations	Beneficiaries without ADL limitations	Beneficiaries in poor health
1P OLS—untransformed	0.97	1.30	0.40	2.70	0.78	3.18
2P OLS—lognormal retransformation	1.98	2.72	0.72	6.95	1.46	8.26
2P OLS—smearing retransformation	1.03	1.43	0.38	3.63	0.76	4.32
2P OLS—2 smearing factors	0.93	1.26	0.38	2.84	0.73	3.32
2P OLS—square root smeared retransformation	1.05	1.20	0.79	1.68	0.98	1.79
2P GLM—constant variance	0.97	1.23	0.56	2.57	0.81	3.18
1P GLM—constant variance	0.98	1.23	0.56	2.57	0.81	3.18
1P GLM—variance proportional to mean squared	1.02	1.37	0.44	3.19	0.79	3.66
1P GLM—variance proportional to mean	0.97	1.24	0.50	2.70	0.78	3.18
Actual	0.97	1.28	0.43	2.70	0.78	3.18

Table 2
Mean square error (MSE), mean absolute prediction error (MAPE), and mean square forecast errors (MSFE) of alternative estimators

	(1) MSE whole sample	(2) MAPE whole sample	(3) Average MSFE over 100 split-samples	(4) Average MAPE over 100 split-samples
1P OLS—untransformed	4.792	1.073	4.849	1.081
2P OLS—lognormal retransformation	10.081	1.704	10.478	1.718
2P OLS—smearing retransformation	5.241	1.080	5.367	1.097
2P OLS—2 smearing factors	4.853	1.032	4.903	1.042
2P OLS—square root smeared retransformation	5.133	1.180	5.158	1.185
2P GLM—constant variance	4.687	1.064	4.924	1.067
1P GLM—constant variance	4.689	1.070	4.923	1.072
1P GLM—variance proportional to mean squared	4.945	1.075	5.038	1.085
1P GLM—variance proportional to mean	4.718	1.052	4.815	1.060

7. Split sample cross-validation

We used cross-validation to evaluate reliably the predictive accuracy of the models, fitting the models to one part of the sample and assessing predictive accuracy on the remaining part of the sample. Thus, we simulated the accuracy with which a model fit to one dataset might predict the costs for individuals in another sample from a similar population. This approach protects us from misleadingly optimistic assessments due to overfitting of complex models. Furthermore, we cannot assess models by comparing standard measures of fit, such as adjusted R^2 values, because of the number of different types of models used, including two-part models. We fit each of the eight models to a randomly sampled half of the data and predicted costs for the other half of the data, repeating the exercise 100 times. All of the models were evaluated on the same split samples to reduce the variance of comparisons among models.

We summarized predictive accuracy using the mean squared forecast error (MSFE), defined as the mean squared difference between predicted and actual costs (Duan et al., 1983). The variance proportional to mean GLM model had the smallest MSFE averaged over replications of the cross-validation (Table 2, column (3)) and also had lower MSFE than the other models in most cross-validation replications (Table 3). The second-best model by either of these measures is the two-part logged OLS with two smearing factors. These cross-validation results confirm Manning and Mullahy's (2001) finding of within-sample overfitting with the constant variance GLM model since those models did not perform as well here as they did in terms of MSE on the entire sample. Although the two-part version of this model performed better than the one-part version in 69 out of 100 samples, the differences between the accuracy of the predictions of the models was extremely small on average, less than one percent. Using an alternative criterion of fit, the mean absolute prediction error, produced similar results. As with the APE for the entire sample, however, the two-part OLS model with two smearing factors performed best, with the GLM variance proportional to mean model close behind. Table 3 confirms that in terms of mean absolute prediction error

Table 3
Split-sample cross-validation of alternative estimators

	1P OLS—untransformed	2P OLS—lognormal retransformation	2P OLS—smearing retransformation	2P OLS—2 smearing factors	2P OLS—square root smeared retransformation	2P GLM—constant variance	1P GLM—constant variance	1P GLM—variance proportional to mean squared	1P GLM—variance proportional to mean
1P OLS—untransformed									
MSFE	–	0	99	71	100	66	68	92	27
MAPE	–	0	70	3	100	9	22	70	0
2P OLS—lognormal retransformation									
MSFE	0	–	0	0	0	0	0	0	0
MAPE	0	–	0	0	0	0	0	0	0
2P OLS—smearing retransformation									
MSFE	1	100	–	0	25	7	6	3	0
MAPE	30	100	–	0	100	18	30	35	9
2P OLS—2 smearing factors									
MSFE	29	100	100	–	100	57	58	92	12
MAPE	97	100	100	–	100	91	93	98	88
2P OLS—square root smeared retransformation									
MSFE	0	100	75	0	–	9	9	23	0
MAPE	0	100	0	0	–	0	0	0	0
2P GLM—constant variance									
MSFE	34	100	93	43	91	–	69	75	8
MAPE	91	100	82	9	100	–	99	95	18
1P GLM—constant variance									
MSFE	32	100	94	42	91	31	–	78	6
MAPE	78	100	70	7	100	1	–	85	8
1P GLM—variance proportional to mean squared									
MSFE	8	100	97	8	77	25	22	–	0
MAPE	30	100	65	2	100	5	15	–	0
1P GLM—variance proportional to mean									
MSFE	73	100	100	88	100	92	94	100	–
MAPE	100	100	100	12	100	82	92	100	–

Figures indicate the number of times out of 100 split samples that the MSFE or MAPE of the row was lower than the MSFE or APE of the column.

(MAPE) the model with two smearing factors and the one-part GLM variance-proportional-to-mean models were the best.

8. Discussion

Four out of the eight alternative models estimated here performed well in terms of calibration of predictions, mean square error, absolute prediction error, and cross-validated forecast error. One was an OLS model that used two smearing factors to correct for heteroscedasticity. Two were constant variance GLM models, in either a one-part or two-part overall formulation. Of these, the two-part model fit the data slightly better than the one-part model, but the difference was very small. The model that best predicted the sample mean and had the lowest mean squared forecast error in the split-sample cross-validation exercise was the variance proportional to mean (Poisson-like) GLM model. The model with two smearing factors had the second lowest mean squared forecast error and the lowest mean absolute prediction error.

We found that the diagnostic tests suggested in the literature were not very helpful in ruling out alternative models, but our example does demonstrate where investments of time by researchers are likely to be fruitful. Plotting the raw and transformed data (and more important, the residuals from models with various transformations) helped us to establish whether the log transformation was a good fit for the data. Plotting predictions of the alternative models against true means by decile was a useful way to see how well their predictions match observed outcomes for ranges of predicted values, while mean squared errors were valuable as a summary of overall goodness of fit. In addition, plotting the predictions by decile showed us that a separate smearing factor for the top decile would improve our predictions.[2] In our case the split-sample cross-validation exercise was informative since it showed that the constant variance GLM models were inferior to the mean proportional to variance model and validated other comparisons.

It is informative to examine the fit of the models across the range of predicted values. As mentioned above, when the mean function does not perfectly fit the data, the choice among the GLM variance functions reflects a compromise between efficiency and fit. All of our GLM models employ a log link function, and if this link fit perfectly then the variance proportional to standard deviation model would be the most efficient. However, the equal variance model fit as well across the range of predicted values because it fit better at the top end of the distribution. This illustrates the importance of examining the performance

[2] The separate smearing factor we calculated for the top decile corrects for the effects of both the lack of fit of the link function and heteroscedasticity in the residual variance of the OLS model on that decile. A third smearing factor (or more) might be added to improve fit in other parts of the range, but examining the fit of this model in other deciles confirmed lack of fit in the top decile was the most serious problem. This focus on fit in the top of the range is based on similar concerns to those of the RAND Health Insurance Experiment researchers, who found that four-equation models that accounted for high inpatient spending were preferable to two-equation models (Newhouse, 1993). This somewhat adhoc fix to the transformed OLS procedure could be expected in general to be suboptimal compared to methods that more directly address lack of fit by fixing the link and/or variance functions, but we have found in practice that the linear predictor is less sensitive to these aspects of the model than the predictive accuracy.

Table 4
Comparison of coefficients from one-part GLM models

	Constant variance	95% Confidence interval		Variance proportional to mean squared	95% Confidence interval		Variance proportional to mean	95% Confidence interval	
Age 70–74	−0.023	−0.165	0.118	−0.329*	−0.532	−0.125	−0.168*	−0.260	−0.075
Age 75–79	−0.250*	−0.409	−0.092	−0.307*	−0.519	−0.095	−0.293*	−0.390	−0.197
Age 80–85	−0.289*	−0.444	−0.134	−0.234*	−0.451	−0.017	−0.279*	−0.375	−0.184
Age 85+	−0.613*	−0.828	−0.399	−0.442*	−0.689	−0.194	−0.526*	−0.641	−0.410
Female	0.218*	0.093	0.342	0.084	−0.117	0.286	0.108*	0.023	0.193
Female*Age 70–74	−0.143	−0.325	0.039	0.240	−0.037	0.516	0.051	−0.071	0.174
Female*Age 75–79	−0.056	−0.253	0.140	0.304*	0.021	0.586	0.150*	0.026	0.275
Female*Age 80–85	−0.189	−0.383	0.005	0.078	−0.207	0.363	−0.027	−0.150	0.096
Female*Age >85	0.012	−0.234	0.258	0.145	−0.168	0.457	0.110	−0.030	0.250
Difficulty with IADLs	0.553*	0.454	0.651	0.499*	0.339	0.659	0.518*	0.459	0.577
Difficulty with ADLs	0.383*	0.305	0.461	0.424*	0.243	0.604	0.400*	0.343	0.457
Heart disease	0.425*	0.348	0.501	0.583*	0.489	0.678	0.487*	0.445	0.530
Cancer (not skin)	0.332*	0.266	0.398	0.445*	0.328	0.562	0.343*	0.298	0.388
Stroke	0.142*	0.072	0.212	0.345*	0.203	0.488	0.228*	0.178	0.278
Diabetes	0.249*	0.181	0.316	0.270*	0.142	0.398	0.258*	0.211	0.305
Asthma, emphysema, COPD	0.026	−0.048	0.099	0.233*	0.099	0.366	0.145*	0.095	0.195
Need equipment to walk	0.228*	0.149	0.307	0.433*	0.291	0.576	0.314*	0.261	0.366
Ever a smoker	0.253*	0.179	0.327	0.256*	0.153	0.359	0.243*	0.198	0.289
Rate health “fair”	0.333*	0.235	0.431	0.418*	0.295	0.541	0.355*	0.303	0.406
Rate health “poor”	0.740*	0.642	0.839	0.709*	0.515	0.902	0.692*	0.631	0.753
Intercept	−0.824*	−0.960	−0.687	−1.112*	−1.288	−0.937	−0.922*	−1.000	−0.844

Notes: Age 65–69 and self-rated health status of good, very good, or excellent are omitted categories. (*) Indicates significant at the 0.05 level. Scheme 1: mean predictions of alternative estimators plotted against actual cost ratios, by decile.

of the models across the distribution and then choosing the variance function that weights deviations at the high versus the low end of the distribution in the way that best balances efficiency and fit.

In this analysis we did not consider alternative link functions for the GLM model. Our results suggest that the log link was fairly satisfactory, but it should not be assumed that this would be true for any cost data. Fortunately, it is fairly easy to introduce other link functions within the GLM framework, such as the power family in which the log link is embedded (Basu and Rathouz, in preperation).

The models that best fit the data in this example might not be the best in other applications. For example, the data used here came from a population with a relatively low percentage of beneficiaries with zero costs (<9%). The two-part models might, therefore, improve precision less than they would for data with a larger percentage of zeroes.

Our analysis focused on prediction rather than on testing hypotheses about effects of covariates. Were the purpose of the study to draw inferences, for example about the effects of beneficiary characteristics on costs, the estimated coefficients from various models would have to be compared. This is possible for the various one-part GLMs, because they all have the same link function, but direct comparisons are not possible for differently-specified models (such as two-part models or transformed OLS models with multiple smearing factors). For example, Table 4 shows that the inferences that would be drawn about the effects of patient characteristics, such as age, gender, and disease are similar across the GLM models. However, there are some differences, most notably the effects of “asthma, emphysema, COPD,” gender, and younger age on costs, which are significant in only two of the three GLMs. The variance-proportional-to-mean model (whose variance functions is intermediate between the other GLMs considered) has estimates for most covariates that fall between the estimates, and usually within the confidence intervals, of the other GLMs. The variation among these fitted models, which differ in the relative emphasis given to goodness of fit at lower or higher predicted costs, suggests that an adjustment of the link function might improve fit over the entire range.

Correct specification of the variance function (or use of a robust variance estimation procedure) is also important to obtain valid standard errors and hypothesis tests for the coefficients. Nonetheless, the one-part GLM models have some important advantages: their predictions are not biased by heteroscedasticity and the coefficients are easier to interpret than those from a two-part model if total costs are of interest.

9. Conclusions

This paper demonstrates that finding the “best” estimator for a given problem and data set can require a researcher to perform a large number of specification checks, some of which add little value. After performing the extensive plotting, testing, and cross-validation exercises described above, we found that at least four of the estimators performed well enough to be used in many applications. Future research could shed light on whether the models that performed best for our sample of Medicare beneficiaries also perform best with younger populations, with longer periods of observation, and/or with other sets of explanatory variables.

Given these results, a researcher considering alternative models where the probability of use per se is not of interest would do well to start with the one-part GLM models. They are easier to estimate in current statistical packages than the two-part OLS models and avoid the problem of having to make post hoc adjustments for heteroscedasticity (common in models of health care costs and use) to remove biases in predicted means for the entire population and relevant subgroups. Furthermore, zero observations can be accommodated without difficulty within this framework.

The researcher taking this route must choose among GLM mean (link) and variance functions to maximize efficiency and fit. First, the researcher should determine if a log link or some other link best approximates the mean function. To do this, the fit of alternative GLM estimators should be evaluated by calculating mean squared error and by plotting mean predictions by decile against mean observed values. Second, the researcher should choose the best variance function, applying the Park test described in Section 5 to residuals from a good candidate model.

If none of the alternative GLM links gives a good fit across the range of the data, then the researcher could try a two-part version of the best-fitting GLM model in order to increase the flexibility of the specification. The researcher interested in modeling the probability of nonzero use per se should proceed directly to this step.

If the best-fitting one- or two-part GLM models cannot be estimated with adequate precision, use of transformed OLS models trades robustness for some improvement in efficiency. If preliminary results with GLM models suggest that residual variance is not proportional to the mean squared, then predictions from log-transformed OLS models are likely to be affected by heteroscedasticity. If the data are homoscedastic on the transformed scale, however, or there are theoretically justified ways to correct for heteroscedasticity, then the researcher could proceed to use the two-part transformed OLS model.

If there are reasons to use the OLS models but no obvious candidates for correcting for heteroscedasticity are present, then the use of smearing factors for different parts of the distribution should be considered. In the case examined in this paper using a separate smearing factor for the top end of the distribution more accurately predicts the high-cost cases than a single smearing factor. With this type of adjustment, comparisons of predicted group means are also made more accurate. With such adjustments, however, the coefficients of the fitted models are difficult to interpret since the mean predictions are also affected by the smearing factors.

Given our finding that many of the models perform equally well, we can advocate the method outlined above rather than the traditional exercise of trying a large number of diagnostic tests and model permutations. By starting with the one-part GLM models, and working through the course of action we describe, researchers should find a model that suits their purposes in the fewest steps.

Acknowledgements

We thank Paul D. Cleary, Richard Frank, Emmett Keeler, Joseph Newhouse, and Kathy Swartz for helpful comments, Larry Zaborski and Kanika Kapur for assistance, and the Commonwealth Fund and the Center for Medicare and Medicaid Services (contract

500-95-007) for support. Melinda Beeuwkes Buntin acknowledges support from the Harvard/Sloan Center for the Study of the Managed Care Industry.

References

Adler, G.S., 1994. A profile of the Medicare Current Beneficiary Survey. Health Care Financing Review 15 (4), 153–163.

Basu, A., Rathouz, P. Estimating marginal and incremental effects on health outcomes using flexible. Link and variable function models: University of Chicago, in preperation.

Blough, D.K., Madden, C.W., Hornbrook, M.C., 1999. Modeling risk using generalized linear models. Journal of Health Economics 18, 153–171.

Duan, N., 1983. Smearing estimate: a nonparametric retransformation method. Journal of the American Statistical Association 78, 605–610.

Duan, N., Manning, W.G., Morris, C.N., Newhouse, J.P., 1983. A comparison of alternative models for the demand for medical care. Journal of Business and Economic Statistics 1 (2), 115–126.

Duan, N., Manning, W.G., Morris, C.N., et al., 1984. Choosing between the sample-selection model and the multipart model. Journal of Business and Economic Statistics 2, 283–289.

Ettner, S.L., Frank, R.G., McGuire, T.G., Newhouse, J.P., Notman, E.H., 1998. Risk adjustment of mental heath and substance abuse payments. Inquiry 35 (2), 223–239.

Greene, W.H., 1993, Econometric Analysis, 2nd ed. Englewood Cliffs, Prentice Hall, NJ.

Jones, A.M., 2000. In: Cuyler, A.J., Newhouse, J.P. (Eds.), Chapter Six: Health Econometrics in Handbook of Health Economics, vol. 1, Elsevier.

Manning, W.G., Morris, C.N., Newhouse, J.P., et al., 1981. A two-part model of the demand for medical care: preliminary results from the Health Insurance Study. In: van der Gaag, J., Perlman, M. (Eds.), Health, Economics, and Health Economics. North Holland, Amsterdam, pp. 103–123.

Manning, W.G., 1998. The logged dependent variable, heteroscedasticity, and the retransformation problem. Journal of Health Economics 17, 283–295.

Manning, W.G., Mullahy, J., 2001. Estimating log models: to transform or not to transform? Journal of Health Economics 20 (4), 461–494.

McCullagh, P., Nelder, J.A., 1989. Generalized Linear Models, 2nd ed. Chapman and Hall, London.

Mullahy, J., 1998. Much ado about two: reconsidering retransformation and two-part model in health econometrics. Journal of Health Economics 17, 247–281.

Nelder, J.A., Wedderburn, R.W.M., 1972. Generalized linear models. Journal of the Royal Statistical Society, Series A 135, 370–384.

Newhouse, J.P. 1993. Free-For-All: Health Insurance, Medical Costs, and Health Outcomes: The Results of the Health Insurance Experiment. Harvard University Press, Cambridge.

Park, R., 1966. Estimation with heteroscedastic error terms. Econometrica 34, 888.

Zaslavsky, A., Beeuwkes Buntin, M., 2002. Using survey measures to assess risk selection among Medicare managed care plans. Inquiry 39, 138–151.