HEDG
HEALTH, ECONOMETRICS AND DATA GROUP
THE UNIVERSITY of York

WP 11/31

Applying Beta-type Size Distributions to Healthcare Cost Regressions

Andrew Jones
James Lomas
Nigel Rice

October 2011

york.ac.uk/res/herc/hedgwp

Applying Beta-type Size Distributions to Healthcare Cost Regressions

Andrew M. Jones [a] James Lomas [a,b,*] Nigel Rice [b]

[a] *Department of Economics and Related Studies, University of York, YO10 5DD, UK*

[b] *Centre for Health Economics, University of York, YO10 5DD, UK*

November 7, 2011

Summary

This paper extends the literature on modelling healthcare cost data by applying the Generalised Beta of the Second Kind (GB2) distribution to UK data. A quasi-experimental design, estimating models on a subset of the data and evaluating performance on another subset, is used to compare this distribution with its nested and limiting cases. We find that the GB2 may be a useful tool for choosing an appropriate distribution to apply, with the Beta-2 (B2) distribution and Generalised Gamma (GG) distribution performing the best with this dataset.

JEL classification: C1; C5

Key words: Health econometrics; Generalised beta of the second kind; Generalised gamma; Skewed outcomes; Healthcare cost data

---

*Corresponding author: Tel.: +44 1904 32 1411.
E-mail address: james.lomas@york.ac.uk

1 Introduction

Modelling healthcare costs is of primary importance in health economics and health services research for, broadly speaking, two reasons. Firstly, cost-effectiveness analyses require statistical methods of modelling cost data, and secondly such methods are used for risk-adjustment purposes in either insurance schemes (particularly in the US) or devolving budgets to areas or healthcare organisations (in the case of UK). As regards cost-effectiveness analysis, there is some debate as to the usefulness of parametric methods - given the typically smaller sample sizes available for data analysis (Briggs et al., 2005). In risk-adjustment, regression models, applied to large datasets, are used to predict costs for individuals or groups of patients (typically over a long period such as a year), adjusting for healthcare needs using sociodemographic information, such as age and gender as well as morbidity characteristics.

Healthcare cost data are highly non-normal and as such present a number of statistical challenges. Costs cannot be negative, and the distribution is asymmetric and right-hand skewed. In addition, cost data possess long thick right-hand tails with some patients exhibiting extremely large costs, owing to clinical complications, comorbidities or rare and costly events. Furthermore, errors are likely to be heteroskedastic and responses to covariates non-linear. In what follows we focus on modelling expenditures for users of health care and do not consider the problem of zero expenditure observations: for a review of methods to deal with this problem see Jones (2000).

Various approaches have been used in health economics to model cost data. Linear regression of costs is used, for example in Person Based Resource Allocation in the UK (see for example Dixon et al., 2009, 2011). However this method may be sensitive to extreme observations, and may incorrectly assume constant additive responses to covariates. By reducing skewness, transforming the dependent variable may improve performance, with applied work often using a log (or less frequently square root) transformation to reduce the effect of extreme observations (Jones, 2000). Policymakers, however, require estimates on the raw scale leading to the additional challenge of re-transformation. However, such retransformations are likely to be problematic in the case of heteroskedastic errors (Manning et al., 2005; Duan, 1983). Alternatively, it is possible to use inherently non-linear specifications, which have the benefit of estimating effects on the natural scale of costs. These include the GLM family of models as well as exponential conditional mean (ECM) models, often considered for count and duration data, but which can be applied more widely to positive dependent variables. Recently, less parametric approaches including Finite Mixture Models and Conditional Density Estimators (Deb and Burgess, 2003; Gilleskie and Mroz, 2004) have been employed on cost data, which allow the researcher to control for heterogeneity encountered in the data. In principle, these approaches

should offer increased robustness, but be less efficient than fully parametric approaches, since they do not impose an entire distribution upon the observed sample. The alternative is to fit a more flexible parametric distribution believed to approximate the data generating process behind the observed sample.

In this paper, we explore the use of the Generalised Beta of the Second Kind (GB2), and its nested distributions, on health care expenditure data. Mullahy (2009) considers the use of the Singh-Maddala distribution (SM) in order to control the heavy right hand tail of cost data, which is nested within the GB2. Jones (2011) suggests the use of GB2 as part of a comparison of many different methods for modelling US healthcare costs.

The four-parameter GB2 represents an extension to the flexible parametric distributions to be used to model healthcare costs in the health economics literature. Fitting a four-parameter distribution with healthcare costs has precedent with Holly and Pentsak (2006) fitting a Pearson IV distribution in order to address the problems of skewness and kurtosis. Manning et al. (2005) present the three-parameter generalised gamma distribution (GG) as a flexible distribution, which can then be used to select between its one- and two-parameter nested distributions. The GG is a limiting case of the GB2 family of distributions.

The flexibility of Beta-type size distributions has led to many different applications ranging from unemployment duration to fire losses faced by a university, as well as, notably, healthcare costs in the actuarial literature (McDonald and Butler, 1987; Cummins et al., 1990; Sun et al., 2008). Cummins et al. (1990) finds that although the four parameter GB2 fits the data well, the parsimonious one parameter exponential distribution performs only slightly worse. This reaffirms that a flexible distribution is not a substitute for finding the correct distribution for a particular empirical application (Manning et al., 2005). The main application of the Beta-type distributions has been to U.S incomes, with one comprehensive study comparing 15 Gamma- and Beta- type distributions (Bordley et al., 1997). They conclude that GB2 provides the best fit of all models with up to four parameters. Indeed Parker (1999) develops a model from neo-classical principles which predicts that incomes will be distributed according to a GB2 distribution.

This paper contributes to the literature on modelling health care costs by comparing GB2 models including nested distributions, specified using a log-link, and GG models in a quasi-experimental design using UK administrative data (Hospital Episode Statistics). Healthcare cost data from a financial year is divided into an 'estimation' and a 'validation' set. Models are estimated on the former, and evaluated on the latter, with emphasis on bias, accuracy and goodness of fit. Marginal effects, although of interest, are not the primary concern in this paper - given the nature of the methodology used. We evaluate performance at different sample sizes, and present response surfaces as a summary of results following the methodology adopted

by Deb and Burgess (2003).

In general, results show GG as the best fit for this dataset and modelling approach, although Beta-2 (B2) offers the least biased results of those tested. When models are estimated on data with highest levels of costs excluded, the Lomax (L) distribution provides the best fit with SM showing the least bias. We find that relative performance of models changes little with increasing sample size, which is in contrast to the results in Deb and Burgess (2003).

2 Empirical Models

We estimate 11 regression models in total, with each model having either two, three or four fundamental scale and shape parameters to estimate. Seven of the models are estimated with only the scale parameter specified as an exponential function of covariates. These are the four-parameter GB2, its three-parameter nested distributions (Singh-Maddala (SM), Dagum (D), Beta-2 (B2)), its two- parameter nested distributions (Lomax (L), Fisk) and the limiting case of the three-parameter Generalised Gamma (GG). In addition, we estimate four models where the shape parameter is also allowed to vary with covariates. These are two 'heteroskedastic' GG models, where $\ln \sigma$ is a function of covariates (see Manning et al, 2005), as well as two analogous 'heteroskedastic' models for the GB2, allowing $\ln a$ to vary with covariates (Sun et al., 2008)[1]. All models are estimated using maximum likelihood techniques.

2.1 Generalised Beta of the Second Kind

The probability density function (1) and first moment (2) of the GB2 distribution are as follows (McDonald, 1984):

$$f(y) = \frac{a y^{ap-1}}{b^{ap} B(p,q)[1 + (\frac{y}{b})^a]^{(p+q)}} \tag{1}$$

where $B(u,v) = \Gamma(u)\Gamma(v)/\Gamma(u+v)$ is the Beta function, and $\Gamma(\cdot)$ is the Gamma function.

$$E(y) = b\left[\frac{\Gamma(p+\frac{1}{a})\Gamma(q-\frac{1}{a})}{\Gamma(p)\Gamma(q)}\right] \tag{2}$$

b is a scale parameter, and a, p and q are shape parameters. Kleiber and Kotz (2003) describe a as influencing the kurtosis of the distribution ('thinness of the tails'), with p and q influencing the degree of skew of the distribution.

[1] See Sections 2.1 and 2.3 for explanation of parameters.

In Figure 1, we present probability density functions of the GB2, setting $b = 1$, and varying values of a, p and q - taken from Kleiber and Kotz (2003) and formatted.

With the Stata module `gb2fit` developed by Jenkins (2009), it is possible to fit the generalised beta of the second kind distribution to outcome data, specifying the distribution parameters as either constant scalars or linear functions of covariates. This code is employed by Jones (2011) to estimate the GB2 on US cost data from the MEPS dataset, allowing the value of b to vary linearly with covariates.

Here we specify $b = \exp(x_i'\beta)$ and treat the remainder of parameters as scalars giving a mean proportional to an exponential function of the covariates. This ensures that predictions are always positive and has a precedent in the costs literature.[2]

2.2 Nested Distributions and Limiting Cases

We estimate each of the nested distributions of GB2, with the exception of the Inverse-Lomax distribution, which is theoretically unable to produce estimates of the mean, and is rarely referred to in the modelling literature (Kleiber and Kotz, 2003). Kleiber and Kotz (2003) give a thorough account of these models including the associated probability density functions and moments of the distributions. We note that SM has been discussed in the health care costs literature as a method to deal with the heavy-tailed nature of cost data (Mullahy, 2009). See Figure 2 for diagram showing relationships between the GB2 and its nested and limiting cases.

[2]Other link functions would be possible such as a square root link, choosing between different forms could then be aided by testing Pearson correlation coefficients and using Pregibon's link test (Pregibon, 1980).

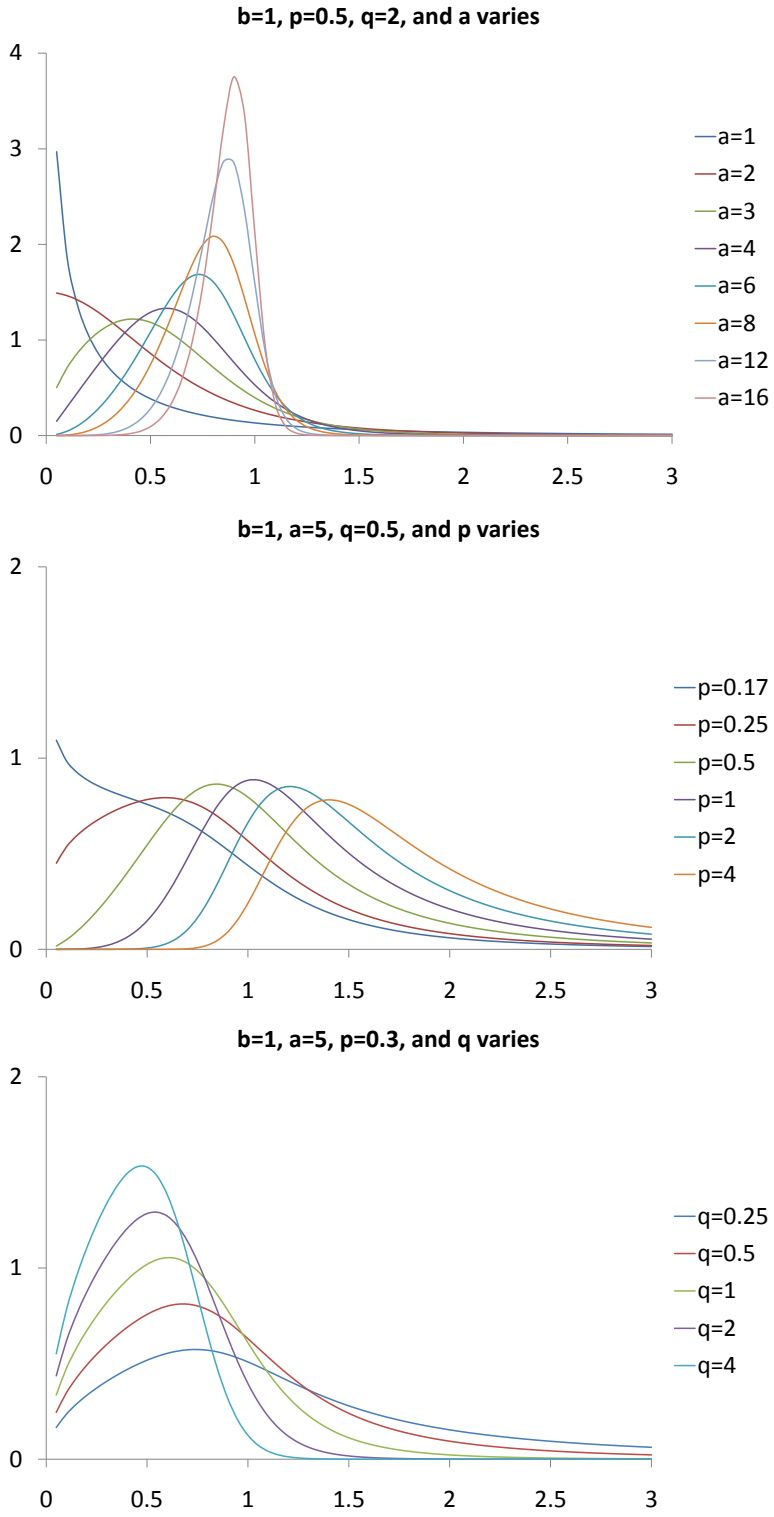
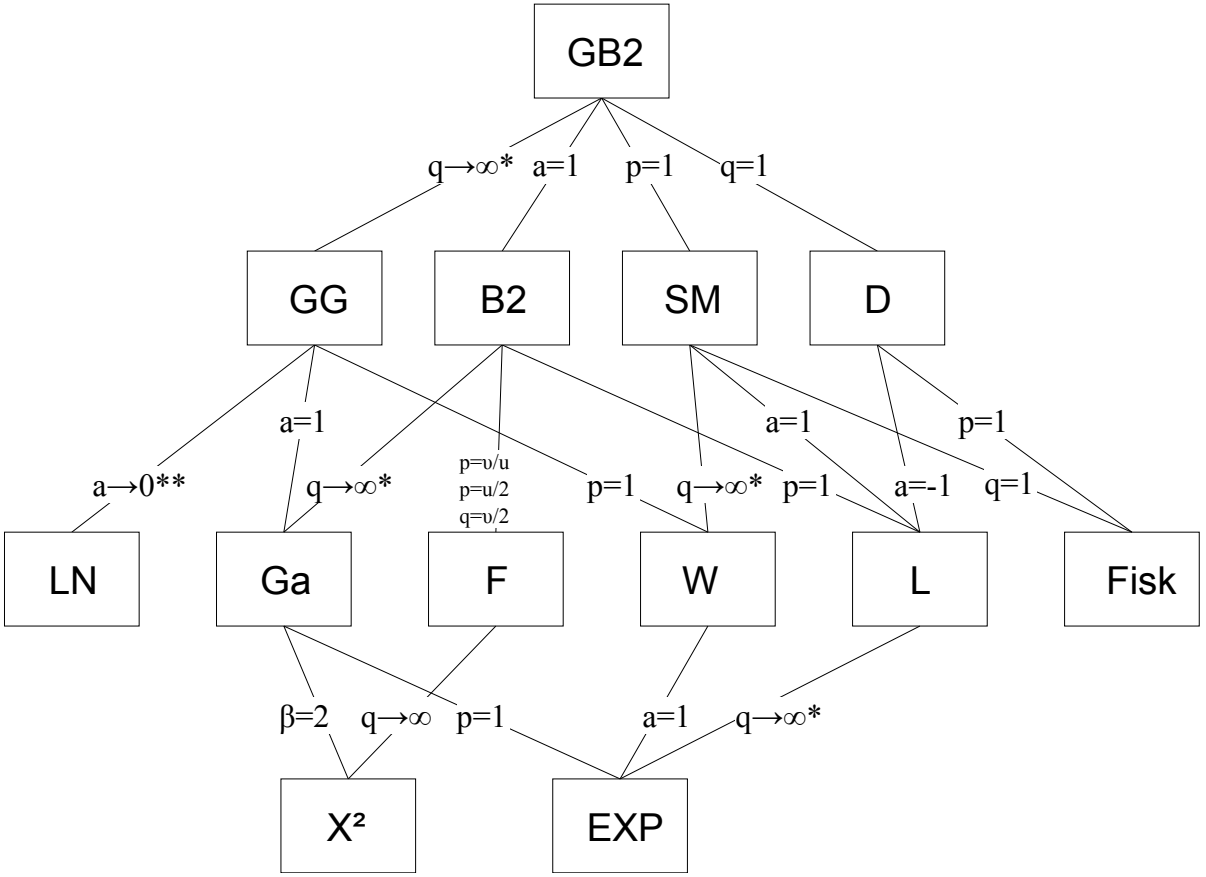


Figure 1: Probability density functions for GB2 with different parameter values



*with $b = \beta q^{1/a}$

**with $b = \left(\sigma^2 a^2\right)^{1/a}$, $p = \left(a\mu + 1\right)/\sigma^2 a^2$

LN=Log-normal distribution, Ga=Gamma distribution, F=F-distribution, W=Weibull distribution, EXP=Exponential distribution, X^2=Chi-square distribution

Figure 2: (adapted from McDonald and Xu, 1995) shows the relationship between Beta-type and Gamma-type size distributions employed in this paper

The GG distribution[3] is a limiting case of the GB2 distribution, where $b = q^{1/a}\beta$ and $q \to \infty$, such that:

$$f(y) = \frac{ay^{ap-1}e^{-(y/\beta)^a}}{\beta^{ap}\Gamma(p)} \tag{3}$$

If $\beta \sim InvGG(a, b, q)$, then y is distributed according to a GB2 distribution. That is:

$$GG(a, \beta, p) \bigwedge_{\beta} InvGG(a, b, q) = GB2(a, b, p, q) \tag{4}$$

GB2 can be said to be a "scale mixture of generalized gamma distributions with inverse generalized gamma weights" (Venter, 1983; McDonald and Butler, 1987, cited in Kleiber and Kotz, 2003). Indeed all of the nested distributions of the GB2 distribution can be interpreted as a scaled mixture distribution, as shown in Table 1 adapted from McDonald and Butler (1987).

Distribution	**Structural distribution**	**Mixing distribution**
$GB2(y; a, b, p, q)$	$GG(y; a, \beta, p)$	$InvGG(\beta; a, b, q)$
$SM(y; a, b, q)$	$Wei(y; a, \beta)$	$InvGG(\beta; a, b, q)$
$D(y; a, b, p)$	$GG(y; a, \beta, p)$	$InvWei(\beta; a, b)$
$B2(y; b, p, q)$	$Ga(y; \beta, p)$	$InvGa(\beta; b, q)$
$L(y; b, q)$	$Exp(y; \beta)$	$InvGa(\beta; b, q)$
$Fisk(y; a, b)$	$Wei(y; a, \beta)$	$InvWei(\beta; a, b)$

Table 1: Beta-type distributions as scaled mixture distributions

2.3 'Heteroskedastic' Models

We extend the modelling of the distributions to account for heteroskedasticity, by specifying one of the shape parameters as a function of covariates.

Manning et al. (2005) parameterise the GG as follows:

$$f(y) = \frac{\gamma^{\gamma}}{\sigma y\sqrt{\gamma}\Gamma(\gamma)} \exp[z\sqrt{\gamma} - u] \tag{5}$$

where $\gamma = |\kappa|^{-2}$, $z = sign(\kappa)\{\ln(y) - \mu\}/\sigma$, and $u = \gamma \exp(|\kappa| z)$. They suggest modelling $\ln \sigma$ as a function of covariates when κ is small. We run two 'heteroskedastic' specifications for GG where the choice of covariates entering into the function differs, regardless of the magnitude of κ. We also run GB2 models where $\ln a$ is a function of covariates, as an analogous extension to the model, and to illustrate the ability of the GB2 distribution to model covariates as influencing both a scale and shape parameter.

[3] GG is estimated using the parameterisation given in Cox et al. (2007), consistent with Stata `streg` command.

3 Data and Choice of Variables

We use individual level data from England on the use of healthcare services to assess the comparative performance of the various regression models. Individual-level information on healthcare utilisation is taken from the Hospital Episode Statistics (HES) for the financial year 2007-2008. HES is a large administrative dataset administered by the NHS Information Centre, containing information on all inpatient episodes, outpatient visits and A&E attendances for all patients on admitted to English NHS hospitals. Information is collected via medical records.

The expenditure variable used throughout is individual patient annual NHS hospital expenditure for all spells finishing in the financial year 2007-2008. Costs generated by inpatient utilisation of NHS facilities were included using the data on reference cost tariffs from 2008-2009.[4] These are then applied to the most expensive episode within the spell of the patient. All episodes falling within the financial year are summed for each patient. Costs for mental and maternity health spells together with private sector spells were excluded.[5] For purposes of our analysis we focus on positive expenditures only, ignoring non-users of services (leaving 600,000 observations).

Table 2 indicates the challenges of modelling cost data: the observed costs are heavily right-hand skewed, with the mean far in excess of the median. They are also highly kurtotic, implying that Beta-type distributions may be useful in trying to model the thickness of the tails. Figure 3 displays a histogram for raw and transformed data, including a Epanechnikov kernel plot. Even after log transformation, the distribution exhibits right-hand skewness.

We specify a set of covariates to model costs using those set out in the risk adjustment literature used in the UK for resource allocation purposes. Hence we adopt a parsimonious model based on the Person-Based Resource Allocation model (PBRA) currently used to allocate healthcare resources across general practices in England. We include sociodemographic information in the form of age and gender, as well as morbidity markers (based on ICD classification). This specification mirrors common practice in the comparative literature on econometric approaches to healthcare cost data, for example Deb and Burgess (2003), which uses morbidity markers in the form of Diagnostic Cost Groups.

[4]For the purposes of this study outpatient visits were excluded.

[5]The dataset was constructed to model the determinants of individual healthcare use as part of a wider project considering the allocation of NHS resources to primary care providers. Data for mental healthcare is poor since a lot of care is undertaken in the community (and hence not recorded in HES), and also since healthcare budgets for this type of care are constructed using separate formulae. Maternity services are excluded since they are unlikely to be heavily determined by morbidity characteristics.

	Full sample
N	600,000
Mean	£2,609
Median	£1,126
Standard deviation	5,011
Skewness	11.82
Kurtosis	259.18
Maximum	£206,893
99th percentile	£18,886
95th percentile	£8,980
90th percentile	£6,033
75th percentile	£2,722
25th percentile	£610
10th percentile	£446
5th percentile	£406
1st percentile	£347
Minimum	£217

Table 2: Descriptive statistics for hospital costs

In order to control for age and gender, we use a cubic term for age together with interactions with gender, as well as a gender dummy. The average age in the full sample is just under 51 years of age and around 54 percent are female.

Morbidity information is available through the HES dataset, adapted from the ICD10 chapters (WHO, 2007) - see Appendix A for further details. The morbidity indicators are coded 1 if the individuals had one or more hospital spells in the financial year with any diagnosis in the relevant subset of ICD10 chapters. Accordingly, the indicators do not indicate the severity of the condition, merely its presence or absence.

'Heteroskedastic' models were constructed allowing one of the shape parameters to vary according to two sets of covariates. One allowed the shape parameter to vary with age, and the other set included the number of morbidity characteristics exhibited by the individual. Whilst it is, in principle, possible to estimate the shape parameter as a function of all covariates, preliminary work found that models with fewer covariates had better convergence performance.

4 Methodology

Comparative studies on US health care cost data regressions fall into two broad categories: those using actual expenditures (Deb and Burgess, 2003; Veazie et al., 2003; Buntin and Zaslavsky, 2004; Basu et al., 2006;

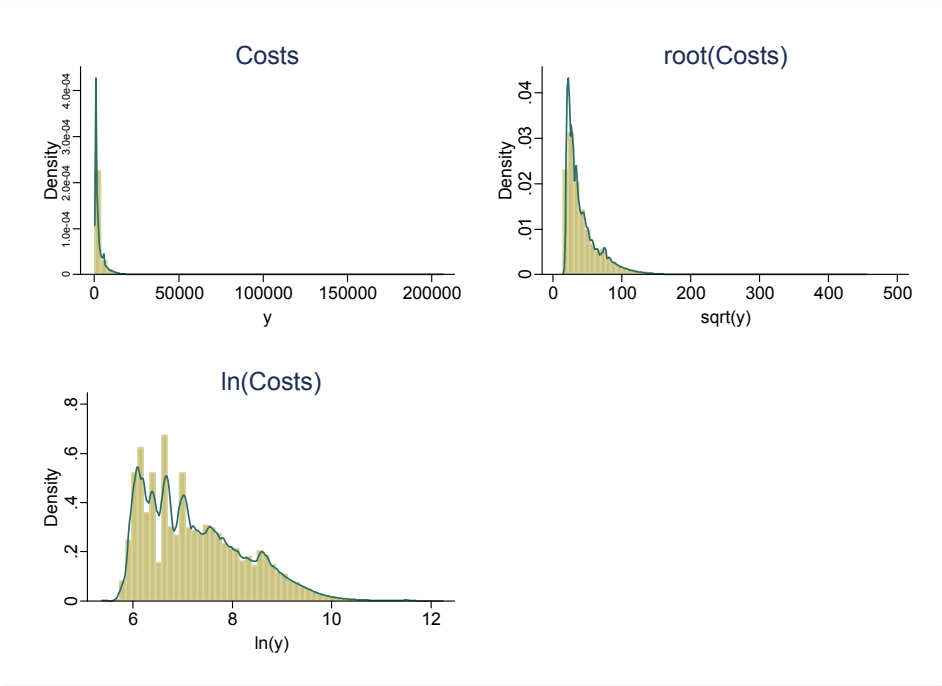


Figure 3: Histogram plots of costs

Hill and Miller, 2010) and those that synthesise expenditures using known distributional forms (Manning et al., 2005; Basu et al., 2004). A few key lessons emerge from this literature. Firstly, there is no one model that dominates in all respects and there seems to be a tradeoff between bias and precision (Veazie et al., 2003). Secondly, that the preferred model is likely to vary with the sample size of data on which the model is estimated (Deb and Burgess, 2003) and will also vary across datasets (Hill and Miller, 2010). It has also been noted that a more flexible model is not necessarily an adequate replacement for the correct model (Manning et al., 2005).

4.1 Quasi-Experimental Design

Our study fits in the category of those using actual expenditures. We exploit the large amount of observations that are available through the HES data by using a quasi-experimental design similar to Deb and Burgess (2003). The full set of observations (600,000) is randomly divided into two equally sized sub-populations: an 'estimation' (300,000) and a 'validation' set (300,000). From within the 'estimation' set we randomly draw, 100 times with replacement, samples of size N_s ($N_s \in 5,000; 10,000; 50,000; 100,000$). The model is estimated on the sample and performance evaluated on both the 'estimation' sample and the full 'validation' set. To ensure the results are replicable, we set a seed number for splitting the dataset into two sets

and randomly drawing samples.

In this way the design follows Monte Carlo principles of resampling, only using actual cost data as opposed to hypothetical known distributions. Since it is known that costs do not follow a single parametric distribution, this may be preferable - assuming that we have sufficient data to represent all the features of the distribution of healthcare costs. In addition we do not influence, a priori, our results towards any one distribution. Furthermore, evaluating on observations not used in estimation guards against over-fitting and embodies the predictive nature of resource allocation budgeting. Our design means that we do not know the true marginal effects of each covariate, as would be the case using synthesised expenditures. This is often the focus of comparative work. In this paper, however, we concentrate on the ability of the models to predict costs for each observation in terms of predictive bias, accuracy and goodness of fit.

4.2 Evaluation of Performance

4.2.1 Estimation Sample Tests

As with other flexible models, one benefit is the ability to choose between the more restrictive nested and limiting case distributions. In order to evaluate competing models we test the restrictions imposed by each nested model of the GB2 distribution using a Wald test and report rejection rates at the 5, 1 and 0.1 percent significance levels (see Table 3 and Appendix B). These are carried out on all samples where the GB2 successfully estimates, and rejection rates reported as a percentage of successful estimations. To compare beta-type models with the GG models (a limiting case and not a linear restriction of a parameter), we use Akaike and Bayesian Information Criteria. As a summary statistic we calculate the average of the log likelihood of models estimated over all samples where the models estimate successfully. AIC and BIC are then calculated, imposing the appropriate penalty upon the summary log likelihood, given the number of coefficients estimated.

We also graph average prediction error (MPE - see next section) by deciles of predicted level of costs, analogous to Hosmer-Lemeshow and Pearson correlation coefficient tests, which allows us to visually assess any patterns in bias by decile of predicted level of cost. In the literature, this kind of assessment is used to decide between link functions; for our purposes, the link function is set as a log-link for all models for demonstrative purposes, and so this interpretation is more about the appropriateness of the shape parameters in influencing the conditional means of each competing distribution.[6]

[6]The functional form of the shape parameters will determine the efficiency of the estimators, but our focus is purely on predictive performance.

4.2.2 Validation Sample Tests

In health economics, the estimated conditional mean cost is usually the most useful to policymakers (Arrow and Lind, 1970). This can also be the case in risk-adjustment formulae, where the goal is often to estimate the expected cost of an individual to the healthcare provider over a certain period of time (Rice and Smith, 2002). Accordingly, we use our models to estimate predicted costs over the year for individuals ($\hat{y}_i = E(y_i|x_i)$) and evaluate performance on metrics designed to reflect the bias (mean prediction error, MPE), accuracy (mean absolute prediction error, MAPE) and goodness of fit (R^2 and root mean square error, RMSE) of these predictions. MPE can be thought of as measuring the accuracy of predictions at an aggregate level, while MAPE is a measure of the accuracy of individual predictions. In addition, we evaluate the variability of bias across replication samples (absolute deviations of mean prediction error, ADMPE). Note that R^2 is calculated by an auxiliary regression of actual levels of costs on predicted values. [7] These are evaluated on the full 'validation' set. Formulae for calculating these metrics are provided below.[8] Only samples where all 11 models are successfully estimated are included for evaluation, and model performance according to each criterion is calculated as an average over all included samples.

$$MPE_{msr} = \frac{\sum (y_i - \hat{y}_i)}{N_s} \tag{6}$$

$$MAPE_{msr} = \frac{\sum |y_i - \hat{y}_i|}{N_s} \tag{7}$$

$$RMSE_{msr} = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{N_s}} \tag{8}$$

$$R^2_{msr} = 1 - \frac{\sum (y_i - (\alpha_{AUX} + \beta_{AUX}\hat{y}_i))^2}{\sum (y_i - \hat{y}_i)^2} \tag{9}$$

$$ADMPE_{msr} = \left| MPE_{msr} - \frac{\sum_{r=1}^{N_r} MPE_{msr}}{N_r} \right| \tag{10}$$

In order to understand how the different model specifications performed, it is necessary to evaluate predictions at different levels of cost. We assess MPE and MAPE for deciles for levels of actual costs, in order to evaluate the degree to which flexible parametric models can model heavily right hand tailed data, which partly motivates their usage.

[7] In equation 9 coefficients from the 'auxiliary' regression are denoted with $_{AUX}$ subscript.

[8] Where $_m$ denotes the model used, $_s$ the sample size used, and $_r$ the replication.

Following Deb and Burgess (2003) we construct response surfaces. These produce polynomial approximations to the relationship between the predicted values and the sample size of the experiment, N_s. For our purposes, we estimate the following regression for each model and for each metric of performance (illustrated below for the mean prediction error).

$$MPE_{msr} = \alpha_m + \beta_m \left(\frac{1}{N_s}\right) + u_{msr} \tag{11}$$

We specify the relationship between MPE and the inverse of the sample size, to reflect increased accuracy with increasing sample size. In particular, the value of α_m represents the value of MPE to which the model approaches asymptotically with increasing sample size. For the metrics that cannot be negative, we use the log function of the value as the dependent variable, for example in the case of mean absolute prediction error, we estimate:

$$\ln\left(MAPE_{msr}\right) = \alpha_m + \beta_m \left(\frac{1}{N_s}\right) + u_{msr} \tag{12}$$

With the log specification, differences in estimates are to be interpreted as percentage differences, as opposed to absolute differences.

5 Results and Discussion

5.1 Estimation Sample Results

Where GB2 was estimated[9], we tested the restrictions required for each of the nested Beta-type size distributions using a Wald test. In Table 3 we present the percentage of replicates where the null hypothesis was rejected at a 5% level of significance. Accordingly a higher percentage indicates greater evidence against the nested model being appropriate.

	Sample size (number of replications)			
Nested model	**5,000 (83)**	**10,000 (85)**	**50,000 (97)**	**100,000 (100)**
SM (1 restriction)	75%	100%	100%	100%
D (1 restriction)	31%	92%	100%	100%
B2 (1 restriction)	46%	68%	100%	100%
L (2 restrictions)	99%	100%	100%	100%
Fisk (2 restrictions)	60%	100%	100%	100%

Table 3: Results of tests on nested model restrictions (percentage rejected at 5% significance level)

Note that with increasing sample size, there is a greater probability that any hypothesis will be rejected, as demonstrated clearly here - with all

[9]See Appendix B for discussion of computational performance and convergence.

restrictions rejected with sample size 50,000 or over. The restrictions are rejected least often with $q = 1$ (Dagum) followed by $a = 1$ (B2) at a sample size of 5,000. With samples of size 10,000 we reject the B2 restriction the least times followed by Dagum. Results for significance levels 1% and 0.1% are presented in Appendix B. While, naturally rejection rates are smaller, these results follow a similar ordering to Table 3.

		Sample size			
		5,000	**100,00**	**50,000**	**100,000**
GB2H(morb)	**AIC**	83781	167527	837232†	1674147†
	BIC	84015	167787	837550‡	1674490‡
GB2H(age)	**AIC**	83776†	167521†	837265	1674210
	BIC	84011	167780	837582	1674553
GB2	**AIC**	84016	168013	839671	1679039
	BIC	84244	168265	839979	1679372
GGH(morb)	**AIC**	83780	167526	837246	1674171
	BIC	84008‡	167778‡	837555	1674504
GGH(age)	**AIC**	83781	167528	837318	1674326
	BIC	84009	167781	837627	1674659
GG	**AIC**	84032	168046	839849	1679405
	BIC	84253	168291	840149	1679728
SM	**AIC**	84061	168103	840131	1679963
	BIC	84282	168349	840431	1680287
D	**AIC**	84022	168026	839740	1679178
	BIC	84244	168271	840040	1679501
B2	**AIC**	84018	168019	839704	1679108
	BIC	84240	168264	840004	1679431
L	**AIC**	85830	171639	857911	1715535
	BIC	86045	171877	858202	1715849
Fisk	**AIC**	84144	168278	840995	1681704
	BIC	84359	168516	841286	1682017

GB2H and GGH represent 'heteroskedastic' versions of the GB2 and GG, where the shape parameter is allowed to vary by the variable named in the subsequent brackets. 'morb' is the number of morbidity characteristics recorded for the patient.
† lowest AIC, ‡ lowest BIC

Table 4: Values for each model's average AIC and BIC at each sample size

Results presented in Table 4 show that the more flexible models perform well according to AIC and BIC. On the whole, models with more estimated parameters perform better than those with fewer. As such, the 2-parameter

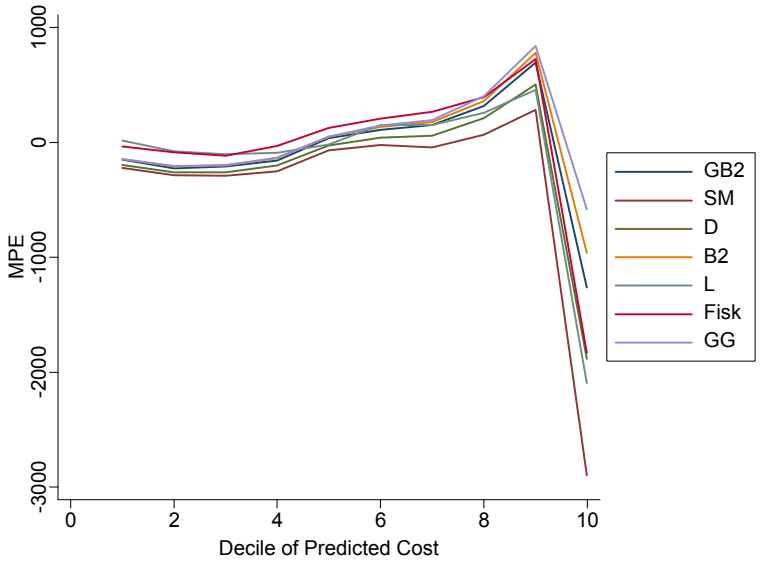


Figure 4: Mean prediction error for each model by each decile of predicted level of cost

L and Fisk perform poorly. Of the 3-parameter models, B2 performs the best at all sample sizes. The 4-parameter GB2 performs better than B2 at all sample sizes according to average AIC, and at samples of size 50,000 and 100,000 with average BIC. The 'heteroskedastic' models perform well and, even after imposing a penalty for extra estimated coefficients, are judged to fit better than models where only one parameter is allowed to vary with covariates. This demonstrates potential for these models, even though the 'heteroskedastic' version of the GG model is estimated in all cases regardless of whether κ is small or not, where Manning et al. (2005) suggests using this version in the event of small κ.

Figure 4 plots MPE against deciles of predicted costs on the estimation set. As such, it represents a visual attempt to determine structural bias in the model - as tested statistically using Pearson correlation coefficients and Pregibon link tests. In general the models appear to follow a similar pattern including large over-estimations on average in the highest decile of predicted costs, although this is exhibited less in the case of the GG model.

5.2 Validation Sample Results

Relative performance in terms of R^2, RMSE, MAPE and MPE varies little with sample size. Hence, for simplicity, we present the results for the sample size 5,000 only, with the remainder of the results found in Appendix B. We only include samples in the analysis for which all models converged.

As expected, results in Table 5 imply a trade-off between bias, as measured by MPE, and accuracy, as measured by MAPE, since no model performs best in both of these dimensions. We display the best performing values in bold font.

	$\mathbf{R^2}$	**RMSE**	**MAPE**	**MPE**	**ADMPE**
GB2	0.18244	4814.76	1790.71	-64.87	57.31
SM	0.17404	5184.83	1937.76	-370.06	65.74
D	0.17933	4935.73	1850.51	-196.90	58.62
B2	0.18455	4757.09	1766.23	**-9.13**	52.60
L	0.18689	4948.54	1805.03	-132.97	59.93
Fisk	0.17580	4980.72	1782.17	-35.34	53.89
GG	**0.18783**	**4687.46**	**1745.31**	41.81	**52.22**
GB2H(age)	0.16140	5338.99	1967.19	-373.38	92.46
GB2H(morb)	0.02526	702526.88	22188.24	-20908.09	11527.68
GGH(age)	0.16599	5128.15	1904.60	-266.78	75.30
GGH(morb)	0.03494	1002436.90	36032.56	-34699.41	38362.88

Table 5: Results of model performance, when all converged, sample size 5,000

B2 gives the least biased estimates, overestimating by £9.31 (0.4% of the mean of the full sample) on average over the replications. In this regard, SM performs the worst, except for the 'heteroskedastic' models. GG gives the most accurate results with B2 the second most accurate and SM the least accurate (excluding the 'heteroskedastic' models). In the case of GG, the mean absolute prediction error, averaged over replications, is £1745.31 (66.9% of the mean of the full sample); for SM this is £1937.76 (74.3% of the mean of the full sample). Once again SM performs the worst. In terms of goodness of fit, R^2 and RMSE, GG performs the best; B2 also performs well. It is worth highlighting that D does not perform especially well according to its performance on the tests using the validation sample, despite the strong performance in tests relating to its log-likelihood. In addition, the results for the 'heteroskedastic' models are very poor. This indicates that these models are not appropriate for predicting costs with this dataset.

In order to investigate the sensitivity of our models to extremely high costs, we trim the highest 5% of costs from the estimation dataset. Convergence performance is poor for 'heteroskedastic' models, and so these results are omitted. Results are shown in Table 6 with best performing values displayed in bold font.

In terms of goodness of fit, L achieves the highest R^2 and lowest RMSE. All models now underestimate costs on average, and relative performance is different to results with the untrimmed estimation dataset. B2 is no longer the least biased and SM has the lowest mean prediction error (having had

	R^2	RMSE	MAPE	MPE	ADMPE
GB2	0.20099	4507.00	1622.79	506.22	37.10
SM	0.18607	4581.56	1687.82	**240.66**	46.44
D	0.19122	4534.22	1669.74	310.65	37.62
B2	0.19786	4512.64	1629.16	473.03	**33.17**
L	**0.20496**	**4495.52**	**1608.57**	490.37	35.78
Fisk	0.18874	4556.86	1635.37	401.85	41.09
GG	0.18821	4616.42	1656.22	318.87	43.60

Table 6: Results of model performance with trimmed estimation set, when all converged, sample size 5,000

the largest MPE in Table 5).

Graphs in Figures 5 and 6 show the performance metrics for each decile of the actual level of costs of the validation set. This enables us to see how well our models peform on out-of-sample observations by each decile. We do not include 'heteroskedastic' models.

Looking at the mean prediction error, we can see which models predict well on average for each decile of costs being considered. It appears as though there is a consistent pattern where the highest costs are underpredicted by all models, and the lowest 8 deciles are overpredicted by all models on average. The models (e.g. SM) that predict higher costs in general are the most biased in overpredicting in low deciles, but then have the best performance in the highest cost decile. The converse applies to models that better predict lower costs in general, such as GB2. The ranking of models' performance over deciles changes somewhat - it is worth noting that GG does particularly badly in that it underpredicts costs in the last decile.

In terms of fitting the individual expenditures, we present the mean average prediction error for each decile of actual costs. It is clear that the models which perform the best on average in terms of MPE in the highest decile of costs actually perform worse in terms of the mean absolute prediction error. This suggests models, such as SM, whilst predicting high costs well on average, do a relatively poor job of predicting individual costs within the highest decile.

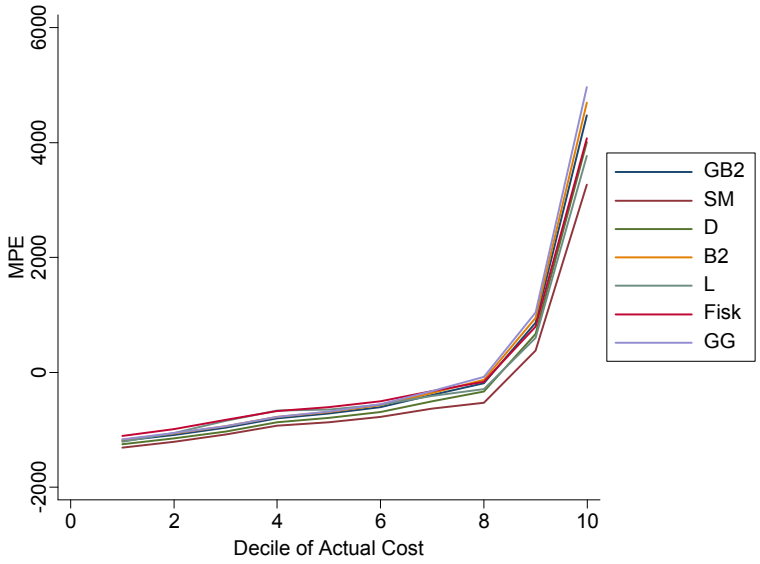


Figure 5: Mean prediction error for each model by each decile of level of cost

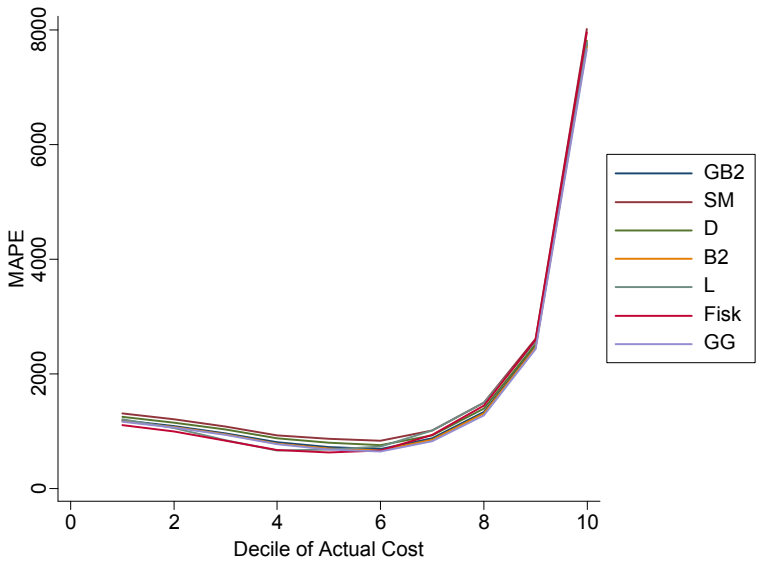


Figure 6: Mean absolute prediction error for each model by each decile of level of cost

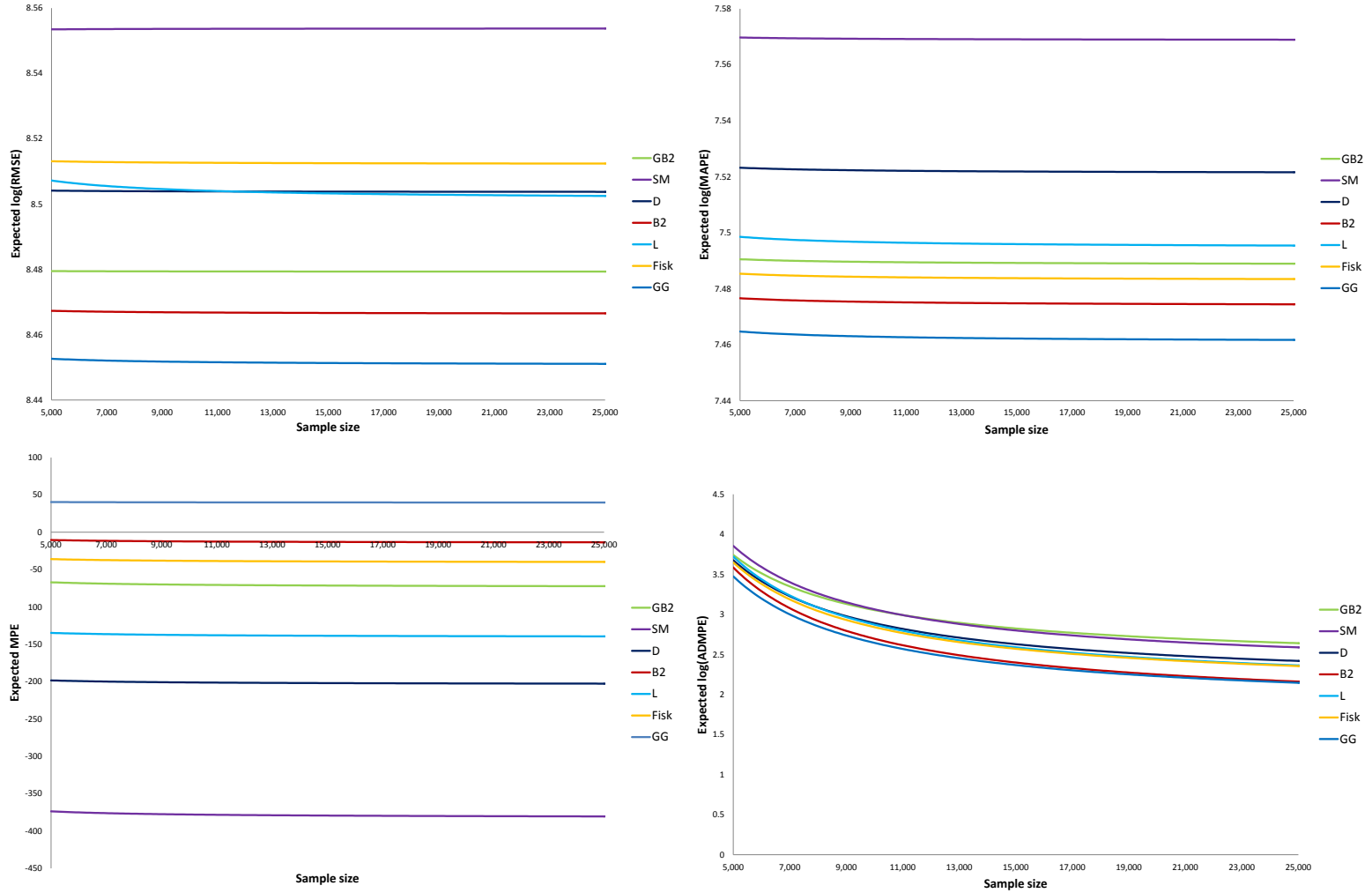


Figure 7: Reponse surfaces for log(RMSE), log(MAPE), log(ADMPE), MPE (clockwise from top left) against sample size, constructed evaluating performance on 'validation' set

As outlined in the methodology, we conduct the procedure with different sample sizes. We find that performance of the models is largely unaffected by sample size, with the same models performing consistently better in terms of predicted costs. Some improvement is observed for all models with LADMPE, suggesting that bias becomes steadier at a sample size of around 10,000. However, on the whole, these results suggest that these models are as appropriate at smaller sample sizes (from 5,000) as larger ones. See Figure 7 for the response surfaces. Coefficient estimates for the constant term for all metrics of performance were all significantly different from zero at the 5 percent level implying that, as the sample size approaches infinity, the measures of performance do not converge to zero.

6 Conclusions and Extensions

We estimate the GB2 on UK healthcare data using maximum likelihood estimation, specifying the conditional mean as an exponential function of covariates, and evaluated its (and its nested and limiting case models') performance using a quasi-experimental design. It is possible to estimate other link functions with this type of models and should be investigated as part of future work.

The results suggest that there may be potential for beta-type distributions in predicting individual healthcare costs: in particular B2 exhibits less bias than other models, without losing much accuracy. In a spirit similar to Manning et al. (2005), the GB2 could be used as a flexible distribution to select among competing distributions it nests.

A further issue that needs to be investigated is the sensitivity of results to the specification of the mean function, especially since this seems to have a bearing upon the convergence performance of certain models, e.g. 'heteroskedastic' specifications of the GB2. Furthermore, it is reasonable to question whether it is appropriate to test performance across estimators when each distribution is estimated on the same set of covariates, although this would seem appropriate when comparing nested and limiting cases of models each using the same link function[10].

Acknowledgements

The authors gratefully acknowledge funding from the Economic and Social Research Council (ESRC) under grant reference RES-060-25-0045. We are grateful to John Mullahy for insightful comments, to members of the Health, Econometrics and Data Group (HEDG) at the University of York for useful discussions, as well as to participants of the HEDG seminar series, especially John Forbes, for helpful suggestions.

[10]Will Manning, cited in Jones, 2011.

References

Arrow KJ, Lind RC. 1970. Uncertainty and the evaluation of public investment decisions. *The American Economic Review* **60**: 364–378.

Basu A, Arondekar BV, Rathouz PJ. 2006. Scale of interest versus scale of estimation: comparing alternative estimators for the incremental costs of a comorbidity. *Health Economics* **15**: 1091–1107.

Basu A, Manning WG, Mullahy J. 2004. Comparing alternative models: log vs cox proportional hazard? *Health Economics* **13**: 749–765.

Bordley R, McDonald J, Mantrala A. 1997. Something new, something old: Parametric models for the size of distribution of income. *Journal of Income Distribution* **6**: 91–103.

Briggs A, Nixon R, Dixon S, Thompson S. 2005. Parametric modelling of cost data: some simulation evidence. *Health Economics* **14**: 421–428.

Buntin MB, Zaslavsky AM. 2004. Too much ado about two-part models and transformation?: Comparing methods of modeling medicare expenditures. *Journal of Health Economics* **23**: 525–542.

Cox C, Chu H, Schneider MF, Munoz A. 2007. Parametric survival analysis and taxonomy of hazard functions for the generalized gamma distribution. *Statistics in Medicine* **26**: 4352–4374.

Cummins JD, Dionne G, McDonald JB, Pritchett BM. 1990. Applications of the GB2 family of distributions in modeling insurance loss processes. *Insurance: Mathematics and Economics* **9**: 257–272.

Deb P, Burgess JF. 2003. A quasi-experimental comparison of econometric models for health care expenditures. *Hunter College Department of Economics Working Papers* **212**.

Dixon J, Asaria P, Georghiou T, Billings J, Gravelle H, Martin S, Rice N, Smith P, Wennberg D, DeLorenzo M, Siegal M, Russell R, Filipova N. 2009. Developing a person based resource allocation formula for general practices in england. *Report to the Department of Health* .

Dixon J, Gravelle H, Smith P, Bardsley M, Martin S, Rice N, Georghiou T, Dushieko M, Sanderson C, Billings J, Delorenzo M. 2011. Calculating budgets for commissioning general practices in england. *British Medical Journal* In Press.

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

Gilleskie DB, Mroz TA. 2004. A flexible approach for estimating the effects of covariates on health expenditures. *Journal of Health Economics* **23**: 391–418.

Hill SC, Miller GE. 2010. Health expenditure estimation and functional form: applications of the generalized gamma and extended estimating equations models. *Health economics* **19**: 608–627.

Holly A, Pentsak Y. 2006. Maximum likelihood estimation of the conditional mean $e(y|x)$ for skewed dependent variables in four-parameter families of distribution Technical report, Institute of Health Economics and Management (IEMS), University of Lausanne, Switzerland.

Jenkins S. 2009. GB2FIT: stata module to fit generalized beta of the second kind distribution by maximum likelihood Statistical software components, Boston College Department of Economics.

Jones AM. 2000. Health econometrics. In Culyer AJ, Newhouse JP (eds.) *Handbook of Health Economics*, volume Volume 1, Part 1. Elsevier, 265–344.

Jones AM. 2011. Models for health care. In Clements MP, Hendry DF (eds.) *Oxford Handbook of Economic Forecasting*. Oxford University Press.

Kleiber C, Kotz S. 2003. *Statistical size distributions in economics and actuarial sciences*. Wiley-IEEE.

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

McDonald JB. 1984. Some generalized functions for the size distribution of income. *Econometrica* **52**: 647–663.

McDonald JB, Butler RJ. 1987. Some generalized mixture distributions with an application to unemployment duration. *The Review of Economics and Statistics* **69**: 232–240.

McDonald JB, Xu YJ. 1995. A generalization of the beta distribution with applications. *Journal of Econometrics* **69**: 427–428.

Mullahy J. 2009. Econometric modeling of health care costs and expenditures: a survey of analytical issues and related policy considerations. *Medical Care* **47**: S104–108.

Parker SC. 1999. The generalised beta as a model for the distribution of earnings. *Economics Letters* **62**: 197–200.

Pregibon D. 1980. Goodness of link tests for generalized linear models. *Journal of the Royal Statistical Society* **29**: 14–23.

Rice N, Smith PC. 2002. Strategic resource allocation and funding decisions. In Mossialos E, Dixon A, Figueras J, Kutzin J (eds.) *Funding health care: options for Europe*. Open University Press.

Sun J, Frees EW, Rosenberg MA. 2008. Heavy-tailed longitudinal data modeling using copulas. *Insurance: Mathematics and Economics* **42**: 817–830.

Veazie PJ, Manning WG, Kane RL. 2003. Improving risk adjustment for medicare capitated reimbursement using nonlinear models. *Medical Care* **41**: 741–752.

Venter GG. 1983. Transformed beta and gamma distributions and aggregate losses. In *Proceedings of the Casualty Actuarial Society*, volume 70. 156–193.

WHO. 2007. International statistical classification of diseases and related health problems, 10th revision, version for 2007 .

Appendix A

We use the variables shown in Table 7 to construct our regression models. They are based on the ICD10 chapters, which are given in Table 8.

Variable name	Variable description
epiA	Intestinal infectious diseases, Tuberculosis, Certain zoonotic bacterial diseases, Other bacterial diseases, Infections with a predominantly sexual mode of transmission, Other spirochaetal diseases, Other diseases caused by chlamydiae, Rickettsioses, Viral infections of the central nervous system, Arthropod-borne viral fevers and viral haemorrhagic fevers
epiB	Viral infections characterized by skin and mucous membrane lesions, Viral hepatitis, HIV disease, Other viral diseases, Mycoses, Protozoal diseases, Helminthiases, Pediculosis, acaiasis and other infestations, Sequelae of infectious and parasitic diseases, Bacterial, viral and other infectious agents, Other infectious diseases
epiC	Malignant neoplasms
epiD	In situ neoplasms, Benign neoplasms, Neoplasms of uncertain or unknown behaviour and III
epiE	IV
epiF	V
epiG	VI
epiH	VII and VIII
epiI	IX
epiJ	X
epiK	XI
epiL	XII
epiM	XIII
epiN	XIV
epiOP	XV and XVI
epiQ	XVII
epiR	XVIII
epiS	Injuries to the head, Injuries to the neck, Injuries to the thorax, Injuries to the abdomen, lower back, lumbar spine and pelvis, Injuries to the shoulder and upper arm, Injuries to the elbow and forearm, Injuries to the wrist and hand, Injuries to the hip and thigh, Injuries to the knee and lower leg, Injuries to the ankle and foot
epiT	Injuries involving multiple body regions, Injuries to unspecified part of trunk, limb or body region, Effects of foreign body entering through natural orifice, Burns and Corrosions, Frostbite, Poisoning by drugs, medicaments and biological substances, Toxic effects of substances chiefly nonmedicinal as to source, Other and unspecified effects of external causes, Certain early complications of trauma, Comlications of surgical and medical care, not elsewhere classified, Sequelae of injuries, of poisoning and of other consequences of external causes
epiU	XXII
epiV	Transport accidents
epiW	Falls, Exposure to inanimate mechanical forces, Exposure to animate mechanical forces, Accidental drowning and submersion, Other accidental threats to breathing, Exposure to electric current, radiation and extreme ambient air temperature and pressure
epiX	Exposure to smoke, fire and flames, Contact with heat and hot substances, Contact with venomous animals and plants, Exposure to forces of nature, Accidental poisoning by and exposure to noxious substances, Overexertion, travel and privation, Accidental exposure to other and unspecified factors, Intentional self-harm, Assault by drugs, medicaments and biological substances, Assault by corrosive substance, Assault by pesticides, Assault by gases and vapours, Assault by other specified chemicals and noxious substances, Assault by unspecified chemical or noxious substance, Assault by hanging, strangulation and suffocation, Assault by drowning and submersion, Assault by handgun discharge, Assault by rifle, shotgun and larger firearm discharge, Assault by other and unspecified firearm discharge, Assault by explosive material, Assault by smoke, fire and flames, Assault by steam, hot vapours and hot objects, Assault by sharp object
epiY	Assault by blunt object, Assault by pushing from high place, Assault by pushing or placing victim before moving object, Assault by crashing of motor vehicle, Assault by bodily force, Sexual assault by bodily force, Neglect and abandonment, Other maltreatment syndromes, Assault by other specified means, Assault by unspecified means, Event of undetermined intent, Legal intervention and operations of war, Complications of medical and surgical care, Sequelae of external causes of morbidity and mortality, Supplementary factors related to causes of morbidity and mortality classified else
epiZ	XXI

Table 7: Classification of morbidity characteristics

ICD10 codes beginning with U were dropped because there were no observations in the 600,000 used. Only a small number (334) were found of those beginning with P and so these were combined with those beginning with O - owing to the clinical similarities.

Chapter	Blocks	Title
I	A00-B99	Certain infectious and parasitic diseases
II	C00-D48	Neoplasms
III	D50-D89	Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism
IV	E00-E90	Endocrine, nutritional and metabolic diseases
V	F00-F99	Mental and behavioural disorders
VI	G00-G99	Diseases of the nervous system
VII	H00-H59	Diseases of the eye and adnexa
VIII	H60-H95	Diseases of the ear and mastoid process
IX	I00-I99	Diseases of the circulatory system
X	J00-J99	Diseases of the respiratory system
XI	K00-K93	Diseases of the digestive system
XII	L00-L99	Diseases of the skin and subcutaneous tissue
XIII	M00-M99	Diseases of the musculoskeletal system and connective tissue
XIV	N00-N99	Diseases of the genitourinary system
XV	O00-O99	Pregnancy, childbirth and the puerperium
XVI	P00-P96	Certain conditions originating in the perinatal period
XVII	Q00-Q99	Congenital malformations, deformations and chromosomal abnormalities
XVIII	R00-R99	Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified
XIX	S00-T98	Injury, poisoning and certain other consequences of external causes
XX	V01-Y98	External causes of morbidity and mortality
XXI	Z00-Z99	Factors influencing health status and contact with health services
XXII	U00-U99	Codes for special purposes

Table 8: ICD10 chapter codes

Appendix B

Convergence of all models was good, see Table 9 for results:

	Sample size			
Model	**5,000**	**10,000**	**50,000**	**100,000**
GB2	83%	85%	97%	100%
SM	100%	100%	100%	100%
D	100%	100%	100%	100%
B2	84%	83%	81%	85%
L	100%	100%	100%	100%
Fisk	100%	100%	100%	100%
GG	100%	100%	100%	100%
GB2H(age)	69%	74%	87%	94%
GB2H(morb)	77%	69%	74%	80%
GGH(age)	97%	99%	100%	100%
GGH(morb)	98%	99%	100%	100%

Table 9: % models converged by sample size

	Sample size			
Nested model	**5,000**	**10,000**	**50,000**	**100,000**
SM (1 restriction)	16%	93%	100%	100%
D (1 restriction)	0%	52%	100%	100%
B2 (1 restriction)	24%	55%	100%	100%
L (2 restrictions)	99%	100%	100%	100%
Fisk (2 restrictions)	36%	85%	100%	100%

Table 10: Results of tests on nested model restrictions (percentage rejected at 1% significance level)

	Sample size			
Nested model	**5,000**	**10,000**	**50,000**	**100,000**
SM (1 restriction)	0%	60%	100%	100%
D (1 restriction)	0%	0%	100%	100%
B2 (1 restriction)	7%	24%	99%	100%
L (2 restrictions)	98%	100%	100%	100%
Fisk (2 restrictions)	18%	49%	100%	100%

Table 11: Results of tests on nested model restrictions (percentage rejected at 0.1% significance level)

	$\mathbf{R^2}$	**RMSE**	**MAPE**	**MPE**	**ADMPE**
GB2	0.18369	4816.73	1790.38	-74.52	37.01
SM	0.17550	5191.21	1940.67	-385.50	41.14
D	0.18078	4934.93	1849.13	-203.84	27.04
B2	0.18593	4754.89	1764.06	-14.63	24.64
L	0.18840	4945.53	1803.36	-141.20	25.67
Fisk	0.17730	4978.53	1779.08	-39.21	26.70
GG	0.18927	4683.64	1742.46	37.45	23.17
GB2H(age)	0.16282	5334.20	1966.90	-381.27	54.19
GB2H(morb)	0.02010	690955.49	29837.54	-28567.98	24255.43
GGH(age)	0.16741	5128.23	1904.87	-276.22	32.81
GGH(morb)	0.04346	333886.41	18317.55	-16994.79	8808.54

Table 12: Results of model performance, when all converged, sample size 10,000

	$\mathbf{R^2}$	**RMSE**	**MAPE**	**MPE**	**ADMPE**
GB2	0.18361	4817.63	1788.67	-74.41	20.59
SM	0.17524	5192.18	1937.62	-381.93	22.15
D	0.18060	4938.11	1848.34	-205.19	17.28
B2	0.18591	4755.67	1762.84	-15.67	15.40
L	0.18950	4926.12	1799.87	-141.31	18.50
Fisk	0.17711	4980.83	1779.17	-42.32	15.87
GG	0.18945	4681.75	1740.32	37.90	14.99
GB2H(age)	0.16348	5311.73	1957.56	-370.26	28.17
GB2H(morb)	0.02238	1267589.80	53066.93	-51792.75	52993.69
GGH(age)	0.16854	5096.06	1892.49	-258.85	19.81
GGH(morb)	0.04385	442534.20	19522.71	-18195.57	8584.18

Table 13: Results of model performance, when all converged, sample size 50,000

	$\mathbf{R^2}$	**RMSE**	**MAPE**	**MPE**	**ADMPE**
GB2	0.18390	4810.78	1786.66	-70.87	11.83
SM	0.17557	5180.75	1935.58	-379.25	11.33
D	0.18091	4928.87	1845.94	-201.24	9.30
B2	0.18620	4749.75	1760.90	-12.01	8.34
L	0.18989	4918.76	1797.71	-138.29	8.94
Fisk	0.17748	4971.31	1776.90	-38.59	8.84
GG	0.18976	4676.53	1738.35	41.83	8.23
GB2H(age)	0.16349	5310.19	1957.06	-369.55	17.90
GB2H(morb)	0.01753	1335959.90	39015.46	-37742.48	28193.42
GGH(age)	0.16866	5090.25	1890.37	-255.07	11.32
GGH(morb)	0.04429	482306.56	35221.91	-33895.06	31492.27

Table 14: Results of model performance, when all converged, sample size 100,000