Biostatistics (2004), **1**, 1, *pp.* 1–23
Printed in Great Britain

Analyzing Health Care Costs: A Comparison of Statistical Methods Motivated by Medicare Colorectal Cancer Charges

By MICHAEL GRISWOLD

Department of Biostatistics,
Johns Hopkins University, Bloomberg School of Public Health
615 N. Wolfe st., Baltimore, MD. 21205
e-mail: mgriswol@jhsph.edu

GIOVANNI PARMIGIANI

Departments of Oncology, Biostatistics and Pathology
Johns Hopkins University, School of Medicine
550 North Broadway, suite 1103, Baltimore, MD. 21205

ARNIE POTOSKY

Division of Cancer Control and Population Sciences, National Cancer Institute
Get Address

and JOSEPH LIPSCOMB

Outcomes Research Branch, ARP, DCCPS, National Cancer Institute
Get Address

Summary

Datasets characterized by highly non-Gaussian distributions pose interesting challenges for prediction and comparison goals. Health care expenditure data is a common example where point masses and severe skewness often complicate analyses. Parametric approaches can improve efficiency characteristics of estimators but may sacrifice robustness in the process. We examine a variety of models commonly used to compensate for complex distributions and illustrate techniques for evaluating the competing models. The discussion is motivated by Medicare colorectal cancer charges. *Results, conclusions & suggestions summary here.* An extended description and additional information is available on the project website: http://biosun01.biostat.jhsph.edu/project/seermed/seermed.htm

1. Introduction

1.1. Motivation

Cost data is typically characterized by distributions that are difficult to describe using standard approaches. For example, costs are both highly skewed, (a result of a few patients incurring disproportionately high costs relative to the majority of patients), and can present point masses ("lumpiness")

at lower cost values, (a result of patients incurring no costs, minimal costs, or program-standard cost amounts) (Figure 1). The comparison of costs between groups of patients, and more generally the formulation of prediction models to describe determinants of variation in costs, can become challenging in these circumstances. Simple estimators, such as the difference in sample means for the group comparison example, can be inefficient and sensitive to individual observations, especially when sample sizes in the groups are substantially different. Common parametric approaches, such as the lognormal model, rely heavily on the mathematical form of the specified distribution and can fail to fit the data adequately even after the transformations are applied. Additionally, modelling in transformed scales, such as the logarithm, can create severe bias adjustment problems for estimating cost expectations. Weighed against these difficulties is the importance of accurate cost prediction for public health resource planning, for cost-effectiveness evaluations on subject specific interventions, and for investigating cost allocation differences among subpopulations of service users.

1.2. A Bit of Background

The following constitutes a short review of statistical methods commonly used for dealing with complex health care cost distributions. The logic underlying usage of each model is discussed to flavor the subsequent comparisons in the results and discussion sections.

Ordinary Least Squares: "A billion here and a billion there, and soon you're talking about real money." – *usually attributed to Senator Everett McKinley Dirksen* –. Databases used in health expenditure analyses tend to be large. The Law of Large Numbers then implies that mean cost estimates should be close to their respective population average costs. By the Gauss-Markov theorem, coefficient estimates under OLS will have the smallest variance among the class of linear unbiased estimators. One suggestion found in the literature then, is to not worry specifically about the form of the distribution with large data sets, since OLS may perform quite respectably in relation to more sophisticated models, (particularly when average expenditure estimation is the goal). The drawback is that inferences based on OLS standard error estimates may be suspect, since ignoring the shape of the distribution may lead to overstated significance via inaccurate standard errors and confidence intervals. Diehr et al. (1999) recommend OLS only when the goal is future cost prediction.

Lognormal Models: Aitchison and Brown (1957) give historical background on the use of this model, commonly applied to make skewed data 'look more normal' whereupon linear regression techniques can be applied with more confidence. Often lost in the proverbial mix are subtleties such as (1.) simple exponentiating of the estimated linear predictor leads to estimates of the median rather than the mean, (whereas the mean is often the *central* interest), and (2.) changes in explanatory variables lead to proportional rather than additive changes in the response. When the additional complication of zero values in the dependent variable arises, a common solution is to add 1, (or some arbitrary small number), before taking the log. Estimates based on the lognormal method can have better procedural properties, such as more appropriate variability estimates than simple analysis on the untransformed scale, and there are many cases where a lognormal model is eminently justified. However, as Manning (1998) eloquently relates:

> Although such estimates may be more precise and robust, no one is interested in log model results on the log scale per se. Congress does not appropriate log dollars. First Bank will not cash a check for log dollars. Instead, the log scale results must be retransformed to the original scale so that one can comment on the average or total response to a covariate x. There is a very real danger that the log scale results may provide a very misleading,

incomplete, and biased estimate of the impact of covariates on the untransformed scale, which is usually the scale of ultimate interest.

If the data actually are lognormally distributed, then the average response on the original scale is a function of both the mean and the variance on the log-transformed scale. Original scale expected costs can therefore be calculated by including an estimate of the residual variability when re-exponentiating.

Smearing: If the data are not lognormally distributed, the techniques above can give severely biased estimates for the expected cost. Duan (1983) developed a nonparametric solution for this problem by constructing a "smearing" estimate that distributes (smears) the excess error about observations when converting back to original scale means. In particular, the empirical error distribution is used when estimating the expectation instead of the assumed lognormal distribution. Manning (1998) investigates estimation for groups having heteroscedastic error terms and recommends the use of separate smearing coefficients in such cases.

Two-stage models: Tobin (1958) examined decompositions such as $E(Cost) = E(Cost|Cost > 0) \cdot Pr(Cost > 0)$, to account for limited dependent variables, (data with clusters of minimum values like zero costs). The tobit model, as it is called, constrains the parameters in the two stages however, and Cragg (1971) has an early treatment allowing a separable mixture (the parameters being determined independently for each stage.) Mullahy (1998) discuss problems arising from two-stage models when transformations are used for the conditional expectation stage.

Generalized Linear Models: Using Generalized Linear Models (McCullagh and Nelder 1989) to address health cost issues has been outlined by Blough et al. (1995) and Blough and Ramsey (2000) who approach estimation with a quasi-likelihood framework. They suggest constructing "profile-extended-quasi-likelihood" surfaces for estimating the optimal link and variance functions to employ in the GLM. We prefer to use a log-link function for clearer parameter interpretations, which in turn leads to specifying a "Gamma-class" distribution, ($Var(y|x) \propto E(y|x)^2$) as the "natural" error choice. An insightful discussion is given in Manning and Mullahy (2001), who perform simulations and present an example comparing three log-link GLM's and two lognormal models (with homoscedastic and heteroscedastic smearing factors) for positive expenditures. Appealing aspects of GLMs include their flexibility in modelling mean-variance relationships and their avoidance of retransformation issues when the mean is of primary interest.

Cox PHM: Survival analysis techniques have recently been a topic of debate for expenditure modelling, the Cox PHM being of particular interest. Two motivations appear in the literature for using the PHM with cost data. The first, highlighted by Lipscomb et al. (1998), is that the PHM relaxes assumptions about the specific error term distribution, offering more flexibility than purely parametric models. The second, discussed in Dudley et al. (1993), Fenn et al. (1995), & Fenn et al. (1996) is that censoring issues may need to be addressed if observations are terminated early by either the end of data collection or by a competing process (such as death). While the first motivation has not generated much controversy, criticism for the second has founded a good deal of current research. Standard survival techniques are not applicable to "censored" cost data because the censoring is informative, (those who are lost to follow up will tend to have smaller expenditures.) Lin et al. (1997) and Etzioni et al. (1999) give in-depth discussions on the issue of induced informative censoring in survival cost models and Jain and Strawderman (2002) give an inventory of published work concerning solutions (as well as their own). The SEER Medicare colorectal cancer dataset we analyze contains a full year of expenditure observations for all subjects.

Comparison Papers: Assorted articles have been devoted to comparing methods for health cost analysis. The following provide discussions of the problems inherent in expenditure modelling, as well as careful considerations into when and where their suggestions may be appropriate. We present their results only briefly here, to outline the separate and various findings. Dudley et al. (1993) considered linear regression, lognormal regression, logistic regression, a Cox PHM, and a Weibull survival model applied to coronary artery bypass graft surgery costs. The authors advocated the Cox PHM, but concluded "we are unable to determine unequivocally which method of analysis is 'best' for analyzing the importance of clinical factors upon cost". Using Medicare ischemic stroke costs, Lipscomb et al. (1998) compared 1- & 2-stage linear regression, 1- & 2-stage lognormal regression (with and without a smearing adjustment), and a Cox PHM. Although their focus was on how to assess candidate models rather than on the actual candidate models per se, the authors do state "For deriving the predictive distribution of cost, the log-transformed two-part and proportional-hazards models are superior. For deriving predicted mean or median cost, these two models and the commonly used log-transformed linear model all perform about the same." Diehr et al. (1999) evaluate 1- & 2-stage linear regression, 1- & 2-stage lognormal regression (with and without a smearing adjustment), and 1- & 2-stage Gamma (GLM) regression with data from the Washington State Basic Health Plan. The authors recommend the two-stage lognormal model for "understanding the system", the one-stage Gamma model for "understanding the effect of individual covariates on total costs", and one-stage linear regression for "prediction of future costs".

To say a consensus has not been reached would not be entirely fair, since each of the previous comparison papers rightly state that choice of statistical method should depend on the scientific goals and specific data in question. However, the diversity of comparisons appraised thus far and the lack of consistency in results/recommendations calls for further discussion. In addition, previous comparison papers base their model discussions and recommendations on a single set of chosen explanatory variables, with little emphasis on how the explanatory variable set was chosen. Our discourse highlights the interplay between variable selection and model choice.

Giov Sugg. - we should also point out that there may be specificas of cancer data that have not been addressed in the literature (Joe and Arnie may help here)

1.3. Goals of paper and statistical challenges

The goal of this paper is to explore the potential of a variety of techniques to reliably address the obstacles in cost estimation. Specifically, we examine variations in costs and patterns of resource use for various demographic configurations in the last year of life of colorectal cancer patients. We use this application as a springboard for comparing methodological approaches, and for tailoring statistical methods to the specific needs of cancer research questions. Our investigations focus on methods for regression modelling, variable selection, and cost prediction with an assortment of error distributions. We additionally demonstrate a hybrid cross-validation/bootstrap method for evaluating competing models. Models examined include one- and two-stage Gaussian regression modelling, one- and two-stage LogNormal modelling (with and without smearing adjustments), one- and two-stage Gamma regression modelling, and proportional hazard modelling.

Statistical challenges include characterizing similarities and differences between different classes of models, demonstrating approaches for adapting models to specific contexts, illustrating techniques for evaluating a set of contending models, and incorporating covariate profile assessments into such evaluations.

2. Data Description

We use colorectal cancer cost data from the Surveillance, Epidemiology and End Results (SEER)-Medicare-linked database as the motivating example throughout our methodologic discussions. Warren et al. (2002) describe in detail the overall materials and methods used in constructing the SEER-Medicare database and Brown et al. (2002) discusses various descriptive cost estimation techniques and limitations with a focus on colorectal and breast cancers. Our subset of costs for colorectal cancer patients is outlined below and is similar to those in Brown et al. (1999) and Etzioni et al. (2001).

Patient Population & Sample: Patients diagnosed with colon or rectal cancer as their first primary cancer comprise our population of colorectal cancer patients. Additionally, patients selected for analysis were entitled to both Medicare part A (inpatient hospital, skilled nursing, home and hospice care) & Medicare part B (physician services, outpatient care, and medical equipment) payments sometime during the calender period Jan. 1986 - Dec. 1998 and had a full year of reimbursement data observed for the 12 months preceding death. In all, 44006 patients, (23101 females and 20905 males) were examined. SEER registries and their contributions included the states of San Francisco (4330), Connecticut(7445), Michigan(7311), Hawaii(1020), Iowa(7324), New Mexico(1508), Seattle(4617), Utah(1498), Georgia/Rural(2457), San Jose(1434), and Los Angeles(5062).

Dependent Variable: We take on a "Medicare" or "Governmental" perspective (payments made for colorectal cancer patients) rather than a societal one (additional cost burdens due to colorectal cancer) and adopt total Medicare payments (the sum of both A & B Medicare reimbursements) in the terminal disease phase (the patient-specific 12 month period preceding death) as our analysis variable. Terminal phase costs are particularly relevant to concerns on end-of-life care in the health care system. As in previous studies on this database, costs (reimbursements) were adjusted to 1994 constant dollars using time and geographical adjustment factors from the Centers for Medicare and Medicaid Services (CMS), (previously the Health Care Financing Administration (HCFA)). Figure 1 shows the marginal distribution of costs in the last year of life for colorectal cancer patients, as well as marginal fits for each of the statistical models described below. Note in particular the substantial concentrations of small expenditures in the left tail, and the extreme skewness towards large expenditures in the right.

Explanatory Variables: Our substantive concerns focus primarily on the effects of Gender and Ethnicity on costs in the last year of life. We additionally investigate the effects of age, time from diagnosis till death; Socio-Economic factors such as marital status, median income, and percent of high school graduates, (the last two measured at the census tract level); severity factors such as cancer stage at initial diagnosis, number of other primary cancers diagnosed within time from diagnosis to death and whether the patient died from cancer. Specific confounder/covariate subsets (covariate profiles) examined in this article are described below:

1. "Basic Profile": The basic covariates of interest are included as linear effects in this profile and simple regression terms were added for each of the following explanatory variables: gender, ethnicity, age at death, basic geographic location (SEER registry), cancer stage, number of distinct tumors diagnosed in the patient, an indicator for cancer being the cause of death, the number of months from diagnosis to death, census tract median income, census tract high school graduation percent and marital status.

2. "Full Profile": Included in this profile are all of the basic covariates of interest discussed above, as well as all other explanatory variables considered in our study, including: interactions for gender by ethnicity and gender by age at death, linear spline terms for Diagnosis to Death

Months larger than 1 and 2 years, linear spline terms for Census Tract Median Income larger than $12,000 and $20,000, and linear spline terms for Census Tract highschool graduation rates being larger than 35%. (The spline terms were added to account for nonlinear continuous covariate effects observed in the training sample.)

3. "Significance Profile": Covariates from the "Full Model" that were *statistically* significant at the $\alpha = 0{\cdot}05$ level in one or more of the cost models were included in this profile. In many articles, this is the only covariate profile presented.

4. "Significance, No Income profile": This profile is the same as Profile 3 without the census tract median income variables. Results in the training sample from the PHM fluctuated substantially under covariate profile 3. Results were more stable when this variable was dropped.

5. "Geographically adjusted Gender, Ethnicity and Age Profile": This profile examines the main explanatory variables of interest, after adjusting for basic geographic information (as represented by SEER registry) and age effects.

6. "Gender*Ethnicity Profile": This profile contains indicator variables for Gender="Male", Ethnicity="Black", Ethnicity="Other" and the interactions of these indicators. White Females are thus the baseline comparison group in this profile, and combinations of the 5 indicators examine total differences between the 6 Gender and Ethnicity Groups, unadjusted for any other effects.

Table 1 in Appendix I lists the explanatory variables included in the 6 covariate profiles. A limitation of the SEER-Medicare database is that it contains only fee-for-service (FFS) claims data, which do not capture all sources of rendered medical services. There are no Medicare claims when a beneficiary receives services covered by, but not billed to medicare; thus costs of non-FFS services are not present in the database. Similarly, costs for HMO enrollees are not available since HMOs have historically not been required to submit specific service claims. For an in-depth discussion on these and other SEER-Medicare database limitations, see Warren et al. (2002).

3. Statistical Methods

3.1. General Overview

We used a dual cross-validation framework, where 10% of the data was set aside for external validation and a "purist" attitude was adopted in locking this data away until all aspects of modelling (and complications therein) had been resolved. Within the remaining 90% training sample, we applied k-fold cross-validation with k=10%, to answer all the questions one must confront in any data analysis, including variable selection, nonlinear covariate function investigation, knot placements for smoothing, mixture model cutpoint values, etc. Models were therefore initially built up on 81% of the full dataset, with k-fold cross-validation evaluations made on 9% of the full dataset until we felt comfortable with our modelling decisions. The final models were then applied to the complete 90% training sample to obtain final parameter estimates. The 10% external validation sample was then 'unlocked' and used to evaluate the parameters and predictions from our final models. A visual representation of this dual cross-validation is available on the project website and is worth approximately 5 times as much as the preceding paragraph, applying the classical 1 picture = 1000 words formula.

3.2. Model Descriptions

Suppose we are interested in predicting costs at the individual level, using a set of covariates (explanatory variables), X_i, which are believed to capture important differences in the cost distributions. A systematic component for the prediction, linear in the parameters but possibly nonlinear in the explanatory variables, may be defined for each subject as:

$$\eta_i = \eta(\beta, X_i) = \beta_0 + \sum_j \beta_j \cdot h_j(x_{ij})$$

where the β parameters are recognized to be model-specific. We wish to determine how changes in the covariates alter the conditional distribution of costs $F(c|X_i) = Pr(C \leq c|X_i)$ particularly through the conditional cost expectation which we shall use for prediction, $\hat{C}_i = E(C|X_i)$. Medians (vs. means) are also commonly used to characterize the central tendencies of distributions but we focus on average expenditures for point estimates and do not consider median prediction in this article, (noting that the health care service that bases their budget on median cost predictions may soon be out of business.)

Normal Regression:
Universally applied in research, the familiar Gaussian regression model specifies:

$$C_i = \eta_i + \epsilon_i,$$

where C_i denotes the i^{th} subject's cost, η_i denotes the systematic component determined by the i^{th} subject's covariate profile, and ϵ_i is a random error component following a Gaussian distribution with mean 0 and constant variance σ^2. The predicted mean cost for this model is simply the estimate of the systematic component, calculated via ordinary least squares or maximum likelihood:

$$\hat{C}_i^N = \hat{\eta}_i,$$

and the Cumulative Distribution Function, (CDF), for costs is simply:

$$F^N(c|\hat{\eta}_i, \hat{\sigma}) = \Phi\{[c - \hat{\eta}_i]/\hat{\sigma}\},$$

where Φ is the CDF of the standard normal distribution.

2-Stage Normal Regression:
Two-stage models draw on the idea that $E(C) = E(C|C > 0) \cdot Pr(C > 0)$ to address large concentrations of zero cost expenditures. The complete specification is then a mixture of two models, the first describing the probability of having any expenditure, and the second describing the average cost among those having positive expenditures. A model is called separable if $\Theta = (\Psi, \Lambda)$ represents parameters describing the complete distribution of costs and the factorization applies: $f(C|\Theta) = f(C|P, \Psi) \cdot f(P|\Lambda)$, where $P = I(C > 0)$ is an indicator function denoting positive or zero cost. Note that the Tobit model is not separable. A standard two-part separable mixture model, with a gaussian distribution for positive costs, can be characterized by:

$$C_i = \begin{cases} 0 & \text{with probability } (1 - p_i) \\ \eta_i + \epsilon_i & \text{with probability } p_i \end{cases}$$

where $\epsilon_i \sim N(0, \sigma^2)$. The probability of a positive expenditure, p_i, is usually modelled with either probit or logistic regression. Opting for the latter, we would specify:

$$Pr(C_i > 0) = p_i = \{1 + exp(-\zeta_i)\}^{-1}$$

(or equivalently, $logit(p_i) = \zeta_i$), where ζ_i is the analogous systematic component for predicting a positive cost:

$$\zeta_i = \gamma_0 + \sum_j \gamma_j \cdot h_j^*(x_{ij})$$

The h^* notation simply denotes that the functions of the explanatory variables (as well as the set of explanatory variables themselves) in ζ_i may be different than in η_i. Under this model, the predictor for an individual's expected cost would be:

$$\begin{aligned} \hat{C}_i &= E(C|\hat{\eta}_i, \hat{\sigma}, \hat{\zeta}_i) \\ &= E(C|C > 0, \hat{\eta}_i, \hat{\sigma}^2) \cdot Pr(C > 0|\hat{\zeta}_i) \\ &= \hat{\eta}_i \cdot \{1 + exp(-\hat{\zeta}_i)\}^{-1} \end{aligned}$$

Mixture models are useful in a variety of ways, allowing us to adapt error distributions to fit the data more closely, and aiding in understanding separate aspects of the expenditure system. These models can easily be made more flexible, (see for example Manning et al. (1987) for a four-part model); one extension in particular allows a non-zero cutpoint, τ, to be defined for splitting the two-stages: $E(C) = E(C|C \leq \tau) \cdot Pr(C \leq \tau) + E(C|C > \tau) \cdot Pr(C > \tau)$. Under such a formulation, we require the additional specification of a model for costs when they are *below* the cutpoint, as well as the usual two models for (1.) the distribution of costs when they are *above* the cutpoint and (2.) the probability of having a cost above the cutpoint. One goal of this expanded two-stage model is to balance parsimony and interpretability with fitting the error distribution in a more complicated, but hopefully tighter manner. Our intention was to use the two-part model to address the concentrations of "minimal" costs, (measured in dollars), in the left tail of Figure 1, and it was found that the simple specification with $\tau = \$0$ left an unsatisfying "lump" of minimal costs remaining. We therefore used our 90% training sample to investigate alternate values for τ that retained model interpretability, but accounted for the observed "minimal" cost concentrations. For the purposes of demonstrating how to adapt models to specific contexts, we use $\tau = \$1000$, and a Uniform(0,$1000) distribution for the "minimal" costs, arguing both that in the last year of life for colorectal cancer patients, $1000 could still be considered a "minimal" cost, and that costs below $1000 were not found to vary substantially between demographic groups of interest. The simple uniform specification for the minimal cost also allows explanatory variable parameter interpretations for expenditure size to focus on the cost group with larger expenditures. More complicated mixture models can be fit using a variety of finite mixture specifications for the lower expenditures, but our aims are to illustrate the technique and we do not compare additional lower expenditure mixture specifications in this manuscript.

As discussed, our two-part Normal mixture model may be specified as:

$$C_i = \begin{cases} \text{Uniform}(0, 1000) & \text{for } C_i \leq 1000 \quad \text{with probability } (1 - p_i) \\ \text{Normal}(\eta_i, \sigma^2) & \text{for } C_i > 1000 \quad \text{with probability } p_i \end{cases}$$

Where:

$$Pr(C_i > 1000) = p_i = \{1 + exp(-\zeta_i)\}^{-1}$$

and ζ_i is the systematic component for predicting a "large" cost as above. The predictor for an individual's expected cost is:

$$\begin{aligned} \hat{C}_i^{N2} &= E(C|\hat{\eta}_i, \hat{\sigma}^2, \hat{\zeta}_i) \\ &= E(C|C \leq 1000) \cdot Pr(C \leq 1000|\hat{\zeta}_i) + E(C|C > 1000, \hat{\eta}_i, \hat{\sigma}^2) \cdot Pr(C > 1000|\hat{\zeta}_i) \\ &= 500 \cdot \{1 + exp(\hat{\zeta_i})\}^{-1} + \hat{\eta}_i \cdot \{1 + exp(-\hat{\zeta_i})\}^{-1} \end{aligned}$$

and the CDF for costs under the two-stage Normal model is:

$$F^{N2}(c|\hat{\eta}_i, \hat{\sigma}, \hat{\zeta}_i) = \left[\frac{(c/1000)^{I(c\leq 1000)}}{1+exp(\hat{\zeta}_i)}\right] \cdot I(c \geq 0) + \left[\frac{\Phi\{[c-\hat{\eta}_i]/\hat{\sigma}\}}{1+exp(-\hat{\zeta}_i)}\right] \cdot I(c > 1000)$$

where $I(\textit{condition})$ is a 0/1 indicator function for the condition.

Lognormal Regression (with and without smearing):
A simple lognormal regression specifies:

$$log(C_i + 1) = \eta_i + \epsilon_i,$$

where ϵ_i again represents a random error component following a Gaussian distribution with mean 0 and constant variance σ^2. The addition of \$1 to the original cost sets the log-transformed cost to zero when the original cost is zero. If costs are log-normally distributed, then the average cost on the untransformed scale is a function of both the mean and the variance on the transformed scale, and the predicted mean cost on the original scale is:

$$\hat{C}_i^{LN} = exp(\hat{\eta}_i + \hat{\sigma}^2/2) - 1$$

As noted above, this estimate can be severely biased if costs do not follow a lognormal distribution, or if the variability is heterogeneous between groups. The smearing adjustment, (Duan 1983), essentially replaces the variability estimate above with a nonparameteric average-retransformed-residual-error estimate, and the predicted mean cost using a single smearing coefficient would be:

$$\hat{C}_i = exp(\hat{\eta}_i + S) - 1$$

where $S = log[\ \sum_i exp(e_i)/N\ ]$ and $e_i = log(C_i+1) - \hat{\eta}_i$ is the residual from the linear regression of the log-transformed expenditures. Manning (1998) relates that even a simple smearing adjustment may not be adequate if error terms are heteroscedastic between covariate groupings, and we have constructed separate smearing coefficients for the six different gender/ethnicity combinations, (Gender (G) = Male,Female; Ethnicity (E) = White,Black,Other). Our predicted mean cost under the smeared lognormal model is thus:

$$\hat{C}_i^{LNS} = exp(\hat{\eta}_i + S_{G,E}) - 1$$

where $S_{G,E}$ denotes a gender/ethnicity specific smearing coefficient.

The smearing technique is an adjustment for estimating a mean, and does not affect other distributional aspects, (quantiles, etc.), and whether or not the smearing adjustment is used the Cumulative Distribution Function, (CDF), for costs under the lognormal model is:

$$F^{LN}(c|\hat{\eta}_i, \hat{\sigma}) = \Phi\{[log(c+1) - \hat{\eta}_i]/\hat{\sigma}\}$$

2-Stage Lognormal Model (with and without smearing):
We exploit the same ideas discussed in the Normal mixture model above for a lognormal version of the two-stage model. The only adjustment needed is to employ a lognormal model for the large expenditures instead of the Normal. Our lognormal mixture model is thus:

$$C_i = \begin{cases} \text{Uniform}(0, 1000) & \text{for } C_i \leq 1000 \quad \text{with probability } (1-p_i) \\ \text{Lognormal}(\eta_i, \sigma^2) & \text{for } C_i > 1000 \quad \text{with probability } p_i \end{cases},$$

where p_i is identical to that for the Normal mixture. Under this model, the predictor for an individual's expected cost is:

$$\begin{aligned}\hat{C}_i^{LN2} &= E(C|\hat{\eta}_i, \hat{\sigma}^2, \hat{\zeta}_i) \\ &= E(C|C \leq 1000) \cdot Pr(C \leq 1000|\hat{\zeta}_i) + E(C|C > 1000, \hat{\eta}_i, \hat{\sigma}^2) \cdot Pr(C > 1000|\hat{\zeta}_i) \\ &= 500 \cdot \{1 + exp(\hat{\zeta}_i)\}^{-1} + \{exp(\hat{\eta}_i + \hat{\sigma}^2/2) - 1\} \cdot \{1 + exp(-\hat{\zeta}_i)\}^{-1}\end{aligned}$$

The lognormal part in this two-stage model is subject to the same deficiencies in estimating the mean as the one-stage lognormal model and we employ smearing adjustments as before. The predicted mean cost under the two-stage lognormal model with smearing is then:

$$\hat{C}_i^{LN2S} = 500 \cdot \{1 + exp(\hat{\zeta}_i)\}^{-1} + \{exp(\hat{\eta}_i + S_{G,E}) - 1\} \cdot \{1 + exp(-\hat{\zeta}_i)\}^{-1}$$

where the gender and ethnicity specific smearing coefficients, $S_{G,E}$, are calculated from the lognormal model residuals for those with positive expenditures. Smearing again does not affect quantile or distribution estimates, and whether or not the smearing adjustment is used the CDF for costs under the two-stage lognormal model is:

$$F^{LN2}(c|\hat{\eta}_i, \hat{\sigma}, \hat{\zeta}_i) = \left[\frac{(c/1000)^{I(c \leq 1000)}}{1 + exp(\hat{\zeta}_i)}\right] \cdot I(c \geq 0) + \left[\frac{\Phi\{[log(c) - \hat{\eta}_i]/\hat{\sigma}\}}{1 + exp(-\hat{\zeta}_i)}\right] \cdot I(c > 1000)$$

Gamma Model

We adopt a Gamma distribution and log-link for our basic generalized linear model specification for the reasons listed above and because it approximated both the marginal and Gender/Ethnicity conditional densities adequately. For the Gamma distribution, and indeed under any generalized linear model utilizing a log-link, the expected cost is simply:

$$\hat{C}_i^G = exp(\hat{\eta}_i)$$

The corresponding estimated Gamma CDF is:

$$F^G(c|\hat{\eta}_i, \hat{\alpha}) = \frac{\Gamma_c(\hat{\alpha}, \hat{\eta}_i)}{\Gamma(\hat{\alpha})\big(exp(\hat{\eta}_i)\big)^{\hat{\alpha}}}$$

where $\hat{\alpha}$ is the estimate of α, the dispersion parameter for the Gamma distribution, $\Gamma(\cdot)$ is the gamma function, and $\Gamma_c(\hat{\alpha}, \hat{\eta}_i)$ is the incomplete gamma function $\Gamma_c(\hat{\alpha}, \hat{\eta}_i) = \int_0^c \frac{(t\hat{\alpha})^{\hat{\alpha}}}{t} \cdot exp\left(\frac{-t\hat{\alpha}}{exp(\hat{\eta}_i)}\right) dt$.

2-Stage Gamma Model

Similar to the Normal and Lognormal mixture models above, the Gamma version of the two-stage model specifies:

$$C_i = \begin{cases} \text{Uniform}(0, 1000) & \text{for } C_i \leq 1000 \quad \text{with probability } (1 - p_i) \\ \text{Gamma}(\eta_i, \alpha) & \text{for } C_i > 1000 \quad \text{with probability } p_i \end{cases},$$

where p_i is again the probability of an expenditure being above \$1000. The predicted mean cost under the two-stage gamma model is:

$$\hat{C}_i^{G2} = \{exp(\hat{\eta}_i)\} \cdot \{1 + exp(-\hat{\zeta}_i)\}^{-1} + 500 \cdot \{1 + exp(\hat{\zeta}_i)\}^{-1}$$

and the corresponding CDF is:

$$F^{G2}(c|\hat{\eta}_i,\hat{\alpha}) = \left[\frac{(c/1000)^{I(c<1000)}}{1+exp(\hat{\zeta}_i)}\right]\cdot I(c\geq 0) + \left[\frac{\Gamma_c(\hat{\alpha},\hat{\eta}_i)}{\Gamma(\hat{\alpha})\big(exp(\hat{\eta}_i)\big)^{\hat{\alpha}}}\cdot\frac{1}{\{1+exp(-\hat{\zeta}_i)\}}\right]\cdot I(c>1000)$$

Cox PHM

The Cox Proportional Hazards model is a semiparametric model characterized by the "survival" function:

$$Pr(C_i\geq c|\eta_i) = 1-F(c|\eta_i) = S(c|\eta_i) = S_0(c)^{exp(\eta_i)}$$

where $S_0(c)$ is the "baseline" survival cost function (all covariates set to zero). Since the expected value of a random variable is equal to the integral of its survival function, the predicted mean cost under the Cox Proportional Hazards Model is:

$$\hat{C}_i^{PHM} = \int_0^\infty \hat{S}_0(c)^{exp(\hat{\eta}_i)}dc$$

The Cox PHM is appealing for cost research because one is able to estimate the β coefficients for use in $\hat{\eta}_i$ without specifying the survival function for expenditures (i.e. the underlying cost distribution) which, as discussed, is a primary difficulty. To obtain an estimate of the baseline survival function, $\hat{S}_0(c)$, we used the standard product-limit estimator (Kalbfleisch and Prentice (1980)).

With estimates for the β coefficients and the baseline survival function in hand, we can calculate $\hat{C}_i^{PHM}$ and determine the CDF for costs under the proportional hazards model with:

$$F^{PHM}(c|\hat{\eta}_i) = 1-S(c|\hat{\eta}_i) = 1-\hat{S}_0(c)^{exp(\hat{\eta}_i)}$$

3.3. Parameter Interpretations:

Substantive research attempts to say something about how the world works, (for cost research, how the world spends), and relies inherently on the underlying model to drive conclusions. To further understand the models presented above and to aid in choosing an appropriate model, we discuss briefly the related parameter interpretations. Suppose there were only one explanatory variable, a linear predictor would take the form $\eta_i = \beta_0+\beta_1 x_i$, ($\zeta_i = \gamma_0+\gamma_1 x_i$ for our probability models), and the parameter β_1 (γ_1) would take on the following interpretations for each of the different models:

Logistic: γ_1 takes on standard log-odds-ratio interpretations. When comparing the dichotomous grouping of those who spent over \$1000 to those who spent \$1000 or less, there is a multiplicative effect of e^{γ_1} on the odds of spending over \$1000 for each unit increase in x.

Normal: β_1 represents the additive change in the average expenditure for each unit increase in x, (standard Gaussian regression interpretations).

Lognormal *(without smearing):* Since $E[log(C_i+1)] = \beta_0+\beta_1 x_i$, β_1 could be taken as the change in the average $log(C_i+1)$ for each unit increase in x. More useful however is to exploit the fact that, under the lognormal model, $E(C_i+1) = e^{\beta_0+\beta_1 x_i}\cdot e^{\cdot 5\sigma^2}$, which shows e^{β_1} to be the multiplicative increase on $E(C_i+1)$ for each unit increase in x. Thus, for $E(C_i+1)\approx E(C_i)$, e^{β_1} has a relative expenditure interpretation.

Lognormal *(with smearing):* With a single smearing coefficient in the lognormal model, $E(C_i+1) = e^{\beta_0+\beta_1 x_i}\cdot e^S$, where S is the smearing coefficient. Hence, e^{β_1} has a relative expenditure interpretation when a single smearing coefficient is considered. With multiple smearing coefficients, interpretations become more complex. Consider the simple case of two smearing groups. When comparing within a group, e^{β_1} retains its relative expenditure meaning but when comparing between

groups, a one unit increase in x is associated with an increase of $\frac{e^{\beta_0+\beta_1(x_i+1)}\cdot e^{S_1}}{e^{\beta_0+\beta_1 x_i}\cdot e^{S_0}} = e^{\beta_1}\cdot\left(\frac{e^{S_1}}{e^{S_0}}\right)$ and the effect is modified by a multiplicative constant. Recalling that the smeared lognormal model incorporates six separate smearing coefficients, interpretations become complex.

Gamma: One of the benefits of using a Generalized Linear Model is the ease of interpretation after applying the inverse link function. Since we used a log-link with our Gamma GLM, we exponentiate and find that e^{β_1} has a simple relative expenditure interpretation.

PHM: Parameters in the Proportional Hazards Model take on log-hazard-ratio interpretations. A one unit increase in x is associated with a multiplicative increase of e^{β_1} in the expenditure hazard (the probability of having a certain expenditure C, given the expenditure will be at least C.) Note that the direction of the effects represented by these parameters are reversed compared to the other parameterizations. For example, if β_1 is positive, the hazard increases with increasing x. Thus the probability of spending "exactly C and no more" increases, which in turn *decreases* the average expenditure. Likewise, if β_1 is negative, the average expenditure *increases*.

Two-Stage Models: All the previous interpretations hold, except they are now applied only to expenditures greater than $1000.

3.4. Validation Criteria

We adopt a validation algorithm similar to Lipscomb et al. (1998) to incorporate both model estimation error and individual-level error. Within our training dataset we randomly select $B = 100$ bootstrap samples and apply each covariate profile to each of our models within each of the bootstrap samples. The final cost prediction for each profile and model combination is then the average of the 100 predictions constructed from the parameters found in the 100 bootstrap training samples. Bootstrap estimates for the β parameters and their standard errors are similarly the mean and standard deviation of the 100 corresponding $\hat{\beta}$'s, while 95% confidence intervals for the β's are found non-parametrically from the 0.025 and 0.975 percentiles of the 100 $\hat{\beta}$'s. The measures we investigate for evaluating the quality of the preceding models are described below. Each measure attempts to capture a different aspect of model "quality", and is calculated for all model and covariate profile combinations.

- BIAS: The bias is computed as

$$BIAS = \frac{1}{n}\sum_i (C_i - \hat{C}_i)$$

  and provides information on the calibration of the prediction, (i.e. whether the predicted cost is centered around the true cost on average.)

- RMSE: The Root Mean Squared Error is computed as

$$RMSE = \sqrt{\frac{1}{n}\sum_i (C_i - \hat{C}_i)^2}$$

  and is a combination of the calibration (bias) and the precision (variability) of the prediction. Many estimators attempt to trade off a small increase in bias for a large decrease in variability, thus improving their overall RMSE measure.

- MAE: The Mean Absolute Error is computed as

$$MAE = \frac{1}{n}\sum_i \left|C_i - \hat{C}_i\right|$$

Similar to the RMSE, the MAE is a penalty measure of the distance between the predicted cost and the actual cost. While the RMSE penalizes in a quadratic fashion, however, the MAE penalizes in a linear fashion. The contributions to these two error measures for predictions that were, say, \$1, \$10, & \$100 away from their true costs would be \$1, \$100, & \$10,000 respectively for the RMSE, while only \$1, \$10, & \$100 for the MAE. The RMSE therefore penalizes large errors more severely than the MAE.

- LS-rule: The Logarithmic-Scoring rule is computed as:

$$LS = \frac{1}{n} \sum_i -log\big[\hat{f}(C_i)\big]$$

Where $f(C_i)$ is the density corresponding to the CDF for each model and covariate profile combination specified above.

While the BIAS, RMSE & MAE measures concentrate on evaluating the point predictions of cost, the LS-rule evaluates how well the actual cost is represented in the predictive cost distribution. Observed costs with high probabilities of occurring under the predictive distributions contribute small penalties to the LS-rule, while observed costs that lie in the tails of the predicted distributions, (have small probabilities of occurring when calculated under the estimated model parameters) contribute large penalties to the LS-rule. As with the RMSE and MAE measures, lower LS-rule values indicate better predictions. For calculation of the LS-rule, the normal, lognormal and gamma densities have closed forms corresponding to the CDF's given previously. To find a representation of the density under the PHM, Lipscomb et al. (1998) estimated the probability of the cost C_i falling in a \$1 "bucket" with $f^{PHM}(C) = F^{PHM}(C) - F^{PHM}(C-1)$. We construct our estimator of $f^{PHM}(C_i)$ by noting that under the PHM $S(c|\eta_i) = S_0(c)^{exp(\eta_i)}$, and thus:

$$\begin{aligned} f(c|\eta_i) &= -\tfrac{\partial S(c|\eta_i)}{\partial c} \\ &= exp(\eta_i) \cdot S_0(c)^{\{exp(\eta_i)-1\}} \cdot \tfrac{-\partial S_0(c)}{\partial c} \\ &= exp(\eta_i) \cdot S_0(c)^{\{exp(\eta_i)-1\}} \cdot f_0(c) \end{aligned}$$

Having obtained estimates of $S_0(c)$ and η_i as previously described, all that is further required is an estimate of the baseline density function, $f_0(c)$, which is simply the (negative) derivative of the baseline survival function. With our non-parametric estimate $\hat{S}_0(c)$ and knowledge that the true $S_0(c)$ is relatively well behaved, (a non-increasing function from 1 to 0), we smooth $\hat{S}_0(c)$ with a constrained monotonic B-spline smoother with λ degrees of freedom, and estimate $f_0(c)$ by noting that the derivative of a B-spline of degree r is simply a corresponding B-spline of degree $(r-1)$. An estimate of the Cox PHM density is thus:

$$\hat{f}^{PHM}(C_i) = exp(\hat{\eta}_i) \cdot \hat{S}_0(c;\lambda)^{\{exp(\hat{\eta}_i)-1\}} \cdot \hat{f}_0(c;\lambda)$$

We used cross-validation within the training sample to resolve standard smoothing issues such as the behavior of the estimate with respect to the choice of λ. Within our bootstrapped validation framework, we take the final estimate for the density at C_i to be the average $\hat{f}^{PHM}(C_i)$, calculated over the 100 bootstrap samples.

We adopt the framework of Lipscomb et al. (1998) to account for individual-level variability in the validation sample and bootstrap the cross-validation statistics $M = 100$ times within the validation sample. Thus, each validation measure is the average of 100 corresponding bootstrapped measures, calculated from randomly sampling the residuals in the 10% pure validation dataset with replacement.

4. Results

4.1. Substantive Results:

Our substantive questions center on expenditure differences for the six Gender and Ethnicity subgroups defined earlier; this section focuses on the related five β parameters that characterize these differences.

Similarities between models: Figure 2 displays parameter estimates and bootstrapped 95% confidence intervals for Gender/Ethnicity differences (White Females are the reference group) over the six covariate profiles described in section 2, for each of the models described above. The 'adjusted' covariate profiles that include Gender/Ethnicity interaction terms (profiles 2 through 5) demonstrate that in a general sense, white males have the least expenditures, followed by black males, while white females, females of other ethnicities and males of other ethnicities had similar expenditures and black females had the highest estimated expenditures overall. Examining the estimates over a variety of models and over a variety of confounder adjustments lends credibility to these statements. An important question then is whether these observed Gender/Ethnicity differences may be policy related *(Joe & Arnie?)*. The unadjusted parameters (profile 6) tell a somewhat different story, where white males and females of other ethnicities spend more than white females. Since these relationships disappear when adjusting for age and geographic area (profile 5), we know that area and age are modifiers of the effects of Gender/Ethnicity. Thus the adjusted models are more appropriate for discussing overall Gender/Ethnicity expenditure differences. We note that white females were the oldest group with an average age of death of 83 years, while white males had an average age of death of 80 years and females of other ethnicities had an average age of death of 81 years. Covariate profile 1 is the basic profile without interactions and does not distinguish between white, black and other ethnicity males, or white, black and other ethnicity females. The "Male" estimate thus averages over the ethnicity subgroups and yields fairly different results (intuitively so) than the other 'adjusted' covariate profiles; likewise the ethnicity estimates for the "Black" and "Other" parameters average over gender. The Gender/Ethnicity interaction terms are therefore important to include.

Differences between models: The overall patterns in estimates and confidence intervals are similar for all models save the simple logNormal model, which deviates in two ways. First, the estimates differ from what we might expect by examining the other models. White male expenditures are all substantially lower than estimates from the Gamma and Proportional Hazards models, the estimates for black females are relatively higher for the 'Full' and 'Significance' profiles and relatively much lower for the 'Basic' profile. Second, and more importantly, the bootstrap confidence intervals are all much wider compared to those for the Gamma and Proportional Hazards models. Since these models describe roughly the same expenditure relationships (multiplicative increases in expenditures), we would hope to have roughly similar estimates and inferences across these models. Figure 1 gives us some insight into why the simple lognormal may behave differently than the other models; to accommodate both the spike of small expenditures in the left tail and the extreme expenditures in the right tail, the lognormal model must substantially overestimate lower expenditures (below around \$10,000), and underestimate mid-sized expenditures (from about \$10,000 to \$100,000). Incorporating even the simple 2-stage mixture vastly improves the lognormal fit and we note that the differences between the lognormal and Gamma models disappear when we use the 2-stage modelling technique. Consider if we had used only the simple lognormal model and the simple adjustments of the 'Basic' covariate profile for determining Gender/Ethnicity differences in expenditures; we would incorrectly conclude that white females and black females appear to have similar medicare expenditures. Examining results over a wide variety of error distributions and covariate profiles allows us to make a

more accurate and robust statement about the data supporting black females as a group spending substantially more than white females on colorectal cancer in the last year of life.

4.2. Validation Results

Figure 3 shows predictive ability cross-validation results for each covariate profile and error model investigated. The left column contains results for all models while the right column omits extremely poor models to expand the vertical axes and visualize results from acceptable models. The BIAS measures have been left on their original scale, while the remaining measures are shown relative to the 'Gender*Ethnicity' profile in the 2-Stage lognormal Model for display purposes. Therefore, BIAS measures closer to zero are better, and smaller MAE, RMSE, & LS-rule measures are better. Validation measures on their natural scales can be found on the project website. We obtain 95% C.I.'s for each measure by additionally bootstrapping the 10% validation dataset 100 times. After calculating predicted values for validation-set individuals, (based on the bootstrapped parameter estimates calculated from the 90% training-set), we randomly draw predicted values and residuals from the validation-set with replacement until we have 100 validation datasets to calculate cross-validation results from. 95% confidence intervals for cross-validation measures are then found non-parametrically from the 0.025 and 0.975 percentiles of each measure.

Covariate Profile Results: Generally speaking, the 'Full Covariate' and 'Significant Covariate' profiles (with the most information) performed the best and the 'Gender*Ethnicity' profile (with the least information) performed the worst. There appear to be fairly dramatic improvements in the MAE & RMSE measures between the 'Significance Profile' (profile 3) and the 'Significance, No Income profile' (profile 4). Thus, census tract median income remains an important covariate for cost prediction even after adjusting for variables that may have accounted for its effects, such as geographic region.

Modelling Technique Results: The Gaussian models, the Gamma models and the Cox PHM appear to perform similarly towards point prediction based on BIAS, MAE and RMSE measures for this dataset. The PHM performs the best in terms of predicting the distribution based on the LS-Rule, with substantial gains made over all other models for the simple 'Gender*Ethnicity' profile. The 2-stage Gamma model also performed well on the LS-rule, with similar results to the Cox PHM for all profiles except the 'Gender*Ethnicity' profile. The simple logNormal model performed the worst of any of the examined models on all validation criteria for all covariate profiles. The smearing technique reduced the lognormal BIAS substantially, but even with six separate smearing coefficients (one for each of the Gender*Ethnicity subgroups) the simple smeared lognormal model still considerably over-estimates expenditures. Only by incorporating both smearing and the two-stage technique do we obtain estimates with relatively smaller amounts of BIAS. The normal models performed well in terms of point predictions and poorly in terms of the predictive cost distribution; adding the second stage to the simple normal model did not lead to large gains in any of the measures. Figure 4 displays the observed and predicted expenditures in the validation set for the 'significance profile' over the different error distributions. None of the models appear to predict expenditures accurately; gains in point prediction validation measures appear to be gains made in precision (variability) rather than in calibration (bias). The Normal, Gamma, and Proportional Hazard models all demonstrate considerably less variability than the lognormal models.

5. Discussion

We have used the SEER database and the problem of estimating colorectal cancer costs to motivate our discussion of statistical challenges in health cost research. The amount of data available from this resource makes it possible to investigate commonly applied statistical modelling strategies with a substantial validation set, even when using a "purist" cross-validation approach. We have attempted to provide insights into a standard set of statistical models frequently seen in the literature, while demonstrating ways to develop adaptations of these models to specific contexts. Further, we presented a variety of techniques for evaluating a set of candidate models, including both error distribution and covariate profile assessments into these evaluations.

Substantive Discussion? Joe & Arnie?

General Recommendations: "While statisticians will inevitably speak of robust models and robust procedures...robustness should be defined as a scientists' ability to ferret simple, lasting structures from data." - *Robert Miller: discussion of George Box's 1980 JRSSA paper* - Previous comparison papers have advocated specific statistical models for specific scientific purposes. While we agree that scientific goals should be considered in any statistical analysis, we instead acknowledge Dr. Miller's point above and recommend examining a variety of statistical models whatever the scientific purpose. With current statistical computing resources it is fairly simple to obtain estimates for a wide variety of models and explanatory variables, all addressing a common goal. Substantive conclusions are given additional weight once it is determined that inferences are not driven by the underlying model or covariate selection. Our 'simple, lasting structures' here are presented in the form of similar patterns in substantive results across the various examined models. Though specific parameter interpretations will change between models, Figure 2 shows that the data are telling the same general story about how the world works (spends). One might argue then, that if the story is generally the same, why not just use one model? To this we point out that for our SEER Medicare colorectal cancer dataset, the simple lognormal model performed rather poorly on all counts. Without examining multiple models in a variety of settings, we would never know for the next dataset if we had chosen its 'lognormal' equivalent. Additionally, after reviewing results from all models examined, it becomes easier to make specific recommendations. For this dataset, one might choose the simple normal model if the goal was simple point prediction. However, the Gamma model performed equally well on all point prediction measures and much better in predicting the cost distribution, thus it may be a better choice for simple analyses. Suppose instead that we wished to perform a cost-benefit analysis, which depends on both estimated expenditures and the probability that those expenditures will occur; the Proportional Hazards Model may then be a better choice since it also performs similarly in terms of point prediction and it does a better job of estimating the overall expenditure distribution.

Modelling choices that matter (and those that don't):
Variable Selection: After primary explanatory variables of interest, probable confounders and possible non-linear relationships were determined, we set about the familiar process of ascertaining a 'statistically significant' subset to include in a 'final model'. The usual thinking is to minimize a bias-variance trade off: the more predictors we include, the more variable each of the estimates become, while leaving an important predictor out leads to bias in the estimates. Figures 2 and 3 show that, given enough data, the differences between the 'statistically significant' subset and the 'Full' subset of predictors are minimal, leading us to propose that inclusion of the additional variables in the 'Full' subset may be suitable with data such as these. Using our dual cross-validation approach for variable selection was informative for evaluating covariate profile effects.

Error Distribution: It is clear at this point that the lognormal distribution provides a relatively poor method of analyzing this dataset. All of the alternatives are superior, and even the simple normal distribution is an improvement from a variety of perspectives. Yet the first (and perhaps too-frequently last) impulse of many analysts when confronted with skewed data is to simply take the log. Straightforward solutions such as using a GLM can provide considerable improvements in terms of both point and overall distribution predictions. If the data is actually lognormally distributed then using the lognormal model will of course be the optimal thing to do. However, if the data deviate from the lognormal assumption and the lognormal model is used, then the expected cost (i.e. the "object of interest") is no longer the quantity that maximizes the likelihood function, (i.e. the "object of inference"$= \theta_g = exp\{E_g ln(X)/ + \sigma^2/2\}$ where g is the underlying true distribution, Royall and Tsou (2003)). The PHM performs best for error distribution estimation, but a drawback is that for point prediction you must integrate the survival distribution for each unique covariate profile, which can be computationally prohibitive if there are continuous covariates. Finally, we again stress the benefits of examining multiple models to ensure robust conclusions.

References

Aitchison, J. and Brown, J. (1957), *The Lognormal Distribution*, Cambridge University Press.

Blough, D., Madden, C., and Hornbrook, M. (1995), "Modelling risk using generalized additive models," *Journal of Health Economics*, 14, 521–549.

Blough, D. and Ramsey, S. (2000), "Using Generalized Additive Models to Assess Medical Care Costs," *Health Services & Outcome Research Methodology*, 1, 185–202.

Brown, M., Riley, G., Potosky, A., and Etzioni, R. (1999), "Estimating healthcare costs related to cancer treatment from SEER-Mdicare data," *Medical Care*, 37, 1249–1259.

Brown, M., Riley, G., Schussler, N., and Etzioni, R. (2002), "Estimating healthcare costs related to cancer treatment from SEER-Mdicare data," *Medical Care*, 40(8 Suppl), 104–17.

Cragg, J. (1971), "Some statistical models for limited dependent variables with application to the demand for durable goods," *Econometrica*, 39, 829–844.

Diehr, P., Yanez, D., Ash, A., Hornbrook, M., and Lin, D. (1999), "Methods for analyzing health care utilization and costs," *Annual Review of Public Health*, 20, 125–144.

Duan, N. (1983), "Smearing Estimate: A Nonparametric Retransformation Method," *Journal of the American Statistical Association*, 78, 605–610.

Dudley, R., Jr., F. H., Smith, L., Mark, D., Califf, R., Pryor, D., Glower, D., Lipscomb, J., and Hlatky, M. (1993), "Comparison of Analytic Models for Estimating the Effect of Clinical Factors on the Cost of Coronary Artery Bypass Graft Surgery," *Journal of Clinical Epidemiology*, 46, 261–271.

Etzioni, R., Feuer, E., Sullivan, S., Lin, D., Hu, C., and Ramsey, S. (1999), "On the use of survival analysis techniques to estimate medical care costs," *Journal of Health Economics*, 18, 365–380.

Etzioni, R., Ramsey, S., Berry, K., and Brown, M. (2001), "The impact of including future medical care cost when estimating the costs attributable to a disease: a colorectal cancer case study," *Health Economics*, 10, 245–256.

Fenn, P., McGuire, A., Backhouse, M., and Jones, D. (1996), "Modelling programme costs in economic evaluation," *Journal of Health Economics*, 15, 115–125.

Fenn, P., McGuire, A., Phillips, V., Backhouse, M., and Jones, D. (1995), "The analysis of censored treatment cost data in economic evaluation," *Journal of the American Statistical Association*, 33, 851–863.

Jain, A. and Strawderman, R. (2002), "Flexible hazard regression modeling for medical cost data," *Biostatistics*, 3, 101–118.

Kalbfleisch, J. and Prentice, R. (1980), *The Statistical Analysis of Failure Time Data*, John Wiley & Sons, Inc.

Lin, D., Feuer, E., Etzioni, R., and Wax, Y. (1997), "Estimating Medical Costs from Incomplete Follow-Up Data," *Biometrics*, 53, 419–434.

Lipscomb, J., Ancukiewicz, M., Parmigiani, G., Hasselblad, V., Samsa, G., and Matchar, D. (1998), "Predicting the Cost of Illness: A comparison of Alternative Models Applied to Stroke," *Medical Decision Making*, 18 suppl, S39–S56.

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

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

Manning, W., Newhouse, J., Duan, N., Keebler, E., Leibowitz, A., and Marquis, M. S. (1987), "Health Insurance and the Demand for Medical Care: Evidence from a Randomized Experiment," *The American Economic Review*, 77, 251–277.

McCullagh, P. and Nelder, J. (1989), *Generalized Linear Models, 2nd ed.*, Chapman & Hall.

Mullahy, J. (1998), "Much ado about two:reconsidering retransformation and the two-part model in health economics," *Journal of Health Economics*, 17, 247–281.

Royall, R. and Tsou, T.-S. (2003), "Interpreting Statistical Evidence Using Imperfect Models: Robust Adjusted Likelihood Functions," *Journal of the Royal Statistical Society Series B*, 65, 391–404.

Tobin, J. (1958), "Estimation of Relationships for Limited Dependent Variables," *Econometrica*, 25, 24–36.

Warren, J., Klabunde, C., Schrag, D., Bach, P., and Riley, G. (2002), "Overview of the SEER-Medicare data: content, research applications, and generalizability to the United States elderly population," *Medical Care*, 40(8 Suppl), 3–18.

6. Appendix I

			Covariate Profile					
Covariate	Type	Range	1	2	3	4	5	6
Gender	ind	$1 = Male$	X	X	X	X	X	X
Ethnicity	cat	$White, Black, Other$	X	X	X	X	X	X
Gender $\times$ Ethnicity	int			X	X	X	X	X
Age at Death	cont	$66 - 110$	X	X	X	X	X	
Gender $\times$ Age at Death	int			X	X	X	X	
Geography	cat	$10\ SEER\ Registries$[1]	X	X	X	X	X	
Cancer Stage	cat	$1, 2, 3, 4, 5, missing$	X	X	X	X		
Number of Cancers	cat	$1, 2, \geq 3$	X	X	X	X		
Cause of Death	ind	$1 = Cancer$	X	X	X	X		
Months from Diagn. to Death[2]	cont	$0 - 296$	X	X	X	X		
Census Tract Median Income[3]	cont	$1,875 - 171,107$	X	X	X			
Census Tract % HS Grads[4]	cont	$\cdot 11 - \cdot 99$	X	X				
Marital Status	ind	$1 = Married$	X	X				

[1] Type: ind=Indicator Variable, cat=Categorical Variable, cont=Continuous Variable, int=Interaction term(s).

[2] linear spline terms added for Diagnosis to Death Months $>$ 1 year and $>$ 2 years in profiles 2,3,4.

[3] linear spline terms added for Census Tract Median Income $>$ \$12,000 and $>$ \$20,000 in profiles 2,3.

[4] linear spline term added for Census Tract highschool graduation rate $>$ 35% in profile 2.

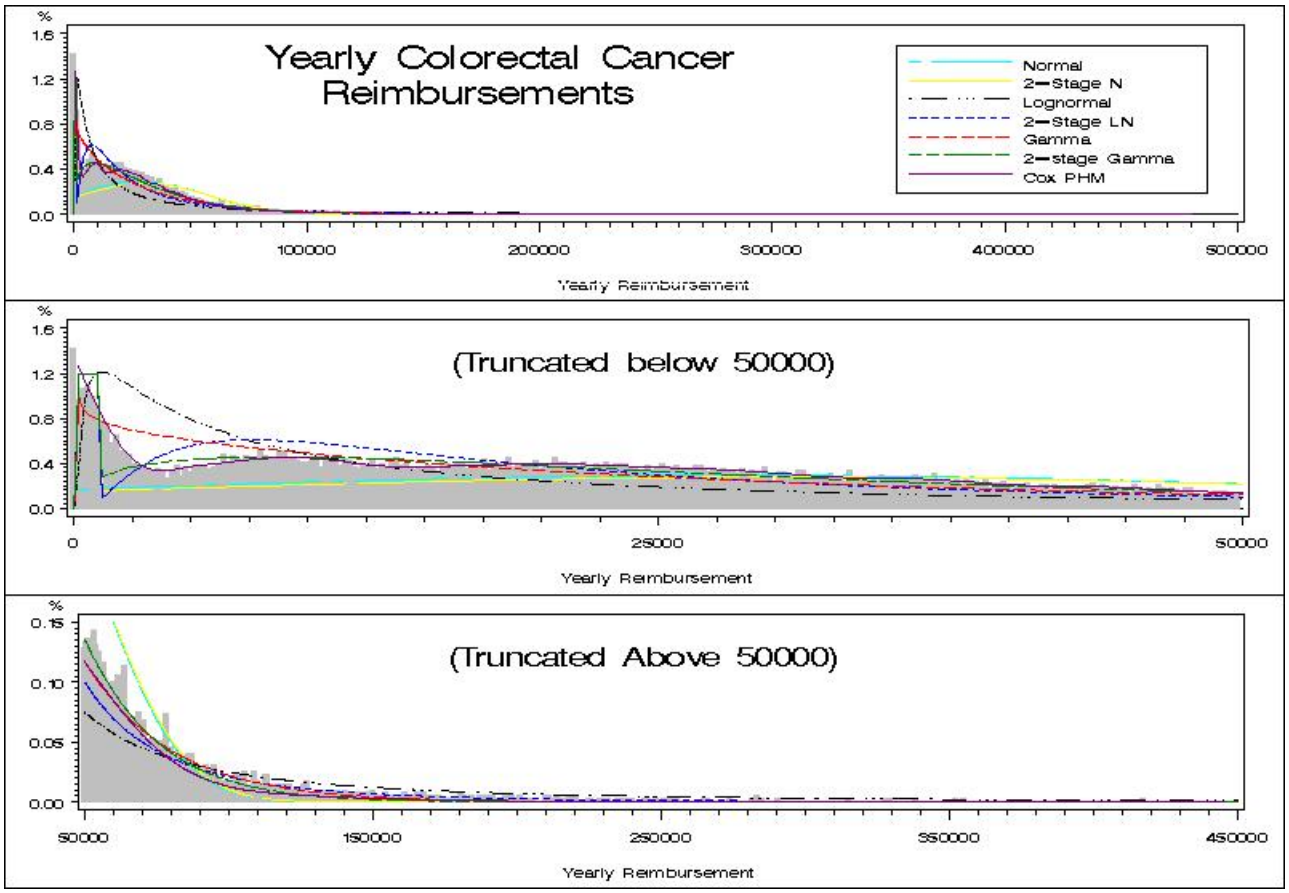

Fig. 1. **SEER Yearly Colorectal Cancer Reimbursements**

[*Received*]

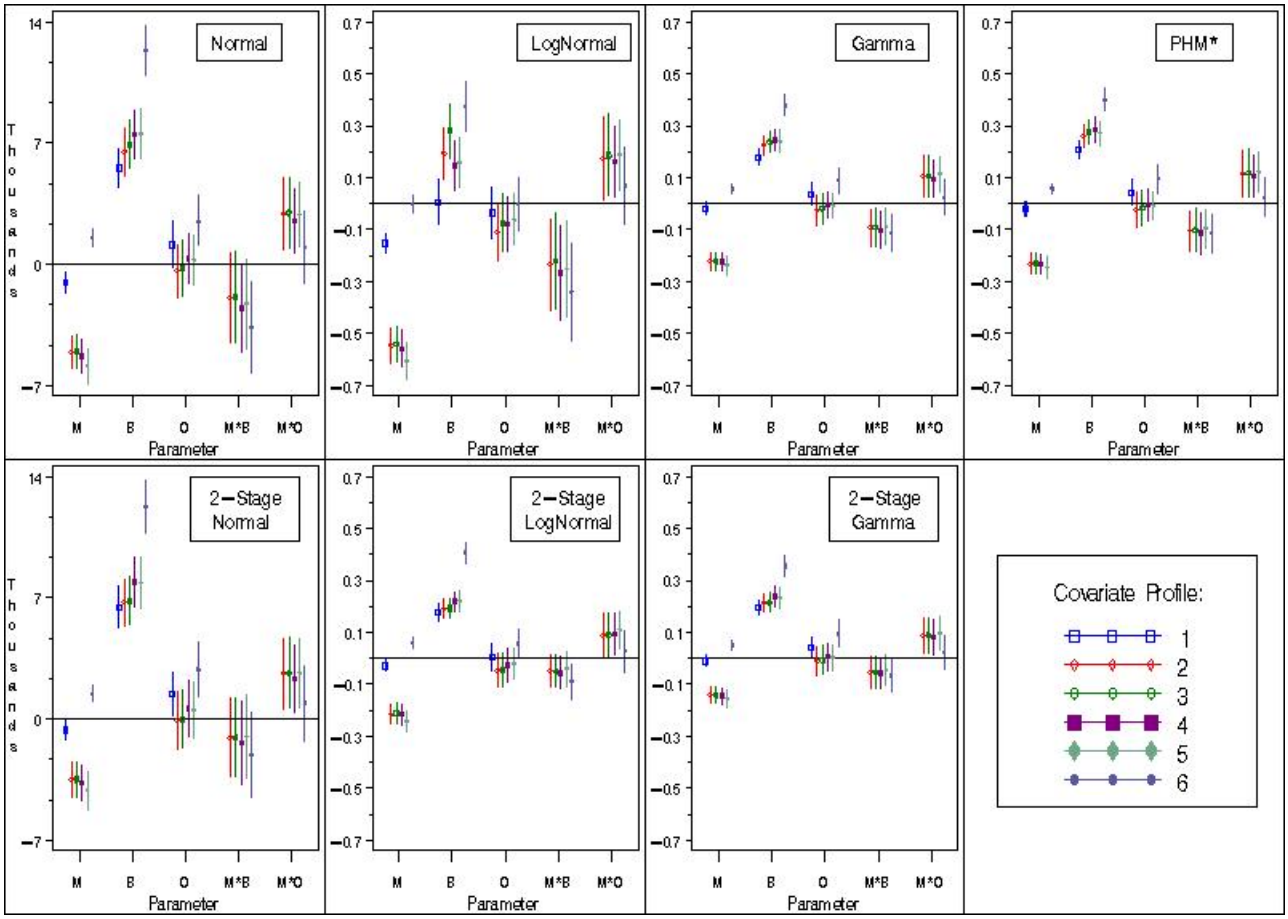

Fig. 2. **Gender/Ethnicity Estimates & 95% C.I.s:** (White Female = Reference Group)

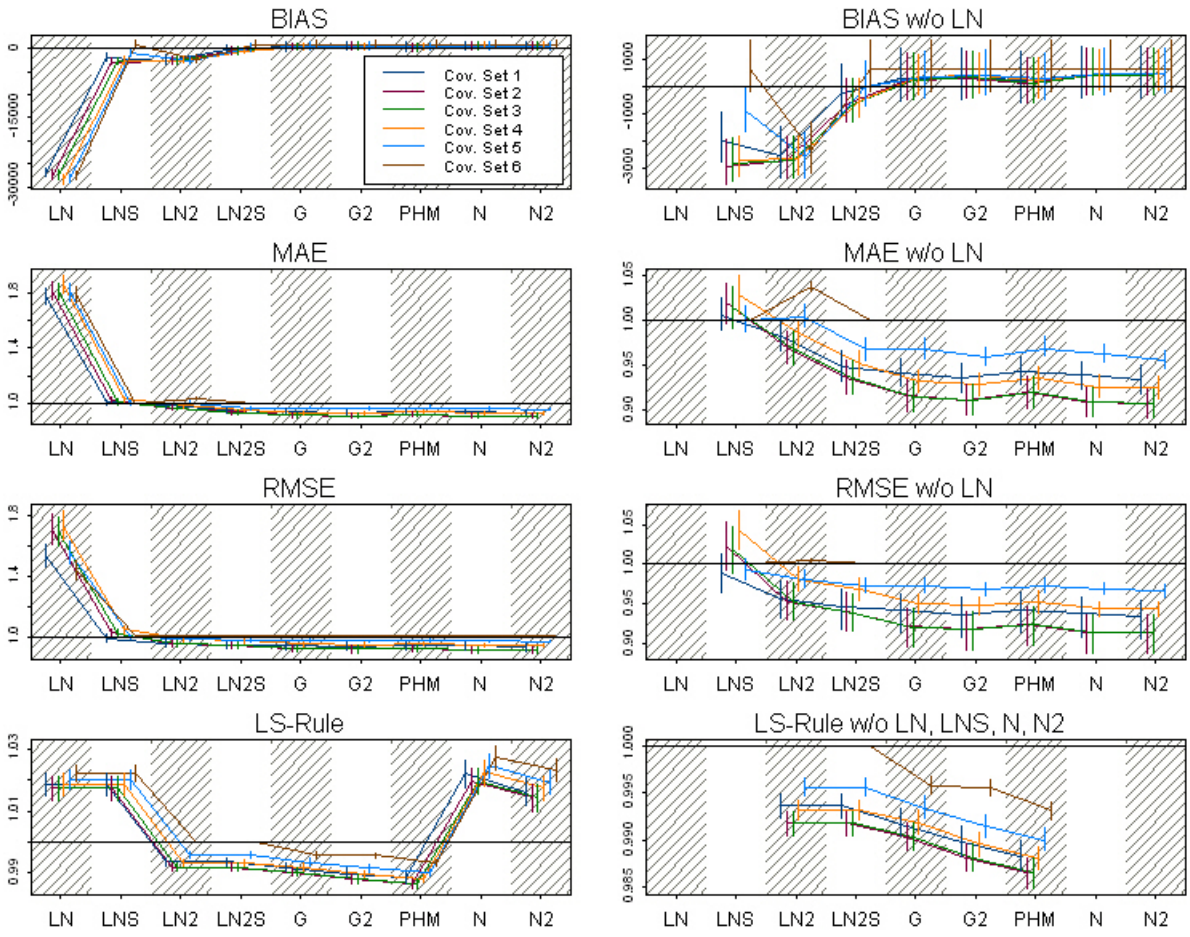

Fig. 3. **Cross Validation Measures:** (MAE, RMSE & LS-Rule measures are Relative to the Gender*Ethnicity Profile in the 2-Stage Lognormal Model)

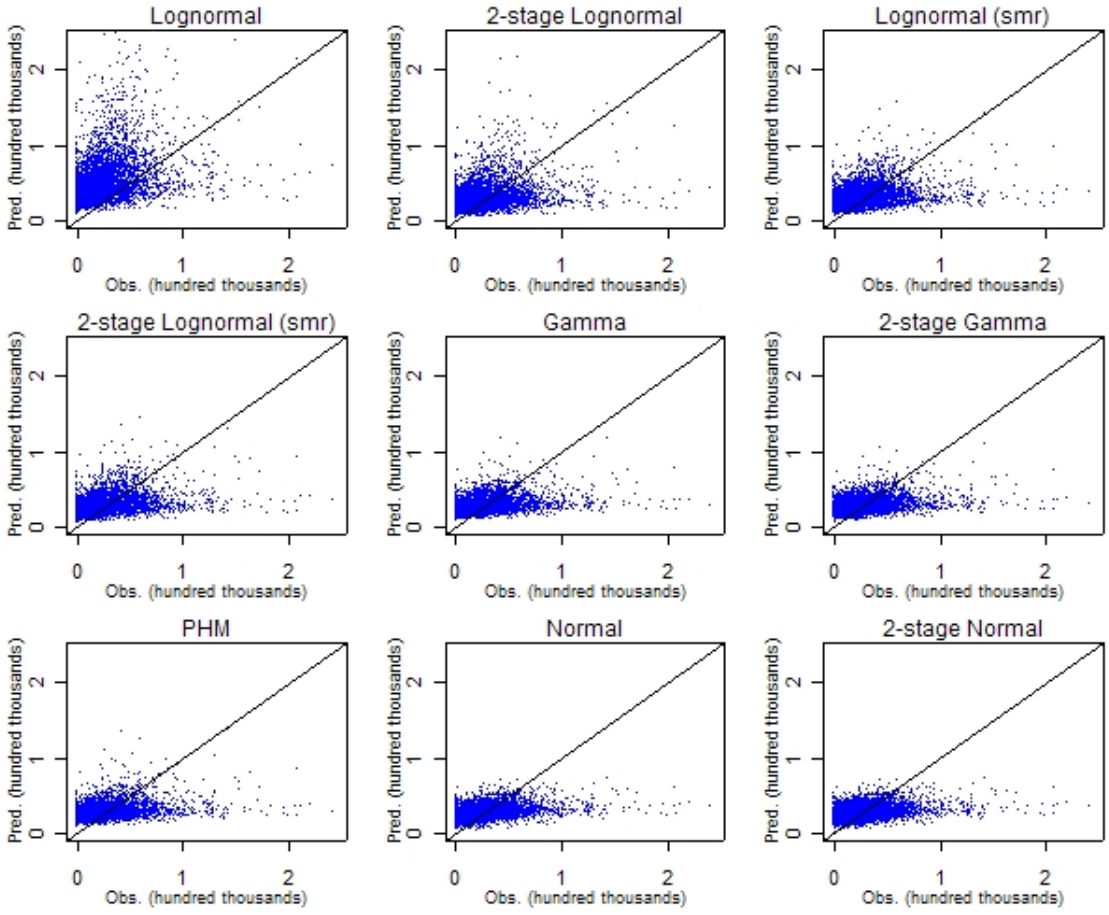

Fig. 4. **Validation Set: Predicted $ vs. Observed $ (Cov. Profile 3)**