Analyzing Health Care Cost Data in the Presence of Heterogeneous Patients: Wouldn't Three Pieces Be Better than One?

By

James F. Burgess, Jr., Ph.D.*

And

Chuan Zhou (chuan.zhou@vanderbilt.edu), Vanderbilt University

Paul A. Fishman (fishman.p@ghc.org), Group Health Cooperative

Li Wang (Li.Wang2@va.gov), VA Puget Sound Health Care System

Kevin L. Sloan (Kevin.Sloan@va.gov), VA Puget Sound Health Care System

February, 2007

*Center for Organization, Leadership, and Management Research, US Dept. of Veterans Affairs, Boston, MA and Department of Health Policy and Management, Boston University School of Public Health, Boston, MA. James.Burgess@va.gov or jfburges@bu.edu.

Preliminary Version for Presentation at the 18th Annual Health Economics Conference (AHEC) at Arizona State University, Tempe, AZ, March 16-17, 2007. Please do not cite or quote without permission. The opinions in this paper are not to be construed as representing those of the US Department of Veterans Affairs. Funding for this work was provided by VA Health Services Research and Development IIR grant 00-077-3 to Kevin Sloan.

ABSTRACT

Rationale: Accurate estimation and prediction of health care costs play crucial roles in the decisions of health plans and other health care agencies on policies and resource allocation. A particularly problematic issue that underlies difficulties in filling that role is the impact of heterogeneous patients on the process. Most previous attempts to address this issue have not balanced "simplicity and transparency" and "statistical soundness and unbiasedness" in the implicit loss function and decision calculus for choosing methods.

Objectives: Development of any cost model should take into account three features of cost data: 1) cost data are typically non-negative and skewed to the right, 2) health costs are often hierarchical and such hierarchical structures introduce correlations among the cost data, and 3) there is often a significant amount of heteroskedasticity in the data which generates biased estimates if not accounted for. Many methods have been proposed that address one or more of these three difficulties. However, sophisticated methods that do this require a high level of statistical knowledge. are not easily interpretable or understood by managers of health care agencies, and also fail to predict costs accurately across the continuum of heterogeneity in severity of illness. And less sophisticated methods that do this tend to be biased and/or perform very poorly in particular regions of the cost distribution. Our objective is to develop an pproach that employs simple generally understood ordinary least squares methods in a new way to address the empirical features of cost data.

Methodology: Case mix diagnostic and demographic information generally employed for risk adjustment can be used first to classify individual patients into spending type groups, then used again to predict health care costs using a separate ordinary least squares regression in each group. This is a special case of sub-classification models in statistics in a way that is closely related to the mixture model approach that has been applied to this problem by Deb and Burgess (2006), but is computationally much simpler to implement. For comparison, we focus on residual mean square error (RMSE) for overall evaluation, mean absolute prediction error (MAPE) for observation by observation evaluation, and predictive ratios by deciles for evaluation across the distribution of costs. We use a 50/50 split sample validation against an array of alternative methods used in the literature on the population of FY2001 users of the US Department of Veterans Affairs health care system to compare models to a three piece implemenation of the new model proposed here.

Results: In the FY2001 VA patient sample, there are 3,744,264 patients, and in populations this large, the 50/50 validation does not reveal any overfitting or any significant differences in the validation sample vs. the estimation sample. In particular, we compare six other models to our proposed three piece OLS model. We compare two other simple models (OLS and no intercept OLS) and four more complex models, two GLM models (log gaussian and log gamma) and two log

retransformation models (one using the Duan correction, the other using the Zhou correction). OLS models, including the three piece model, have a mean predicted that is closest to the observed mean, and thus also do best using the RMSE criterion. The GLM log gaussian model comes closest to these among the more complicated models. The three piece model does even better on the MAPE criterion (e.g. $2785 vs. $3312 for the OLS). For the predictive ratio results, the 10th decile always is pulled in and estimated accurately and the three piece model actually does slightly worse in that decile. However, a well known problem with the OLS method also is overprediction in the 7th-9th deciles. The three piece model does much better there. And it also addresses the negative prediction problem well in the first three deciles. The more complex models each have their biases in particular deciles. The GLM log gaussian model dramatically overpredicts most of the lower deciles and underpredicts the 10th decile. The other more complicated methods overpredict the 10th decile by ratios exceeding 2.0, though they do better in the other deciles than the GLM log gaussian model.

Conclusion: One problem with ordinary least squares models in general is that they can predict negative cost. Many researchers have estimated no intercept ordinary least squares models and taken other steps to avoid this problem, the approach we propose here accomplishes much the same outcome without the bias problem that no intercept models create and predicts all low cost people better than standard ordinary least squares approaches. This approach also performs much better on all of the evaluation criteria than other more complex statistical approaches. We have taken a simple approach to subclassification, as opposed to more complex methods previously suggested (i.e. mixture models), and this achieves the balance between “simplicity and transparency" and "statistical soundness and unbiasedness" that we established as our objective.

1. Introduction

Accurate estimation and prediction of health care costs play crucial roles in the decisions of health plans and other health care agencies on policies and resource allocation. Typically, these agencies have access to basic demographic information (age/sex) and diagnostic information as predictors of cost outcomes. Concurrent information is used to characterize current cost when agencies are focused on resource allocation or other estimation problems. Retrospective information is used to characterize prospective cost when agencies are designing payment models or other prediction problems. Developing consistent policy useful models has been a considerable challenge in recent years, caused primarily by difficulties in accounting for the impact of heterogeneous patients on the process.

Most of the previous literature has either focused their approaches to health care cost estimation on simplicity and transparency designed to appeal to managers and policymakers or on statistical soundness and unbiasedness designed to address the nature of the underlying statistical distributions of costs. Many transparent and easy to explain solutions based around adapting ordinary least squares (OLS), including the approaches used in developing the Diagnostic Cost Group (DCG) family of models (Ash et al. (1986, 1989); Ellis and Ash (1995); Ellis et al. (1996); and Pope et al. (2000)), violate strict statistical assumptions of unbiasedness by (e.g.) estimating models without an intercept. Alternative statistical approaches attempting to stay within the bounds of statistical soundness have become more and more complicated (Blough et al.

(1999); Zhou et al. (2001); Welsh and Zhou (2006); Manning, Basu, and Mullahy (2005); Basu and Rathouz (2005); and Deb and Burgess (2006)), but without solving some essential aspects of explaining patient heterogeneity.

In particular, if we set as a goal balancing unbiasedness in all portions of the cost distribution against simplicity and transparency for management and policy applications, then we move in a new direction. Deb and Burgess (2006) propose a finite mixture model solution to the cost estimation problem that generally fits into the tendency noted above toward relatively complex modeling. The contribution of this paper is to note that finite mixture models fall into a larger class of models, subclassification models, which take a population and subclassify it into subgroups. Simplicity and transparency then suggests circumventing computational difficulties in finite mixture models through a more direct two step procedure. First, we classify patients with different health care spending potentials into more homogeneous subgroups based on risk scores, a weighted linear combination of demographic and diagnostic instruments. Then, we estimate predicted costs within each piece or subclass with separate OLS regressions. We demonstrate the effectiveness of this simple and transparent approach with a three piece model of total patient cost for U.S. Department of Veterans Affairs (VA) patients in Fiscal Year 2001.

2. Methodology and Framework

The nature of the implicit loss functions that researchers and policymakers use for choosing methods of cost estimation are different. And while we do not

intend to lay out a formal decision calculus here, we do want to define terms and be as clear as possible. A risk instrument is the organization of diagnoses, pharmacy dispenses, demographic information, or procedures into risk categories. A risk measure is a specific unit of medical care or a proxy for care needs such as cost, utilization, or a clinical measure such as mortality or morbidity. A risk model is the prediction that results from estimating risk measures with risk instruments. And then risk adjustment is the application of a risk model to resource allocation, payment, clinical performance or efficiency measurement, provider or facility profile, or other policy/management goal. In this paper we will employ a risk instrument that combines diagnoses and demographic information to apply to a risk measure of VA annual total (inpatient/outpatient/pharmacy) cost using concurrent information. The point of the paper is to evaluate this risk model for cost, without specifically applying it to risk adjustment in a particular context.

2.1 Existing approaches to modeling cost data

Jones (2000) and other researchers have highlighted three features or key difficulties in modeling health care cost data. First, they are typically non-negative and skewed to the right. This is a key part of the patient heterogeneity, which is heterogeneous in the types and amounts of services that are provided. Second, they are hierarchical in structure, with patients nested within providers and providers are nested within facilities or practices, which again are nested into systems. This introduces cost correlations within the data that ideally are accounted for. And finally, Manning (1998), Mullahy (1998), and others have

emphasized the significant amounts of heteroskedasticity in the data and the biases that result in estimation for failure to account for that heteroskedasticity.

Linear regression OLS methods on log-transformed data or generalized linear models (GLM) with log links have been proposed to address the non-normality (Blough et al. (1999); and Manning and Mullahy (2001)). Smearing estimators (Duan, 1983) allow one to make inferences for mean cost on the original dollar scale if transformations are used and yet further complicated methods for improving the smearing approach to adjust for heteroskedasticity (Zhou et al., 2001) have been proposed. Instead of transformation to alter the scale of the data, another approach is to allow a flexible exponential family to be assumed for the error distribution and use different link functions between risk measures like cost and risk instruments like diagnoses. Gamma models as well as Gaussian models with log links have been used to model the skewness and ensure non-negative responses. Manning, Basu and Mullahy (2005) and Basu and Rathouz (2005) provide some guidance and approaches for designing optimum GLM models. Note that one key advantage of GLM models is that the log transformation is on the expected mean, so no retransformation after exponentiation is needed.

In contrast to these more complicated measures, the demand of policymakers for simpler models has led many to use simple OLS models (see Diehr et al., 1999 for a review) that work especially well for the middle of the distribution and pulls in the extreme values relatively well. However, the intercept in OLS models generates predictions of negative cost for a relatively large

number of patients at the low end of the cost distribution, typically more than 10% of them depending on how many patients have very low costs. Pope et al. (2000) document how using a no-intercept OLS model, despite the statistical biases created, addresses the negative cost prediction issue very well, especially when combined with a variable selection process to eliminate negative parameters on diagnostic categories. Their DCG family of models based on the no-intercept model has been very successful at meeting government and private payor needs, including being selected to risk adjust the Medicare managed care program.

2.2 Risk instruments

Gruenberg et al. (1996), among others, documents how much is gained from adding patient level risk information to better determine costs. Risk instruments vary in their approach to using diagnostic, age, and gender information to predict costs. And some add in procedures or pharmacy prescriptions to improve the fit. But commonly, a calendar year is employed as a unit of analysis with a vector of indicator variables for groupings of patient diagnoses and age-sex levels forming the set of independent variables. Some instruments further refine the classification process by creating hierarchies across the diagnostic groups such that, if an individual receives diagnoses that map to more than one grouping within the hierarchy (say diabetes mellitus with and without complications), only that variable corresponding to the most severe/costly grouping is coded as present. Hierarchies attempt to improve classification precision by reducing the overlap among similar diagnostic groups.

In this work, we considered and tested a wide variety of commonly employed risk instruments. We concluded in another paper that the results do not depend heavily on which risk instrument is selected (Fishman et al., 1996). As a result, all of the work in this paper uses the DCG risk instrument to simplify the presentation of the results, though none of the substantive conclusions depends upon which one is chosen.

The available DCG RiskSmart software makes this instrument even more desirable in practice and presents different ways of breaking out disparate risks intuitively related to the approach proposed here. The DCG method was developed to predict Medicare payments (Ash et al. (1999); Ellis and Ash (1995); and Ellis et al. (1996)), and is the mandated case-mix method for adjusting Medicare capitation payments (Pope et al. (2000)). One variant of this model is the DCG/Hierarchical Condition Category model (DCG/HCC model). This model assigns ICD- 9-CM codes to “DxGroups” that are clinically related and similar with respect to levels of resource use. These groups also are aggregated and collapsed, using hierarchies to allow or disallow multiple groups in certain circumstances, into HCCs (Ellis et al. (1996)). The 184 HCCs and the Age/Gender variables are the covariates in the models reported in this paper.

2.3 Subclassification as an approach to addressing heterogeneity

The key motivation for the proposed alternative method in this paper is the generic observation that patients with similar risk instrument characteristics tend to have more homogeneous cost distributions. Thus, we hypothesize that patient populations consist of multiple sub-groups with usage patterns of services

relative to diagnoses that differ. But the number of sub-groups is unknown as is a methodology for identifying and defining them. In some sense, risk instruments like HCC/DCGs are attempting to do this as well, hypothesizing that the increment to cost from falling into an HCC is constant across the spectrum of the population. Conceptually, we proceed differently, we hypothesize a latent variable that identifies differences in expected spending patterns for each individual patient in the sub-class, which is determined by the diagnostic-based risk instrument. This latent variable leads naturally to a finite mixture model framework, which Deb and Burgess (2006) have explored elsewhere in similar data. For comprehensive discussions of how to form finite mixture models in general, see Titterington et al. (1985) and McLachlan and Peel (2000).

In another paper, Zhou et al. (2006), we discuss details of the statistical theory underlying our proposed approach in greater detail, but if we lay out simple straightforward criteria of symmetry (equal proportions of high and low cost), separation (cost groups well separated from each other), and homoskedasticity (reduced variability in each group), we come to a proposed split of 20%/60%/20% into three groups. Intuitively, we are defining a low cost patient group who might receive routine screening and evaluations, but do not have enough diagnoses to be considered sick, the lowest 20% of the risk profile. Then we are defining a parallel symmetric high cost patient group who have severe disease or many complications that are likely to cause them to employ many types of health services or some very specific high cost services, the

highest 20% of the risk profile. The remaining 60% will be considered medium cost patients who fall in between these two groups.

The process to implement this three piece model is as follows:

1) Divide the sample into an estimation and validation subsamples (here 50-50).

2) Run an OLS regression on the estimation sample with cost predicted by the risk instrument and call this the population-level risk regression, generating beta weights we will call risk weights. Generate rank statistics on the predicted cost from this risk regression (we call these risk scores, since they could be normalized to a standard population) and divide the sample into three pieces using the 20/60/20 split of the estimation sample.

3) Run three additional OLS regressions on the estimation sample with cost predicted by the risk instrument again, one regression for each of the three subclassifications. Call this the cost regression, generating beta weights as three sets of cost weights.

4) Take the entire set of coefficient estimates, both risk weights and cost weights, and move to the validation sample. First generate risk scores for each patient using the risk weights, classify patients into the three risk strata or subclasses, and then predict cost using the stratum specific cost weights.

Note that the entire procedure, both steps of subclassification and estimation, is being validated against the other approaches. Also, this is not a full mixture model, since the latent group indicator is defined directly against the risk instrument, rather than being treated as a random variable. Also note this is NOT stratifying based on observed cost, the actual cost varies across the

spectrum in each of the three subclasses. But the key point is that the process is simple and intuitive, and involves OLS at each stage.

2.4 Diagnostics and goodness of fit

Explicit criteria are needed to compare models and are related to the implicit loss function that one employs in trading off attributes of measures. The specific criteria could vary depending on the context of the uses to which one puts the analysis. The tendency in this area of research has been to focus on the raw prediction power, especially R^2, at the expense of a more expansive view of goodness of fit attributes. First, to avoid over-fitting, we conducted a 50-50 split sample validation and will present and evaluate the models only on the validation sample results. Nevertheless, using a large database encompassing the entire VA patient population of over 3 million individuals means that the overfitting issues are minimized and the results are quite similar.

Then, we consider the following goodness of fit criteria. Predicted means are best approximated by OLS types of methods, which work especially precisely in large samples such as we have here. R^2 measures have been the most common primary criterion in the past, although the sum of squares favors a model that fits the highest cost patients well, at the expense of the lower part of the cost distribution. Partly as a result, we believe that it is more important in many applications to characterize costs better across the distribution of costs from high cost to low cost and to measure bias directly both in the population and in the individual observations. In particular, the residual mean square error (RMSE) represents the extent of the bias in the overall sample, while the mean

absolute prediction error (MAPE) gives the total bias in the individual observations.

To assess the accuracy of the prediction across the entire range of the distribution we adapt the approach of Hosmer and Lemeshow (1989) for assessing the accuracy of prediction in logistic regression. We calculate predictive ratios for each decile of the distribution as the sum of predicted cost for all individuals in the decile divided by the sum of actual cost of the individuals in that subgroup. In the literature previously, Mincer and Zarnowitz (1969) have done similar assessments in the aggregate by regressing predicted costs on observed costs and reporting the slope and intercept of those regressions.

3. Data and Results

3.1 Data and models

The VA population and health care system in FY 2001 spent $20,129,000,000 on medical care for 3,744,264 veterans. Justice et al. (2006) discusses the strengths and weaknesses of using VA as a laboratory for research and how results are likely to generalize to other populations. In general, the VA is most closely aligned to the Medicare population, with elderly patients and many patients under 65 with service connected disabilities. It also is predominantly male. Table 1 summarizes the characteristics of the estimation sample and the validation sample relative to the entire sample. Total cost for these purposes is defined as inpatient cost plus outpatient cost plus pharmacy cost. Table 2 summarizes the data for the three piece model by subclass.

We examine five alternatives to our three piece proposed model in this version of the paper (and will shortly add a no-intercept OLS model to that list). A simple OLS model is contrasted against two GLM log models (one Gaussian and one Gamma) and two log-retransformation models, one using the simple Duan smearing estimator and the other using Zhou's heteroskedasticity adjusted estimator.

Figures 1 and 2 depict the distributions of individual total cost for the whole sample, and the three pieces identified by the risk score matching on original and log-scales, respectively. They suggest that the total population consists of multiple groups, thus motivating the mixture model approach. Note the distributions for different risk strata are more symmetric on the log-scale, but this would still require re-transformation to original scale. We have tried fitting separate models on the log-scale and then re-transforming back to the original scale with smear estimators. The results are better than a single piece approach using Duan's method, but still not better than the multi-piece model on the original scale (results not shown).

3.2 Main results

Table 3 contains the goodness of fit comparisons for the six models (no-intercept OLS to be added). OLS models, including the three piece model, have a mean predicted that is closest to the observed mean, and thus also do best using the RMSE criterion. The GLM log gaussian model comes closest to these among the more complicated models. The three piece model does even better on the MAPE criterion (e.g. $2785 vs. $3312 for the OLS). For the predictive ratio

results, the 10th decile always is pulled in and estimated accurately and the three piece model actually does slightly worse in that decile. However, a well known problem with the OLS method also is overprediction in the 7th-9th deciles. The three piece model does much better there. And it also addresses the negative prediction problem well in the first three deciles. The more complex models each have their biases in particular deciles. The GLM log gaussian model dramatically overpredicts most of the lower deciles and underpredicts the 10th decile. The other more complicated methods overpredict the 10th decile by ratios exceeding 2.0, though they do better in the other deciles than the GLM log gaussian model.

It is also interesting to compare the parameter estimates between the OLS and the three pieces of the three piece OLS against the HCCs. Figures 3, 4 and 5 compare the risk instrument variables that are included in each of the three pieces against the corresponding OLS parameters. There are 200 possible explanatory variables, 184 HCCs, 15 age/sex classification variables, and an intercept. Note that 7 of the HCCs document conditions never observed in the veteran population, most of these being neonate, newborn, or birthing diagnoses and only the high cost patient subclass has estimates for all of the 193 other variables. As can be seen from Figure 5, this high cost patient subclass also has the closest parameters to the OLS, since the high cost patients drive the OLS regression, there is only slightly more variability in the high cost estimates. In Figure 4, 166 parameters are estimated and the medium cost patient cost weights are half of the OLS estimates with less than half the variance. Finally, in Figure 3, only half the parameters (102) are estimated, but the mean estimates

are similar even though the OLS estimates have more than twice the variance. There are also fewer negative parameter estimates, to minimize the negative cost estimation effect. In general, across the three pieces, the parameter estimates drop as one goes from the high cost to the medium cost to the low cost subclasses. And in those cases where all three parameters are estimated, the OLS estimate tends to be between the high cost and the medium cost estimate. In HCC cases where only the high cost parameter is estimated, the general OLS estimate and the high cost weight are very close. For example, in HCC 105 (Vascular Disease), the cost weights (low to high) are $399, $528, and $1,987, respectively, while the OLS estimate is $849, while for HCC 104 (Vascular Disease with Complications) there are no patients in the medium or low cost subclasses and the OLS estimate is $12,427 while the high cost subclass cost weight is $13,807.

4. Conclusions and Directions for Future Research

One problem with ordinary least squares models in general is that they can predict negative cost. Many researchers have estimated no intercept ordinary least squares models and taken other steps to avoid this problem, the approach we propose here accomplishes much the same outcome without the bias problem that no intercept models create and predicts all low cost people better than standard ordinary least squares approaches. This approach also performs much better on all of the evaluation criteria than other more complex statistical approaches. We have taken a simple approach to subclassification, as

opposed to more complex methods previously suggested (i.e. mixture models), and this achieves the balance between “simplicity and transparency" and "statistical soundness and unbiasedness" that we established as our objective. Usefulness of this approach requires that a cost-benefit analysis on the cost of doing analyses is part of the implicit loss function for decision making.

Within this same project, we have tested hierarchical modeling to predict facility cost and to profile facility costs. Such analyses are easily appended to what we have presented here, as well as other practical concerns of managers and policymakers. A number of considerations and extensions are possible in addition to adding the no-intercept model, which is another more blunt approach to trying to address the same problem, but without the biased estimation that the no-intercept approach creates. This works on very large datasets. How well would it work on other sizes and disease severity for patient populations? The VA population size and disease severity made a particularly fruitful test case.

Acknowledgements: The research team acknowledges the many intellectual debts which informed the track of this research, especially those of Will Manning in recommending the adaptation of the Hosmer-Lemeshow tests that drove these results to their conclusion, Partha Deb in suggesting sub-classification models to approach this problem in the first place, and Arlene Ash, Randy Ellis, Greg Pope and others who developed the no-intercept OLS approach to the Diagnostic Cost Group models which has been a particularly attractive practical alternative to date to address the problem on which we focus.

REFERENCES

Ash A, Porell F, Gruenberg L, et al. “An Analysis of Alternative AAPCC models using data from the continuous Medicare history sample.” Report prepared for the Health Care Financing Administration. Health Policy Research Consortium, Brandeis/Boston University 1986.

Ash A, Porell F, Gruenberg L, Sawitz E, Beiser A. “Adjusting Medicare capitation payments using prior hospitalization data.” *Health Care Financing Review* 1989; 10:17–29.

Basu A, Rathouz P. “Estimating marginal and incremental effects on health outcomes using flexible link and variance-function parameters.” *Biostatistics* 2005; 6(1): 93-109.

Blough, DK, Madden, CW, Hornbrook, MC. “Modeling risk using generalized linear models.” *Journal of Health Economics* 1999; 18: 153–171.

Deb, P., Burgess Jr., J.F. “A quasi-experimental comparison of statistical models for health care expenditures.” Working manuscript 2006.

Dempster, AP, Laird, NM, Rubin, DB “Maximum likelihood from incomplete data via the EM algorithm.” *Journal of the Royal Statistical Society, Series B* 1977; 39: 1–38.

Diehr, P, Yanez, D, Ash, A, Hornbrook, M, Lin, DY. “Methods for analyzing health care utilization and costs.” *Annual Review of Public Health* 1999; 20: 125–144.

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

Ellis RP, Ash A. “Refinements to the Diagnostic Cost Group (DCG) model.” *Inquiry* 1995; 32:418–429.

Ellis, RP, Pope, GC, Iezzoni, L, Ayanian, JZ, Bates, DW, Burstin, H et al. “Diagnosis-based risk adjustment for Medicare capitation payments.” *Health Care Financing Review* 1996; 17:101–28.

Fishman, P, Sloan, K, Burgess Jr., J, Zhou, C, Wang, L.”Evaluating alternative risk assessment models: Evidence from the US veteran population.” Group Health Center for Health Studies Working Paper 2006; http://www.centerforhealthstudies.org/ctrstaff/fishman.html.

Gruenberg L, Kaganova E, Hornbrook MC. “Improving the AAPCC with health-status measures from the MCBS.” *Health Care Financing Review* 1996; 17: 59–75.

Hosmer, DW, Lemeshow, S. *Applied Logistic Regression*, Wiley, Chichester, UK, 1989.

Jones, A. “Health econometrics.” *Handbook of Health Economics* (Culyer, A, Newhouse, J., Eds.) Elsevier, Amsterdam, 2000.

Justice, AC, Erdos, J, Brandt, C, Conigliaro, J, Tierney, W, Bryant, K. “The Veterans Affairs Healthcare System: A unique laboratory for observational and interventional research.” *Medical Care (suppl.)* 2006; 44: S7–S12.

Manning,WG. “The logged dependent variable, heteroscedasticity, and the retransformation problem.” *Journal of Health Economics* 1998; 17: 283–295.

Manning W, Mullahy J. “Estimating log models: to transform or not to transform?” *Journal of Health Economics* 2001; 20: 461–494.

Manning, WG, Basu, A, Mullahy, WH. “Generalized modeling approaches to risk adjustment of skewed outcomes data.” *Journal of Health Economics* 2005; 24: 465–488.

McLachlan, J, Peel, D. *Finite Mixture Models*. New York, John Wiley & Sons Ltd., 2000.

Mincer, J. and Zarnowitz V. “The Evaluation of Economic Forecasts,” in J. Mincer (Ed.), Economic Forecasts and Expectation, National Bureau of Research, New York, 1969.

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

Pope GC, Ellis RC, Ash AS, Liu CF, et al. “The principal inpatient Diagnostic Cost Group Model for Medicare risk adjustment.” *Health Care Financing Review* 2000; 21:93-118.

Salem-Schatz S, Moore G, Rucker M, Pearson SD. “The case for case-mix adjustment in practice profiling. When good apples look bad.” *JAMA* 1994; 272:871–874.

Titterington, DM, Smith, AFM, Makov, UE, *Statistical analysis of finite mixture distributions*, John Wiley & Sons, 1985.

Welsh AH, Zhou, XH.”Estimating the retransformed mean in a heteroscedastic two-part model.” *Journal of Statistical Planning and Inferences* 2006; 36: 860–881.

Zhou, C, Burgess Jr., JF, Fishman, PA, Wang, L, Sloan, KL. "A multi-piece model for medical care cost data." Working Manuscript 2006.

Zhou, XH, Stroupe, KT, Tierney, W.M. "Regression analysis of health care charges with heteroscedasticity." *Applied Statistics* 2001; 50: 303–312.

TABLE 1: Population Descriptive Statistics

Population Characteristics	All Sample	Estimation Sample	Validation Sample
N	3,744,264	1,871,407	1,872,857
Mean Age, years (SD)	61.9 (14.5)	61.9 (14.5)	61.9 (14.5)
Male (%)	95.6	95.6	95.6
Mean Total Cost (SD)	$4,263 (12,820)	$4,263 (12,714)	$4,264 (12,925)
Mean HCC (SD)	5.6 (3.8)	5.6 (3.8)	5.6 (3.8)

Table 2: Population Descriptive Statistics across Subclassification Risk Categories by Risk Score

Population Characteristics	All Sample	Risk Score in Lower 20%	Risk Score in Middle 60%	Risk Score in Upper 20%
N	3,744,264	747,193	2,247,476	749,595
Mean Age, years (SD)	61.9 (14.5)	64.5 (15.1)	61.1 (14.4)	61.6 (13.9)
Male %	95.6	95.5	95.6	95.9
Mean Total Cost (SD)	$4,263 (12,820)	$603 (2,769)	$1,962 (3,920)	$14,819 (25,034)
Median Total Cost	$1,206	$390	$1,184	$6,595
Mean HCC (SD)	5.6 (3.8)	2.0 (1.4)	5.3 (2.6)	10.1 (4.3)

Table 3: FY01 Total Cost with HCC, Validation Sample (N = 1,872,857)

Diagnostics	Observed Mean Cost = $4,264					
	OLS	3 Piece OLS (20/60/20)	GLM Log Gaussian	GLM Log Gamma	Log-Retransf. Duan	Log-Retransf. Zhou
Mean Predicted	$4,263	$4,260	$5,349	$11,578	$7,730	$8,368
R^2	0.41	0.43	0.35	0.03	0.06	0.01
RMSE	$9,937	$9,726	$10,524	$334,752	$94,224	$554,007
MAPE	$3,312	$2,785	$3,976	$9,713	$5,965	$7,362
Predicted Ratios by Predicted Deciles						
1st	-1.50	0.49	3.62	1.30	1.08	0.78
2nd	-0.60	0.88	3.24	1.22	1.06	0.73
3rd	0.03	0.88	3.12	1.08	1.00	0.69
4th	0.44	0.95	2.34	1.00	1.01	0.69
5th	0.78	1.03	1.90	0.95	0.97	0.67
6th	1.03	1.06	1.78	0.91	0.95	0.65
7th	1.20	1.03	1.87	0.86	0.93	0.63
8th	1.32	0.96	1.57	0.84	0.91	0.62
9th	1.33	0.95	1.17	0.85	0.93	0.64
10th	1.01	1.03	0.80	4.25	2.61	3.15
Mincer-Zarnowitz coefficients						
Constant	2,533	2,426	3,980	-7,075	-337	-7,186
Slope	0.41	0.43	0.32	4.37	1.89	3.65

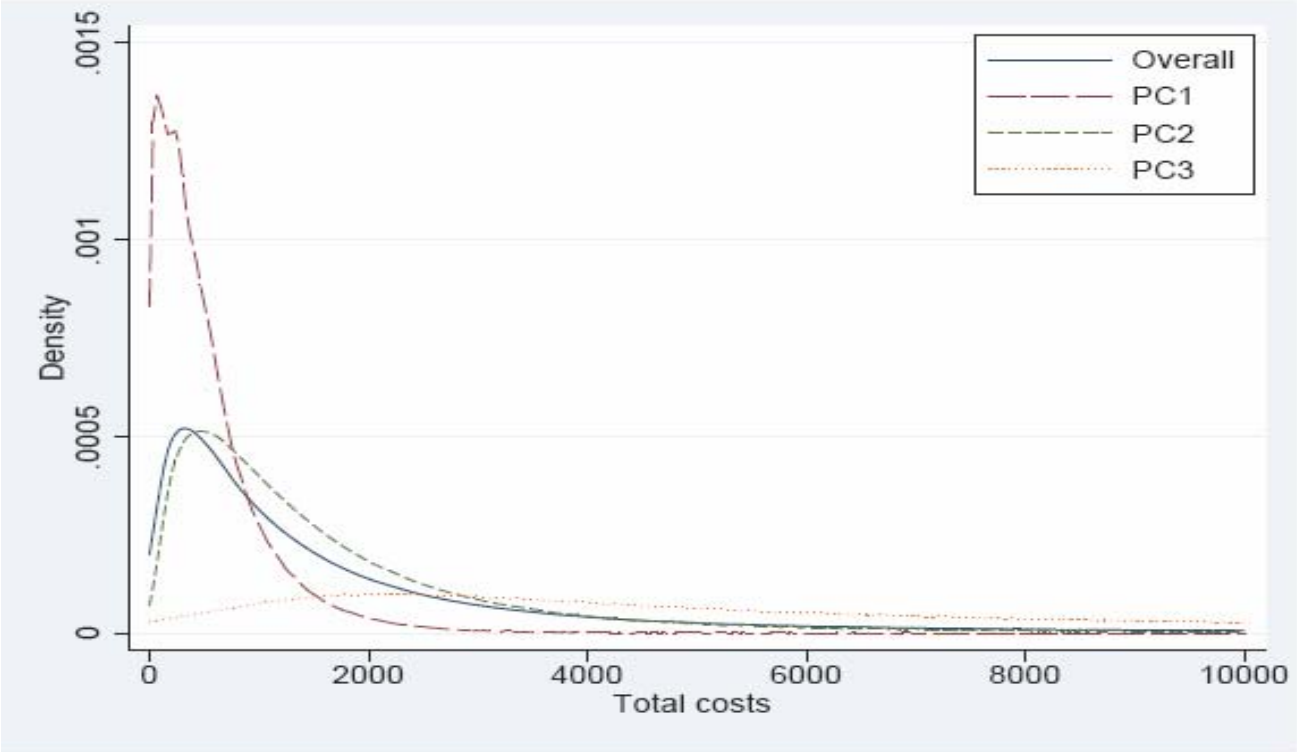

Figure 1. Distribution of overall and piece-wise cost (original scale)

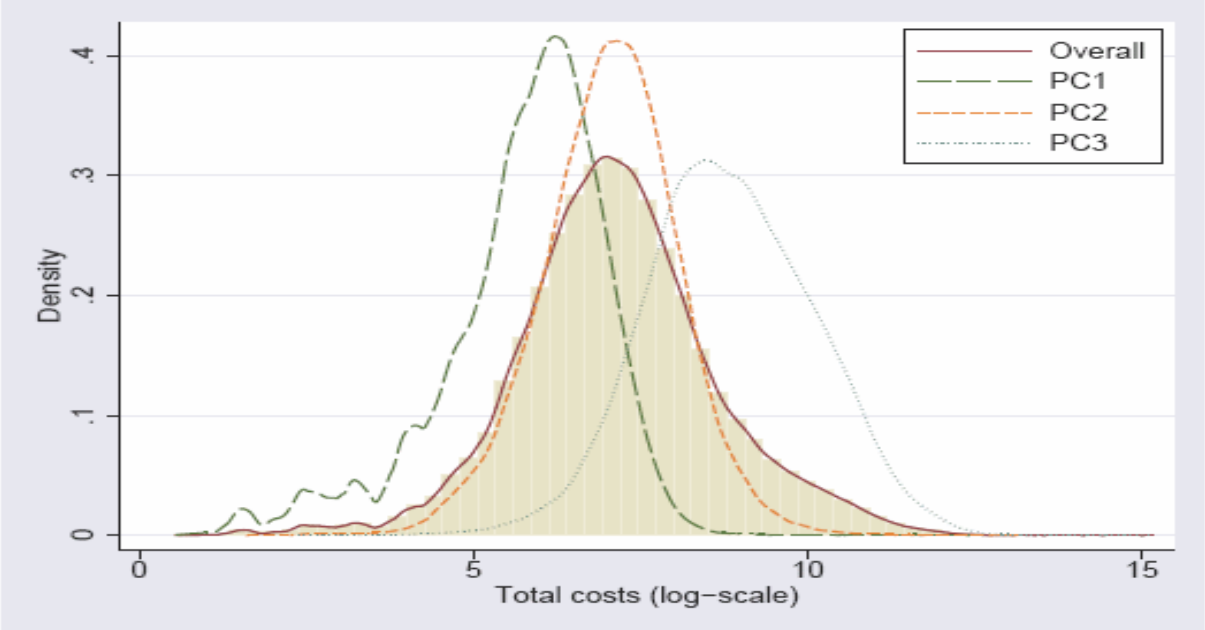

Figure 2. Distribution of overall and piece-wise cost (log scale)

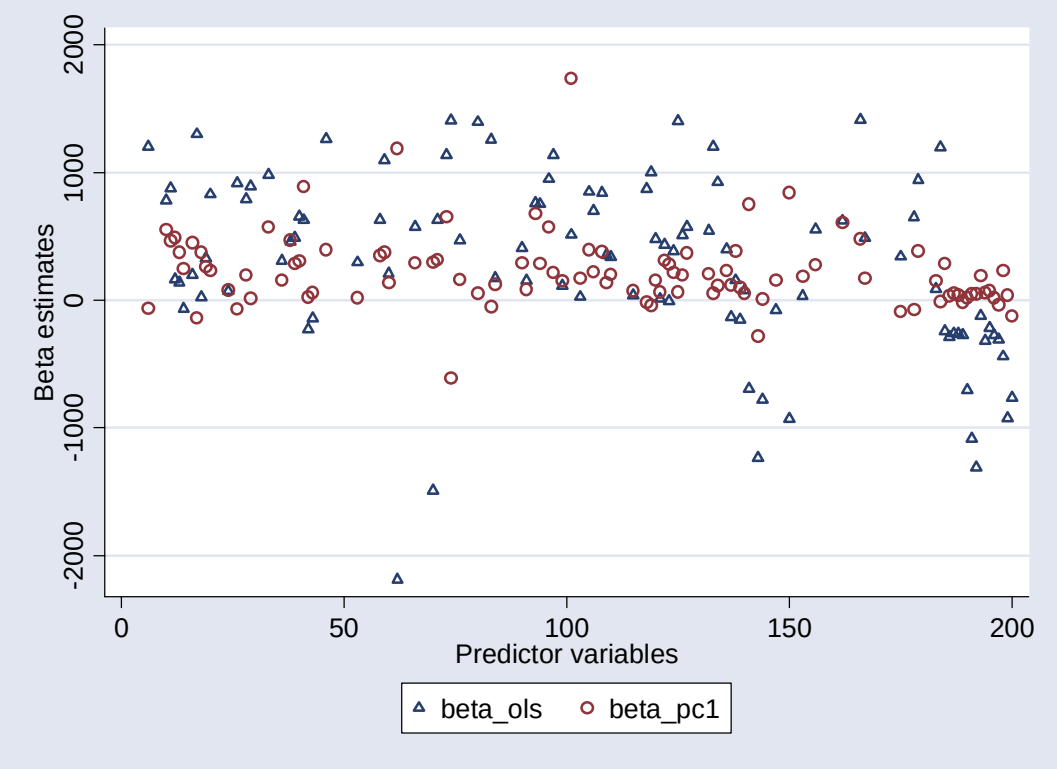

Figure 3 Beta estimates from PC1, lower 20%, vs beta estimates from OLS

Variable	Obs	Mean	Std. Dev.	Min	Max
beta_ols	102	293.66	690.06	-2189.544	1412.253
beta_pc1	102	225.47	291.48	-612.4108	1733.739

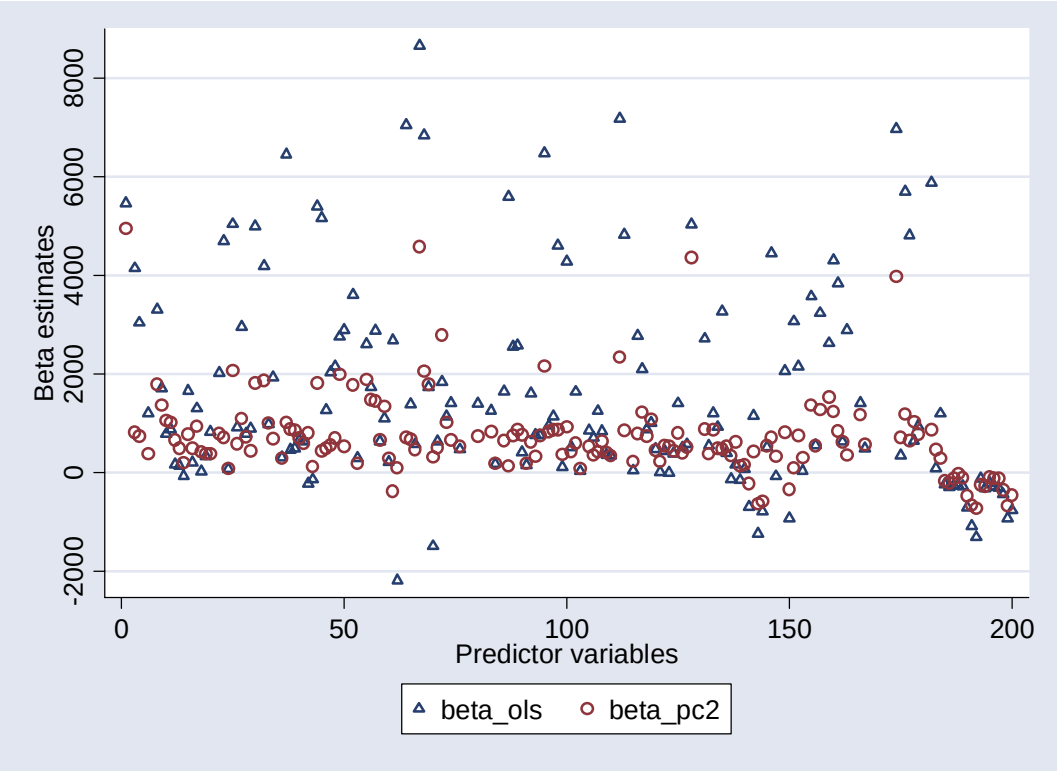

Figure 4 Beta estimates from PC2, middle 60%, vs beta estimates from OLS

Variable	Obs	Mean	Std. Dev.	Min	Max
beta_ols	166	1567.66	2035.735	-2189.54	8653.6
beta_pc2	166	717.19	844.5303	-730.474	4942.672

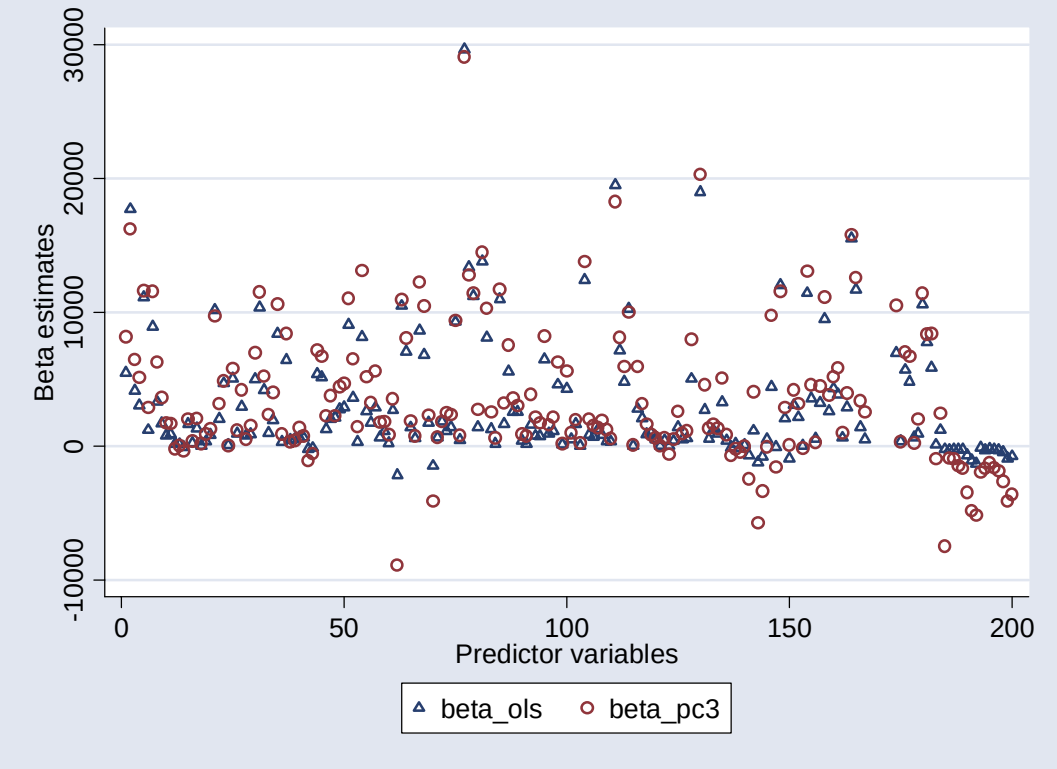

Figure 5 Beta estimates from PC1, upper 20%, vs beta estimates from OLS

Variable	Obs	Mean	Std. Dev.	Min	Max
beta_ols	193	3062.261	4511.819	-2189.54	29663.76
beta_pc3	193	3600.164	5176.652	-8903.77	29131.46