ESSAYS ON ASSESSING METHODS FOR MODELLING THE DISTRIBUTION OF HEALTHCARE COSTS

James Richard Scott Lomas

Ph.D. Thesis

University of York
Department of Economics and Related Studies

August 2014

Abstract

This thesis comprises three essays on assessing methods for modelling the distribution of healthcare costs.

Chapter 2 extends the literature on modelling healthcare cost data by applying the generalised beta of the second kind (GB2) distribution to English hospital inpatient cost data. A quasi-experimental design, estimating models on a sub-population of the data and evaluating performance on another sub-population, is used to compare this distribution with its nested and limiting cases. While, for these data, the beta of the second kind (B2) distribution and generalised gamma (GG) distribution outperform the GB2, our results illustrate that the GB2 can be used as a device for choosing among competing parametric distributions for healthcare cost data.

In Chapter 3, we conduct a quasi-Monte Carlo comparison of the recent developments in parametric and semi-parametric regression methods for healthcare costs, both against each other and against standard practice. The population of English NHS hospital inpatient episodes for the financial year 2007-2008 (summed for each patient: 6,164,114 observations in total) is randomly divided into two equally sized sub-populations to form an estimation set and a validation set. Evaluating out-of-sample using the validation set, a conditional density approximation estimator shows considerable promise in forecasting conditional means, performing best for accuracy of forecasting and amongst the best four (of sixteen compared) for bias and goodness-of-fit. The best performing model for bias is linear regression with square root transformed dependent variable, while a generalised linear model with square root link function and Poisson distribution performs best in terms of goodness-of-fit. Commonly used models utilising a log link are shown to perform badly relative to other models considered in our comparison.

Chapter 4 examines methods for estimating the full conditional distribution of healthcare costs. Understanding the data generating process behind healthcare costs remains a key empirical issue. Although much research to date has focused on the prediction of the conditional mean cost, this can potentially miss important features of the full conditional distribution such as tail probabilities. We conduct a quasi-Monte Carlo experiment using English NHS inpatient data to compare 14 approaches to modelling the distribution of healthcare costs: nine of which are parametric, and have commonly been used to fit healthcare costs, and five others designed specifically to construct a counterfactual distribution. Our results indicate that no one method is clearly dominant and that there is a trade-off between bias and precision of tail probability forecasts. We find that distributional methods demonstrate significant potential, particularly with larger sample sizes where the variability of predictions is reduced. Parametric distributions such as log-normal, generalised gamma and generalised beta of the second kind are found to estimate tail probabilities with high precision, but with varying bias depending upon the cost threshold being considered.

Contents

List of Tables

Beta-type Distributions

Comparison of Developments

Estimating the Full Distribution

List of Figures

Beta-type Distributions

Comparison of Developments

Estimating the Full Distribution

Acknowledgements

First of all I would like to thank my supervisors Andrew Jones and Nigel Rice for providing expert advice, not just in the field of Economics, throughout my time as a Ph.D. student. They are both brilliant mentors and rôle models. I am indebted to the Centre for Health Economics (CHE), which provided the perfect base for my studies, and in particular the Health, Econometrics and Data Group where I was allowed to regularly present my work in order to refine my ideas, interpretations and discussions. I could not fail to mention my fellow Ph.D. students, who have been a great source of help and amusement. In particular, I want to thank Dan, with whom I shared an office for four years, for his amazing friendship and 'IT support'. Thanks to him (again!) and to Seamus for proofreading my thesis. I am very grateful for funding from the Economic and Social Research Council (ESRC) and the Royal Economic Society (RES).

I would especially like to thank Andrew and Nigel for conceiving the idea of my visit to the US in 2012 and the ESRC for funding this. I was hosted by the indefatigable John Mullahy and, in conversations with John and Dave Vanness, my brain was successfully rewired so that I would always think about the whole distribution of a variable as opposed to just its mean. There are too many inspiring discussions to mention them all, but I would like to thank John, Dave, Will Manning and Partha Deb who gave up so much time to discuss ideas with me, and who continue to provide support now. I would also like to thank 'Mr. Walton' and Graham Bates for inspiring me as a sixth form student and as an undergraduate.

I dedicate this thesis to my loving fiancée, Jessica, my mom, dad, brother, nan and bop, and whole family who have never stopped believing in me, and without whom this thesis would not be possible. My mom and dad really must take the majority of the credit for this, having always encouraged me to aim for the top and to 'sock it to 'em!', and a special mention must go to 'Boppa' for his invaluable support with all things maths and stats throughout my life.

Declaration

I confirm that the work presented in this thesis is my own. Funding for my studies was initially provided by the Economic and Social Research Council (ESRC) and was attached to the funding for the Health, Econometrics and Data Group (HEDG) under the Large Grant Scheme (reference: RES-060-25-0045). As such, my work was linked to the research projects of the group and all essays are co-authored, which is explicitly acknowledged and detailed below.

Chapter 2 is written in co-authorship with Professor Andrew M. Jones and Professor Nigel Rice, and has been published as a peer-reviewed research article under the title: *Applying Beta-type Size Distributions to Healthcare Cost Regressions. Journal of Applied Econometrics* 2014; 29(4): 649-670. Figure 2.8 has been corrected for this thesis. An earlier version of this paper were presented at the University of Wisconsin-Madison, the University of Chicago, City University of New York and the University of Oslo. I am the lead author, and carried out the empirical analysis, wrote the draft, made revisions and disseminated the paper. I also contributed to the original research idea and preparation of the data. Professor Andrew M. Jones handled the submission process and is thus corresponding author.

Chapter 3 is written in co-authorship with Professor Andrew M. Jones, Professor Nigel Rice and Peter Moore. It is available as a working paper under the title: *A Quasi-Monte Carlo Comparison of Developments in Parametric and Semi-parametric Regression Methods for Heavy Tailed and Non-Normal Data: with an Application to Healthcare Costs. Health, Econometrics and Data Group Working Paper* 2013; 13/30. Some revisions have been made for the thesis version, reflecting feedback from various presentations. This paper has been presented at the University of Leeds, Brunel University and the University of Oxford. In addition, Professor Nigel Rice has presented this paper at the Fourth Italian Health Econometrics Workshop, University of Padua, and Professor Andrew M. Jones has presented it at the Centre for Health Economics and Policy at the University of Copenhagen. Once again I am the lead author, having carried out the empirical analysis and written the draft, and also having contributed to the original research idea, preparation of the data and dissemination of the paper. Preliminary work on the empirical analysis and literature review had been conducted by Peter Moore. In the end these were replaced with my own review and program code, but the contribution by Peter Moore is acknowledged through his co-authorship.

Chapter 4 is written in co-authorship with Professor Andrew M. Jones and Professor Nigel Rice. It is available as a working paper under the title: *Going Beyond the Mean in Healthcare Cost Regressions: a Comparison of Methods for Estimating the Full Conditional Distribution. Health, Econometrics and Data Group Working Paper* 2014; 14/26. I have made a contribution to the original research idea and preparation of the data. I designed and implemented the empirical analysis and wrote the draft. I presented an earlier version of this paper at the 2014 European Workshop in Econometrics and Health Economics in

Munich and the 2014 Annual Health Econometrics Workshop at the University of Toronto.

Chapter 1

Introduction

This thesis contains three essays on the econometric modelling of healthcare costs. Each essay extends, and is motivated by, recent advances in the existing large body of literature on this topic. The work contained in this thesis contributes to the literature by offering a significant methodological advancement in describing the generalised beta of the second kind as an appropriate distribution to model healthcare costs (chapter 2); a rigorous comparison of state of the art approaches (chapter 3); and an empirical assessment of methods for extending the modelling of healthcare costs to the full distribution (chapter 4).

Healthcare costs (or healthcare expenditures) are of major importance on a macroeconomic, as well as microeconomic, level. Historically, and according to most projections, the quantity of money traded for healthcare has been and is continuing to increase: total expenditure on healthcare for the UK as a percentage of gross domestic product (GDP), increased from 4.0% in 1961 to 9.4% in 2011[1] (Organisation for Economic Co-operation and Development, 2013), and the King's Fund project that the UK "could be spending nearly one-fifth" of its GDP on public provision of health and social care in 50 years' time (Appleby, 2013). This means that the precise estimation of healthcare costs is increasingly important. Because of the economic significance of the results generated through modelling, the sensitivity of results to the various econometric methods is a methodological research question with important implications for policy.

Models of healthcare costs are important for driving policy and include the estima-

[1] Public expenditure accounted for 85.1% and 82.8% of total expenditure for these two years respectively.

tion of key parameters for populating decision models in cost-effectiveness analyses (Hoch et al., 2002); adjusting for healthcare need in resource allocation formulae in publicly funded healthcare systems (Dixon et al., 2011); undertaking risk adjustment in insurance systems (Van de Ven and Ellis, 2000) and assessing the effect of observable lifestyle characteristics such as smoking and obesity on resource use (Johnson et al., 2003; Cawley and Meyerhoefer, 2012; Mora et al., 2014). Understanding how best to model the features of healthcare expenditure data is crucial to informing policy decision making.

The focus of this thesis is the assessment of how best to model healthcare cost data considering the various statistical challenges that this endeavour presents. Healthcare cost data are highly non-normal: values cannot be negative and their distribution is asymmetric and right-hand skewed.[2] In addition, cost data possess long, thick right-hand tails with some patients exhibiting extremely large costs owing to clinical complications, comorbidities or rare events. Furthermore, errors are likely to be heteroskedastic and responses to covariates non-linear.

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 for risk adjustment in the UK (see for example Dixon et al., 2009, 2011). This method may be sensitive to extreme observations, and may incorrectly assume constant additive responses to covariates. Transforming the dependent variable may improve performance by reducing skewness, with applied work often using a log (or less frequently a 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. Such re-transformations are likely to be problematic in the case of heteroskedastic errors on the transformed scale (Duan, 1983; Manning et al., 2005).

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 generalised linear model (GLM) family and exponential conditional mean models, which although often considered for count and duration data, can also be applied more widely to positive dependent variables. Each model within these families makes different assumptions about

[2]There is also usually a mass point of costs at zero £, i.e. non-users of healthcare, which are often dealt with as a first stage of a two-part model, and modelling the positive expenditures forms the second stage of the two-part model (see Jones, 2000).

the distribution of the outcome variable. Within the GLM family, it is only necessary to make assumptions about the functional form of the conditional mean and conditional variance of the distribution. Duration and count models typically require assumptions about the parametric form of the entire distribution. Whilst the GLM family is a natural way to deal with heteroskedasticity, it fails to account explicitly for the issues of skewness and kurtosis, which have implications for the efficiency and robustness of estimators (Mullahy, 2009). More flexible parametric distributions allow for a greater range of estimated skewness and kurtosis coefficients (McDonald et al., 2013). Manning et al. (2005) introduce the generalised gamma distribution for use with healthcare costs, a distribution that features important parametric distributions as nested and special cases, such as the gamma and log-normal distribution, each with precedent as popular choices of distribution for modelling healthcare costs.

Another strand of the literature has emphasised the utility of semi-parametric models including extended estimating equations (Basu et al., 2006), finite mixture models (Deb and Burgess, 2003) and conditional density approximation estimators (Gilleskie and Mroz, 2004), which aim to allow for greater flexibility and make fewer assumptions about the functional form of the whole distribution. The extended estimating equations model adopts the generalised linear models framework and allows for the link and distribution functions to be estimated from data, rather than specified *a priori*: while this may ensure consistency there may be important efficiency losses. Finite mixture models introduce heterogeneity (both observed and unobserved) through mixtures of distributions. Conditional density approximation estimators are implemented by dividing the emprical distribution into discrete intervals and then decomposing the conditional density function into 'discrete hazard rates'. As such these approaches, in principle, offer flexibility in modelling the distribution of costs, particularly when functional forms are not well approximated by parametric distributions.

With so many competing econometric methods to choose from, the applied researcher faces a daunting task in determining which approach is best. As pointed out by Basu et al. (2006), it is unlikely that economic theory will provide any *a priori* "guidance about distributional characteristics and functional forms that may relate the outcome of interest to covariates". Thus, there is a need for empirical comparative work to aid

researchers. Traditionally, econometric methods are compared using a Monte Carlo study. This typically involves simulating data with certain characteristics and comparing the performance of estimators given these characteristics. This approach may be inappropriate for healthcare costs, since we are interested in a very large number of permutations of assumptions underlying the distribution of the outcome variable. In addition, such studies are prone to affording advantage to certain models arising from the chosen distributional assumptions used for generating data. However, studies conducted in this area using this approach have been useful in uncovering specific distributional characteristics that cause problems for econometric methods. Manning and Mullahy (2001), for example, find that certain forms of heavy-tailed data, whilst not affecting consistency, render estimates from GLM approaches imprecise.

Due to improvements in computational capacity and burgeoning availability of large datasets through administrative records and surveys, there is a growing literature of empirical comparative assessments using actual healthcare cost data (Deb and Burgess, 2003; Veazie et al., 2003; Buntin and Zaslavsky, 2004; Basu et al., 2006; Hill and Miller, 2010). An important limitation of these studies is that none of them is a comprehensive evaluation of all econometric approaches. In addition, any synthesis of the existing literature would be inconclusive in terms of which method is most appropriate for a given application. However, these papers show, amongst other things, that there is no commonly dominant method. Different models are preferred according to different performance criteria such as, for example, mean prediction error (bias) and mean absolute prediction error (accuracy). The performance of regression methods is judged almost entirely based upon the conditional mean function, where it is found that distributional characteristics such as the specification of the variance function can impact upon the performance in terms of the conditional mean. Another important aspect of this research is that performance is evaluated, at least in part, on observations not contained within the estimation sample. This is necessary, since competing approaches estimate different numbers of parameters and overfitting the sample is an important concern. Chapter 3 further discusses the methodologies and results of these studies.

The work contained in this thesis uses real data taken from the large English administrative dataset: Hospital Episode Statistics (HES), collected by the NHS Information

Centre.[3] Analysis is carried out using a dataset comprising 6,164,114 separate observations, which represents the population of hospital inpatient healthcare users for the year 2007-2008. Since data is taken from administrative records, there is only information on users of inpatient NHS services, and therefore only contains strictly positive costs. In order to fully exploit the large dataset, it is randomly divided into two equally sized groups – an 'estimation' set and a 'validation' set (each with 3,082,057 observations) – before any analysis is undertaken. Because researchers using observational data from social surveys typically have fewer observations in their datasets than are present in the 'estimation' set used here, smaller samples of different sizes are drawn from within the 'estimation' set and these are used for estimating the models. These samples are used to estimate regression models, which are then later evaluated using the data from the 'validation' set. This methodology is called a quasi-experimental design or a quasi-Monte Carlo study, since it follows Monte Carlo principles of resampling, whilst using actual cost data as opposed to hypothetical known distributions. Using this methodology, the three papers in this thesis each make a significant contribution to the literature surrounding the econometric modelling of healthcare costs.

Chapter 2 is focused around the application of the generalised beta of the second kind distribution[4] (GB2, and its special cases) to modelling healthcare costs. As discussed earlier, in order to model data with heavy tails, Manning et al. (2005) introduce the three-parameter generalised gamma distribution. Whilst this distribution is able to more flexibly model skewness and kurtosis than its limiting and nested cases (such as the Weibull, log-normal and gamma distributions), generalised gamma is itself a limiting case of the four-parameter GB2 (and is therefore more restrictive). Jones (2011) considers the use of GB2 for healthcare costs, but estimates GB2 in such a way as to not feature Manning et al. (2005)'s generalised gamma as a special case. In chapter 2, GB2 is estimated with a log link function, making it directly comparable to a whole host of distributions often applied to healthcare costs. The constraints imposed by certain nested distributions are directly testable, e.g. Fisk (log-logistic) and Singh-Maddala distributions. Using the quasi-Monte Carlo study design, this chapter shows that GB2 offers the best fit according to

[3]Now named the Health and Social Care Information Centre.

[4]This is elsewhere known as the generalised-F distribution.

one of the scores that should in principle capture the fit of the whole distribution (Akaike Information Criterion), however it does not perform best in terms of the bias or accuracy of its conditional mean forecasts (its special cases – beta of the second kind and log-normal distribution – are, respectively, best according to these metrics). In addition, GB2 does not perform best in terms of forecasting the probability of very high costs, which appear to be better-modelled using the more parsimonious log-normal and generalised gamma distributions. This chapter concludes that GB2 offers considerable benefits over and above the existing parametric methods to which it is compared, but that in this illustration it is unclear as to whether it would be the best distribution to apply to the data, since less flexible distributions perform better in key performance criteria. However, GB2 shows promise as an umbrella distribution from which more restrictive functional forms can be chosen through statistical tests.

The contribution of chapter 3 is to provide a systematic comparison of all recent developments in semi-parametric and fully parametric modelling against each other and against standard practice. To begin, the chapter provides a review of comparative work using actual data, exploring the methodology adopted as well as the results attained. From this, it is clear that there is no comprehensive empirical comparison existing in the literature prior to this work. In particular, GB2 had only been compared against other fully parametric approaches, and not with standard practice or semi-parametric methods, and the conditional density approximation estimator had previously not been compared to other techniques using real data. The focus of this chapter is the performance of econometric methods in terms of predicting the conditional mean, which is the focus of most comparative work, and is inherently important in informing policy in healthcare. In line with other comparative studies, no model performs best across all metrics of evaluation, which are bias, accuracy and goodness-of-fit of forecasted conditional means. Using a quasi-Monte Carlo study design, the results indicate that models estimated with a square root link function perform much better than those with log or linear link functions. In addition, more flexibility in the econometric modelling does not necessarily result in better performance. As previously noted, the link function plays a big role, but when this is flexibly estimated from the data using the extended estimating equations approach, such an approach performs less well than models with a square root link function. Among models with a square

root link, additional flexibility in the functional form of the whole distribution is found to lead to improvements in the accuracy of forecasts, with both a two component gamma model and GB2 performing among the top four. The conditional density approximation estimator performs among the top four across all three metrics of performance, and so demonstrates promise. Another interesting result is that commonly used models such as linear regression (on levels of and log-transformed costs), a gamma GLM with log link, and the log-normal distribution, are not among the four best performing models with any of the chosen metrics. In summary, the findings presented in this chapter illustrate the sensitivity of results to the choice of econometric method and provide insight into the best models to consider when interested in bias, accuracy or goodness-of-fit of forecasted conditional means.

Chapter 4 is focused on the issue of fitting the full conditional distribution of healthcare costs. In so doing, it extends the literature on comparative assessment of econometric methods for healthcare costs, since these have almost exclusively investigated performance based on the fit of the conditional mean. As mentioned above, the conditional mean is important for policymakers, but is generally not the only characteristic that is of interest to policymakers (a full discussion can be found in Vanness and Mullahy, 2007). In particular, within health economics there is a particular emphasis on identifying individuals or characteristics of individuals that lead to very large costs. Methods estimated using maximum likelihood within the existing literature on modelling cost data are compared with distributional methods which have been more often applied within labour economics (Fortin et al., 2011). These distributional methods involve running many equations for different points of the distribution (Han and Hausman, 1990; Foresi and Peracchi, 1995; Chernozhukov et al., 2013) or at different quantiles (Machado and Mata, 2005; Melly, 2005; Firpo et al., 2009). Chapter 4 reviews these methods and their applications involving healthcare cost data, and provides details on how each is estimated. The quasi-Monte Carlo approach is used to forecast tail probabilities with each of these methods. Unlike Chapter 2, these tails are throughout the entire distribution of costs, rather than just being very high costs. This therefore analyses the fit of the whole distribution. Through repeated sampling, tail probabilities are evaluated in terms of bias and precision. Distributional methods demonstrate significant potential in modelling tail probabilities, particularly with larger sample

sizes where the variability of predictions is reduced. Parametric distributions such as log-normal, generalised gamma and generalised beta of the second kind are found to estimate tail probabilities with high precision, but with varying bias depending upon which tail probability is being considered.

Chapter 5 compares and contrasts the findings in Chapters 2, 3 and 4, and identifies avenues for future research.

Chapter 2

Beta-type Distributions

Applying Beta-type Size Distributions to Healthcare Cost Regressions

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

Summary

This paper extends the literature on modelling healthcare cost data by applying the generalised beta of the second kind (GB2) distribution to English hospital inpatient cost data. A quasi-experimental design, estimating models on a sub-population of the data and evaluating performance on another sub-population, is used to compare this distribution with its nested and limiting cases. While, for these data, the beta of the second kind (B2) distribution and generalised gamma (GG) distribution outperform the GB2, our results illustrate that the GB2 can be used as a device for choosing among competing parametric distributions for healthcare cost data.

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 3766.
E-mail address: andrew.jones@york.ac.uk

2.1 Introduction

Modelling healthcare costs is of primary importance in health economics and health services research for, broadly speaking, two reasons. Firstly, cost-effectiveness analyses that compare costs of treatments to the health gains achieved, often require statistical methods of modelling cost data, particularly in the absence of clinical trial data. Secondly, such methods are used for risk-adjustment purposes in either insurance schemes (for example in the US) or devolving budgets to healthcare organisations (in the case of the UK). This requires regression methods, applied to large datasets, to predict specific healthcare costs for individuals or groups of patients (typically over a long period such as a year), adjusting for the needs for healthcare using morbidity characteristics, together with socioeconomic information, and age and gender.

Healthcare cost data are highly non-normal and their distributional characteristics 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 positive costs for hospital inpatients and do not consider the problem of the many zero cost observations that would be found in a general population sample of users and non-users of healthcare. Zero costs are often dealt with as a first stage of a two-part model, and modelling the positive expenditures forms the second stage of the two-part model (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 for risk adjustment in the UK (see for example Dixon et al., 2009, 2011). This method may be sensitive to extreme observations, and may incorrectly assume constant additive responses to covariates. Transforming the dependent variable may improve performance by reducing skewness, with applied work often using a log (or less frequently a 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.

Such retransformations, however are likely to be problematic in the case of heteroskedastic errors on the transformed scale (Duan, 1983; Manning et al., 2005).

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 generalised linear model (GLM) family and exponential conditional mean (ECM) models which, although often considered for count and duration data, can also be applied more widely to positive dependent variables. Each model within these families makes different assumptions about the distribution of the outcome variable. Within the GLM family, it is only necessary to make assumptions about the functional form of the mean and variance of the distribution. Duration and count models typically require assumptions about the parametric form of the entire distribution. Whilst the GLM family is a natural way in which to deal with heteroskedasticity, it fails to account explicitly for the issues of skewness and kurtosis which have implications for the efficiency and robustness of estimators (Mullahy, 2009). More flexible distributions allow for a greater range of estimated skewness and kurtosis coefficients (McDonald et al., 2013). Manning et al. (2005) introduced the generalised gamma (GG) distribution for use with healthcare costs, where GG includes important parametric distributions as nested and special cases, such as the gamma (GAMMA) and log-normal (LN) distribution, each with precedent as popular choices of distribution for modelling healthcare costs. In addition, specifying a fully parametric distribution allows us to extract information from our estimated model about moments beyond the mean and to look at the whole range of predicted quantiles. In the case of healthcare costs, a policymaker may be interested in the right-hand tail of the distribution and in knowing the probability of an individual's healthcare cost exceeding a certain threshold, for example if the individual would then become eligible for reinsurance (Deb and Burgess, 2003).

In this paper, we explore the use of the generalised beta of the second kind (GB2), and its nested and limiting case distributions, on English inpatient healthcare cost data. Jones (2011) suggests the use of GB2 as part of a comparison of many different methods for modelling US healthcare costs. 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. Importantly, the GG is also a special limiting case of the GB2. It is well known that the distribution of healthcare costs differs from application to application (Hill

and Miller, 2010), and so the GB2 is well placed to discriminate amongst its competing special case distributions, each of which are potential distributions an applied researcher might choose to use for modelling healthcare costs. The full set list of distributions considered in this paper is as follows (with label in parentheses): generalised beta of the second kind (GB2), Singh-Maddala (SM), Dagum (DAGUM), beta of the second kind (B2), Lomax (LOMAX), Fisk (FISK), generalised gamma (GG), gamma (GAMMA), log-normal (LN), Weibull (WEI) and exponential (EXP).

In addition to the 'umbrella' function served by the GB2 (Cox, 2008), it is also possible that its special or nested case distributions have inadequate flexibility to capture the nature of the underlying data generating process, with the GB2 distribution itself being the best fit. Bordley et al. (1997) find that the GB2 provides the best fit to US incomes, following a comprehensive study of gamma- and beta-type distributions, suggesting that the extra flexibility of the GB2 is required in this case. Each parametric distribution implies certain restrictions upon the skewness and kurtosis coefficients, which often depend upon the distribution parameters. In Figure 2.1, adapted from McDonald et al. (2013), we display the possible combinations of skewness and kurtosis for the GB2 and each of its nested and limiting cases. In particular, the GB2 allows for a greater flexibility of possible skewness-kurtosis estimates than any of its nested or limiting cases; Figure 2.1 shows that the possible skewness-kurtosis spaces for all of the nested or limiting cases of GB2 are enveloped by the skewness-kurtosis space for GB2. Three- and four-parameter distributions have possible skewness-kurtosis spaces[1], whilst two-parameter distributions only allow for a locus of points, and one-parameter distributions (such as the exponential distribution) a single point (skewness equal to 2 and kurtosis equal to 9). We superimpose the skewness and kurtosis coefficients from the quantiles of our hospital inpatient data (described later), which are shown to lie within the GB2 skewness-kurtosis space.

The GB2 and some of its nested cases can also be estimated where the implied population distribution has undefined moments, such as undefined variance, skewness and kurtosis coefficients, for certain ranges of parameter values. While the sample moments must be defined, those of the underlying data generating process may not be. Indeed

[1]In Figure 2.1 we denote an upper bound with subscript U, and lower bound with subscript L.

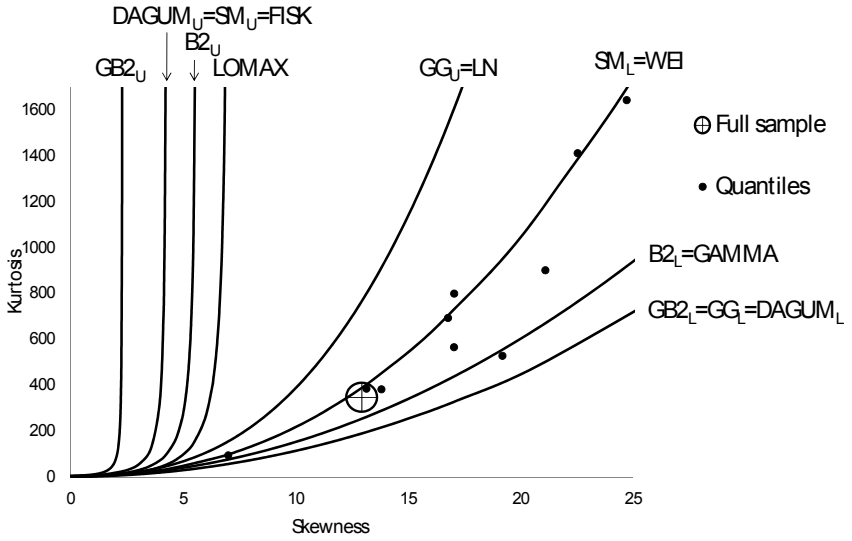


Figure 2.1: Graph showing skewness-kurtosis spaces for each distribution

Note:

The dots shown on Figure 2.1 were generated as follows: the data were divided into ten subsets using the deciles of a simple linear predictor for healthcare costs using the set of regressors introduced later. Figure 2.1 plots the skewness and kurtosis coefficients of actual healthcare costs for each of these subsets, the skewness and kurtosis coefficient of the full estimation sub-population (represented by the larger circle with cross) and theoretically possible skewness-kurtosis spaces and loci for each distribution considered in this paper.

Mandelbrot (1963) discusses such distributions, where the mean is finite, but variance (and measures related to higher moments) are infinite. Some popular distributions used for healthcare costs lack the flexibility to produce undefined moments (e.g. GAMMA, LN, WEI and EXP).

The flexibility of beta-type size distributions makes them appropriate in a wide variety of circumstances, ranging from unemployment duration to fire losses faced by a university, as well as US incomes and nursing home occupancy rates in the actuarial literature (McDonald and Butler, 1987; Cummins et al., 1990; Bordley et al., 1997; Sun et al., 2008). Cummins et al. (1990) find that although the four-parameter GB2 fits the data well, the parsimonious one-parameter EXP 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).

This paper contributes to the literature on modelling healthcare costs by comparing beta-type size distributions – GB2 and its nested and limiting case distributions – specified using a log link, in a quasi-experimental design using English hospital inpatient cost data

(Hospital Episode Statistics). Healthcare cost data from a financial year is divided into an 'estimation' and a 'validation' sub-population. Models are estimated and tested using samples of observations from the former, and their forecasting performance evaluated on the latter, with emphasis on bias, accuracy and goodness of fit of forecasted conditional means, as well as forecasted probability of costs beyond a certain threshold. Marginal effects, although of interest, are not the primary concern in this paper. We evaluate performance at different 'estimation' 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 providing the best fit for these data and model specification, although B2 offers the least biased results of those tested. We find that the relative performance of models changes little with increasing sample size, in contrast to the results in Deb and Burgess (2003).

Other, 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, allowing the researcher to control for heterogeneity encountered in the data. Here we focus only on parametric models for ease of comparison, and to focus on the specific issue of choice of the functional form of the whole distribution.

2.2 Empirical Models

We estimate 11 regression models in total, with each model having up to four fundamental scale and shape parameters to estimate. All models are estimated, using maximum likelihood, with only the scale parameter specified as an exponential function of covariates: the four-parameter generalised beta of the second kind (GB2), its three-parameter nested distributions (Singh-Maddala (SM), Dagum (DAGUM), beta of the second kind (B2)), its two-parameter nested distributions (Lomax (LOMAX), Fisk (FISK)) and the limiting cases of the three-parameter generalised gamma (GG), two-parameter gamma (GAMMA), log-normal (LN) and Weibull (WEI) distributions, and the one-parameter exponential (EXP) distribution.

2.2.1 Generalised Beta of the Second Kind

The probability density function (2.1) and first moment (2.2) of the GB2 distribution are as follows (McDonald, 1984):

$$f(y) = \frac{ay^{ap-1}}{b^{ap}B(p,q)[1+(\frac{y}{b})^a]^{(p+q)}} \tag{2.1}$$

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

$$E(y) = b\left[\frac{\Gamma(p+\frac{1}{a})\Gamma(q-\frac{1}{a})}{\Gamma(p)\Gamma(q)}\right] \tag{2.2}$$

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 the relative values of p and q influencing the degree of skewness in the distribution.

In Figure 2.2, 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). The flexibility of the GB2, as demonstrated by Figure 2.2, is a considerable strength. We further note that GB2 can be negatively skewed, and whilst this is unlikely to be useful for healthcare costs, health-related quality of life measures and hospital occupancy rates are examples of outcomes that may follow a negatively skewed distribution.

With the Stata module `gb2fit` developed by Jenkins (2009), it is possible to fit the GB2 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 the parameters as non-negative 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]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).

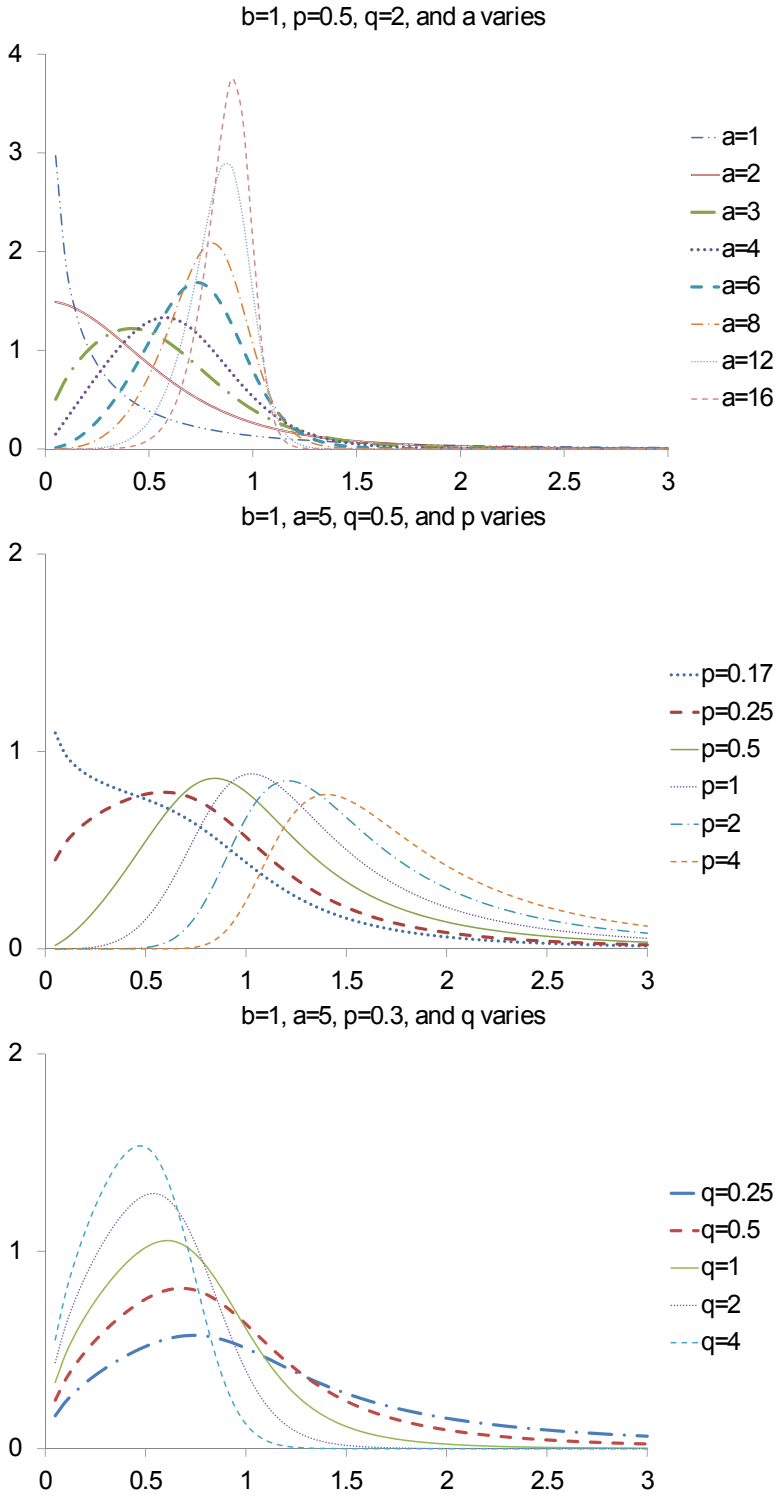
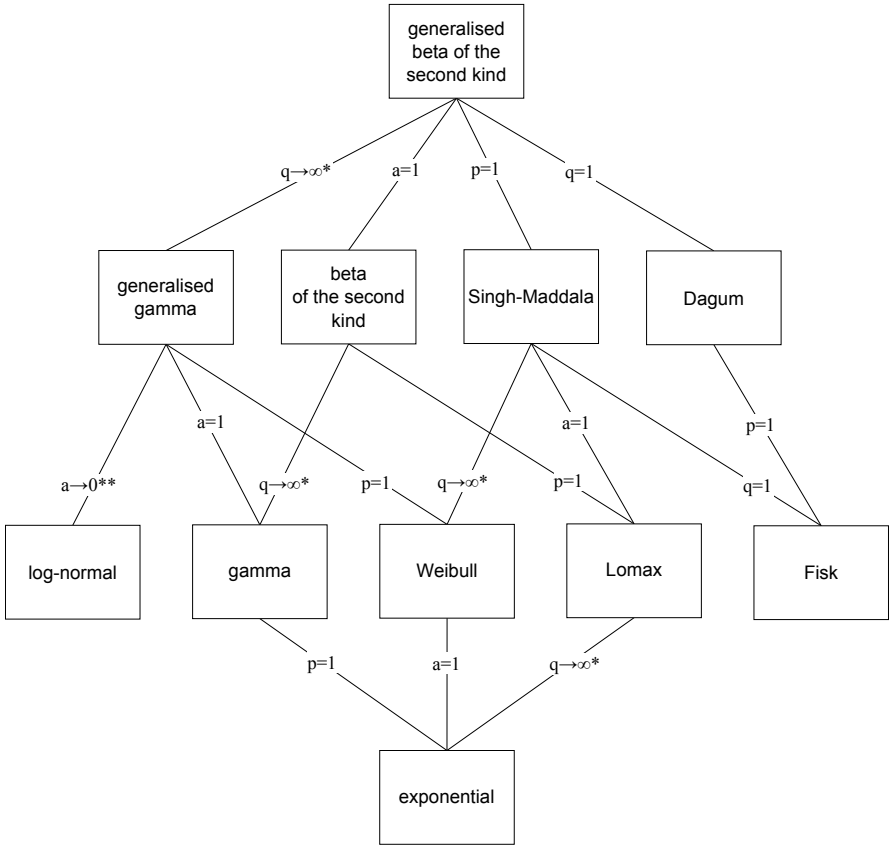


Figure 2.2: Probability density functions for GB2 with different parameter values

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 ($a = 1$ and $q = 1$), which is theoretically unable to produce estimates of the mean, and is rarely referred to in the modelling literature (Kleiber and Kotz, 2003). We note that SM, of the beta-type family of distributions, has been discussed in the healthcare costs literature as a method to deal with the heavy-tailed nature of cost data (Mullahy, 2009). Kleiber and Kotz (2003) provide a thorough account of these distributions and their respective histories in the statistics literature. Table 2.1 shows the probability density functions, cumulative distribution functions and moments of each of the distributions used here and Figure 2.3 the relationships between the GB2 and its nested and limiting cases.

This paper builds on advances made in Manning et al. (2005) in terms of parametric distributions to be applied to healthcare costs. Manning et al. (2005) compare GG with its nested and limiting cases (GAMMA, WEI, EXP and LN). Similarly, we compare GB2 with its nested and limiting cases. Since GG is a limiting case of GB2, we include all the distributions used in Manning et al. (2005) (see Table 2.2). In their paper, Manning et al. (2005) also discuss models where a second parameter is allowed to vary with covariates, termed 'heteroskedastic' models. In principle, the GB2 allows for a second parameter to be specified as a function of covariates. Preliminary work, however, showed such models to perform very poorly and we do not investigate these models in this paper.



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

Figure 2.3: The relationship between beta-type and gamma-type size distributions employed in this paper (adapted from McDonald and Xu, 1995)

	Probability Density Function $\mathbf{f}(\mathbf{y}) =$	$\mathbf{r^{th}}$ **Moment** $\mathbf{E}(\mathbf{Y^r}) =$	**Cumulative Distribution Function** $\mathbf{F}(\mathbf{y}) =$
GB2	$\frac{ay^{ap-1}}{\left(\exp\left(x_i'\beta\right)\right)^{ap}B(p,q)\left[1+\left(\frac{y}{\exp\left(x_i'\beta\right)}\right)^{a}\right]^{(p+q)}}$	$\left(\exp\left(x_i'\beta\right)\right)^{r}\left[\frac{\Gamma\left(p+\frac{r}{a}\right)\Gamma\left(q-\frac{r}{a}\right)}{\Gamma(p)\Gamma(q)}\right]$	$I_Z(p,q)$* where $z=\left(\frac{y}{\exp\left(x_i'\beta\right)}\right)^{a}$
SM	$\frac{aqy^{a-1}}{\left(\exp\left(x_i'\beta\right)\right)^{a}\left[1+\left(\frac{y}{\exp\left(x_i'\beta\right)}\right)^{a}\right]^{(1+q)}}$	$\left(\exp\left(x_i'\beta\right)\right)^{r}\left[\frac{\Gamma\left(1+\frac{r}{a}\right)\Gamma\left(q-\frac{r}{a}\right)}{\Gamma(q)}\right]$	$1-\left[1+\left(\frac{y}{\exp\left(x_i'\beta\right)}\right)^{a}\right]^{-q}$
DAGUM	$\frac{apy^{ap-1}}{\left(\exp\left(x_i'\beta\right)\right)^{ap}\left[1+\left(\frac{y}{\exp\left(x_i'\beta\right)}\right)^{a}\right]^{(p+1)}}$	$\left(\exp\left(x_i'\beta\right)\right)^{r}\left[\frac{\Gamma\left(p+\frac{r}{a}\right)\Gamma\left(1-\frac{r}{a}\right)}{\Gamma(p)}\right]$	$\left[1+\left(\frac{y}{\exp\left(x_i'\beta\right)}\right)^{-a}\right]^{-p}$
B2	$\frac{y^{p-1}}{\left(\exp\left(x_i'\beta\right)\right)^{p}B(p,q)\left[1+\left(\frac{y}{\exp\left(x_i'\beta\right)}\right)\right]^{(p+q)}}$	$\left(\exp\left(x_i'\beta\right)\right)^{r}\left[\frac{\Gamma(p+r)\Gamma(q-r)}{\Gamma(p)\Gamma(q)}\right]$	$I_Z(p,q)$* where $z=\left(\frac{x}{\exp\left(x_i'\beta\right)}\right)$
LOMAX	$\frac{q}{\exp\left(x_i'\beta\right)}\left[1+\left(\frac{y}{\exp\left(x_i'\beta\right)}\right)\right]^{-q}$	$\left(\exp\left(x_i'\beta\right)\right)^{r}\left[r!\frac{\Gamma(q-r)}{\Gamma(q)}\right]$	$1-\left[1+\left(\frac{y}{\exp\left(x_i'\beta\right)}\right)\right]^{-q}$
FISK	$\frac{ay^{a-1}}{\left(\exp\left(x_i'\beta\right)\right)^{a}\left[1+\left(\frac{y}{\exp\left(x_i'\beta\right)}\right)^{a}\right]^{2}}$	$\left(\exp\left(x_i'\beta\right)\right)^{r}\Gamma\left(1+\frac{r}{a}\right)\Gamma\left(1-\frac{r}{a}\right)$	$\left[1+\left(\frac{y}{\exp\left(x_i'\beta\right)}\right)^{-a}\right]^{-1}$

*where $I_Z(p,q)=\frac{1}{B(p,q)}\int_0^z\frac{t^{p-1}}{(1+t)^{p+q}}dt$ is the incomplete beta function ratio.

Table 2.1: Beta-family distribution properties

	Probability Density Function $\mathbf{f}(\mathbf{y}) =$	$\mathbf{r^{th}}$ **Moment** $\mathbf{E}(\mathbf{Y^r}) =$	**Cumulative Distribution Function** $\mathbf{F}(\mathbf{y}) =$
GG	$\frac{\kappa}{\sigma y \Gamma(\kappa^{-2})}\left(\kappa^{-2}\left(\frac{y}{\exp(x_i'\beta)}\right)^{\kappa/\sigma}\right)^{\kappa^{-2}} exp\left(-\kappa^{-2}\left(\frac{y}{\exp(x_i'\beta)}\right)^{\kappa/\sigma}\right)$	$(\exp(x_i'\beta))^r \left(\kappa^{2\sigma/\kappa}\right)^r \frac{\Gamma\left(\kappa^{-2}+\frac{r\sigma}{\kappa}\right)}{\Gamma(\kappa^{-2})}$	if $\kappa > 0$: $\Gamma\left(z;\kappa^{-2}\right)$* if $\kappa < 0$: $1-\Gamma\left(z;\kappa^{-2}\right)$* where $z = \kappa^{-2}\left(\frac{y}{\exp(x_i'\beta)}\right)^{\kappa/\sigma}$
GAMMA	$\frac{1}{y\Gamma(\kappa^{-2})}\left(\kappa^{-2}\left(\frac{y}{\exp(x_i'\beta)}\right)\right)^{\kappa^{-2}} exp\left(-\kappa^{-2}\left(\frac{y}{\exp(x_i'\beta)}\right)\right)$	$(\exp(x_i'\beta))^r \left(\kappa^2\right)^r \frac{\Gamma\left(\kappa^{-2}+r\right)}{\Gamma(\kappa^{-2})}$	if $\kappa > 0$: $\Gamma\left(z;\kappa^{-2}\right)$* if $\kappa < 0$:$1-\Gamma\left(z;\kappa^{-2}\right)$* where $z = \kappa^{-2}\left(\frac{y}{\exp(x_i'\beta)}\right)$
LN	$\frac{1}{\sigma y\sqrt{2\pi}}\exp\left(\frac{-\left(\ln y - x_i'\beta\right)^2}{2\sigma^2}\right)$	$(\exp(x_i'\beta))^r \exp\left(\frac{r^2\sigma^2}{2}\right)$	$\Phi\left(\frac{\ln y - x_i'\beta}{\sigma}\right)$
WEI	$\frac{1}{\sigma y}\left(\frac{y}{\exp(x_i'\beta)}\right)^{\frac{1}{\sigma}} exp\left(-\left(\frac{y}{\exp(x_i'\beta)}\right)^{\frac{1}{\sigma}}\right)$	$(\exp(x_i'\beta))^r \Gamma(1+r\sigma)$	$1-\exp\left(-\left(\frac{y}{\exp(x_i'\beta)}\right)^{\frac{1}{\sigma}}\right)$
EXP	$\frac{1}{\exp(X\beta)}\exp\left(\frac{-y}{\exp(X\beta)}\right)$	$(\exp(x_i'\beta))^r r!$	$1-\exp\left(-\frac{y}{\exp(x_i'\beta)}\right)$

*where $\Gamma\left(z;\kappa^{-2}\right) = \frac{1}{\Gamma(\kappa^{-2})}\int_0^z t^{\left(\kappa^{-2}-1\right)}\exp(-t)dt$.

Table 2.2: Gamma-family distribution properties

2.3 Data and Choice of Variables

We use individual-level data from England on the use of hospital inpatient services to assess the comparative performance of the various parametric distributions. 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[3], containing information on all inpatient episodes, outpatient visits and A&E attendances for all patients admitted to English NHS hospitals. Information is collected via medical records.

The cost variable used throughout is individual patient annual NHS hospital cost 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 applied to the most expensive episode within the spell of an inpatient stay. 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 costs only, ignoring non-users of services. In total this provides us with 6,164,114 observations on healthcare costs. With exception of users of maternity or mental healthcare, this represents the population of hospital inpatient healthcare users for the year 2007-08.

Table 2.3 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 leptokurtic, implying that beta-type distributions may be useful in trying to model the thickness of the tails. Figure 2.4 displays a histogram for raw and transformed data, including an Epanechnikov kernel plot. Even after log transformation, the distribution exhibits right-hand skewness.

Following the model used to inform the distribution of healthcare resources across gen-

[3]Now named the Health and Social Care Information Centre.

[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 incomplete since a lot of care is undertaken in the community and with specialist providers (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, and accordingly for the setting of healthcare budgets are determined using alternative mechanisms.

	Level	Square root	Logarithm
N	6,164,114		
Mean	£2,609.95	43.18	7.25
Median	£1,126	33.56	7.03
Standard deviation	£5,087.96	27.30	1.00
Skewness	13.03	2.84	0.74
Kurtosis	363.18	19.62	2.99
Maximum	£604,701.1	777.63	13.31
99th percentile	£19,015	137.89	13.31
95th percentile	£8,956	94.64	9.10
90th percentile	£6,017	77.57	8.70
75th percentile	£2,722	52.17	7.91
25th percentile	£610	24.70	6.41
10th percentile	£446	21.12	6.10
5th percentile	£406.5	20.16	6.01
1st percentile	£347	18.63	5.85
Minimum	£217	14.73	5.38

Table 2.3: Descriptive statistics for hospital costs

eral practices within Primary Care Trusts in England (Person-Based Resource Allocation: Dixon et al., 2009, 2011), we specify a parsimonious set of covariates to model costs based on morbidity characteristics of individuals (from ICD classifications), together with age and gender. This specification mirrors common practice in the comparative literature on econometric approaches to healthcare cost data, such as Deb and Burgess (2003), who use morbidity markers in the form of Diagnostic Cost Groups. In order to control for age and gender, we use polynomials up to a cubic term for age together with interactions with gender, as well as a gender dummy. The average age in the observed population is just under 52 years of age and around 54 percent of individuals are female. Morbidity information is available through the HES dataset, adapted from the ICD10 chapters (WHO, 2007) – see Appendix A for further details. The 24 different morbidity indicators are coded 1 if an individual 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 represent the severity of the condition, merely its presence or absence.

While 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, this would seem more appropriate when comparing nested and limiting cases of models each using

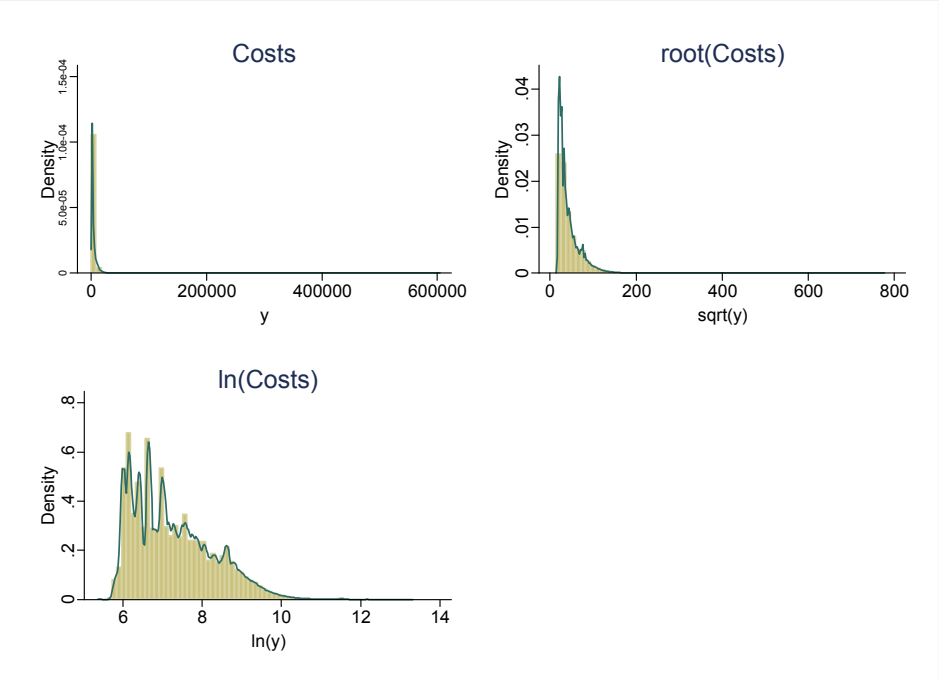


Figure 2.4: Histogram plots of costs

the same link function as is the case here.[6]

2.4 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; 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, the preferred model may vary with the sample size of data on which the model is estimated (Deb and Burgess, 2003) and may 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).

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

2.4.1 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).[7] The population of observations (6,164,114) is randomly divided into two equally sized sub-populations: an 'estimation' set (3,082,057) and a 'validation' set (3,082,057).[8] 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.

By employing this method, the study design follows Monte Carlo principles of resampling, whilst using actual cost data as opposed to hypothetical known distributions. Since it is unlikely that actual costs will follow a simple 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 that are not used in estimation guards against over-fitting and embodies the predictive nature of risk adjustment and resource allocation. Our quasi-experimental 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. Instead, we concentrate on the ability of the models to forecast costs for each observation in terms of predictive bias, accuracy and goodness of fit.

2.4.2 Evaluation of Performance

Estimation Sample Evaluation

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

[7] Such split-sample style modes of evaluation have earlier precedent in the comparitive literature on healthcare costs, see Duan et al. (1983).

[8] Given the size of the dataset, any sub-optimality resulting from the proportions allocated to each set is likely to minimal. To ensure the results are replicable, we set a fixed seed for splitting the dataset into two sets and for randomly drawing samples.

Wald test and report rejection rates at the 5 percent significance level as a percentage of all replications where all models are estimated successfully (see Table 2.4). To compare beta-type models with the gamma-type models (a limiting case of the former and not a linear restriction of a parameter), we use Akaike and Bayesian Information Criteria: for a discussion of the problem with comparing model performance between non-nested (but special) cases, see Vuong (1989). 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 mean prediction error (MPE – see next section) by deciles of predicted levels of costs, analogous to modified Hosmer-Lemeshow and Pearson correlation coefficient tests, allowing 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, and so this interpretation is more about the appropriateness of the shape parameters in influencing the conditional means of each competing distribution.

Validation Set Evaluation

In health economics, the estimated conditional mean cost is often 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, or group of individuals, to the healthcare provider over a certain period of time (Rice and Smith, 2002). Accordingly, we use our models to estimate forecasted mean 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 bias 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 replications (absolute deviations of mean prediction error, ADMPE). Note that R^2 is calculated by an auxiliary regression of actual levels of costs on forecasted

values.[9] These are evaluated on the full 'validation' set. Formulae for calculating these metrics are provided below.[10] Only replications where all 11 models are successfully estimated on the sample are included for evaluation, and model performance according to each criterion is calculated as an average over all included replications.

$$MPE_{msr} = \frac{\sum (y_i - \hat{y}_i)}{N_s} \tag{2.3}$$

$$MAPE_{msr} = \frac{\sum |y_i - \hat{y}_i|}{N_s} \tag{2.4}$$

$$RMSE_{msr} = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{N_s}} \tag{2.5}$$

$$R^2_{msr} = 1 - \frac{\sum (y_i - (\alpha_{AUX} + \beta_{AUX}\hat{y}_i))^2}{\sum (y_i - \hat{y}_i)^2} \tag{2.6}$$

$$ADMPE_{msr} = \left| MPE_{msr} - \frac{\sum_{r=1}^{N_r} MPE_{msr}}{N_r} \right| \tag{2.7}$$

In order to get a greater insight into the performance of different distributions, we evaluate forecasted conditional means at different levels of cost. We assess MPE and MAPE for deciles of actual costs, in order to investigate the degree to which flexible parametric models can model heavily right hand tailed data. This nature of the distribution's skewness and kurtosis partly motivates the usage of flexible parametric models. Furthermore, we can use these results to uncover the locations in the distribution of costs where each model performs well relative to the other models in the comparison.

In addition, we consider the goodness-of-fit of the whole distribution by calculating the estimated probability of having a cost above a chosen threshold level for each observation. Since all of the models considered are fully parametric, it is possible to extract this given estimated parameter values using the cumulative distribution function. Once a probability is estimated for each observation, we take the average of these, giving the amount of mass the estimated distribution allocates beyond the chosen value, which can then be compared

[9] In equation 2.6 coefficients from the 'auxiliary' regression are denoted with $_{AUX}$ subscript.
[10] Where $_m$ denotes the model used, $_s$ the sample size used, and $_r$ the replication.

to the observed empirical mass from the data given by the proportion of costs above the chosen level.

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).

$$MPE_{msr} = \alpha_m^{MPE} + \beta_m^{MPE}\left(\frac{1}{N_s}\right) + u_{msr}^{MPE} \tag{2.8}$$

We specify the relationship to be between MPE and the inverse of the sample size, reflecting that we expect reduced bias as the number of observations increases. In particular, the value of α_m^{MPE} represents the value of MPE to which the model approaches asymptotically with increasing sample size. Here, u_{msr}^{MPE} represents the error term from the regression. For the metrics that cannot be negative, we use the log function of the value as the dependent variable. As such, in the case of mean absolute prediction error we estimate:

$$\ln\left(MAPE_{msr}\right) = \alpha_m^{\ln(MAPE)} + \beta_m^{\ln(MAPE)}\left(\frac{1}{N_s}\right) + u_{msr}^{\ln(MAPE)} \tag{2.9}$$

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

2.5 Results and Discussion

To begin discussing the results, we mainly focus on results using samples with 5,000 observations. We analyse results across increasing sample sizes by summarising and discussing the response surfaces.

2.5.1 Estimation Sample Results

Where all 11 models were successfully estimated[11], we tested the restrictions required

[11] We discarded estimates from two of the samples with 5,000 observations, since Stata was unable to predict means for the LOMAX, owing to a large q parameter. No estimates were discarded at larger sample sizes.

for each of the nested beta-type size distributions using a Wald test. In Table 2.4 we present the percentage of replications where the null hypothesis, of the restriction being valid, was rejected at a 5% level of significance. Accordingly a higher percentage indicates greater evidence against the nested model being appropriate.

	Sample size	
Nested model	**5,000**	**10,000**
SM (1 restriction)	71%	100%
DAGUM (1 restriction)	20%	82%
B2 (1 restriction)	42%	71%
LOMAX (2 restrictions)	98%	100%
FISK (2 restrictions)	53%	100%

Table 2.4: Results of tests on nested model restrictions (percentage rejected at 5% significance level)

The restrictions are rejected least often with $q = 1$ (DAGUM), followed by $a = 1$ (B2) at a sample size of 5,000 (with 10,000 observations, we reject least often $a = 1$ (B2) and then $q = 1$ (DAGUM)[12]). These results might lead to the increased use of these two models over the GB2. Since we use actual data, we cannot observe either where type 1 and type 2 errors occur or their associated losses. Suppose there is little cost associated with using a more flexible model: efficiency loss is small, there is little extra computational demand, and overfitting is hard to detect. In this case, the researcher should use the more flexible model, knowing that any type 1 error (if a more parsimonious model was, in fact, valid) is associated with little cost. In addition, if a more parsimonious model was used, as a result of the restriction not being rejected, the reduced goodness-of-fit could present a large cost due to the type 2 error if the correct model were actually the more flexible form. It is, therefore, necessary to bear these results in mind when looking at the results of goodness-of-fit for forecasting out-of-sample. We do not consider here gains or losses from efficiency or computational demand, although both are relevant in practice.

Results presented in Table 2.5 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 one-parameter EXP model performs the worst according to both AIC and BIC. Of the two-parameter distributions, the FISK model

[12]Percentage rejected at 5% significance level was 100% with sample sizes 50,000 and 100,000.

GB2	**AIC**	**83,981**
	BIC	84,209
SM	**AIC**	84,026
	BIC	84,247
DAGUM	**AIC**	83,986
	BIC	84,208
B2	**AIC**	83,984
	BIC	**84,205**
LOMAX	**AIC**	85,822
	BIC	86,037
FISK	**AIC**	84,122
	BIC	84,337
GG	**AIC**	83,996
	BIC	84,217
GAMMA	**AIC**	85,269
	BIC	85,484
LN	**AIC**	84,141
	BIC	84,356
WEI	**AIC**	85,673
	BIC	85,888
EXP	**AIC**	85,899
	BIC	86,108

Table 2.5: Values for each model's average AIC and BIC at sample size 5,000

is found to perform the best, with the LOMAX model the worst. In general, according to these measures, the two-parameter distributions perform worse than those with three parameters. The best three-parameter distribution, according to this process, is B2: the best overall according to BIC. This is unsurprising from the results in Table 2.4, where we reject the $a = 1$ restriction in 42% of subsets. The GB2, a four-parameter (and thus the most flexible) distribution is the top ranking model in terms of AIC and is third according to BIC.

Figure 2.5 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 (similar to a modified Hosmer-Lemeshow (1980) test) – 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. Because of this, we display the smallest nine deciles of predicted cost on one scale, with a separate scale for the 10th decile (right hand panel of the Figure).

The similarity in pattern of prediction error by decile of predicted cost makes sense

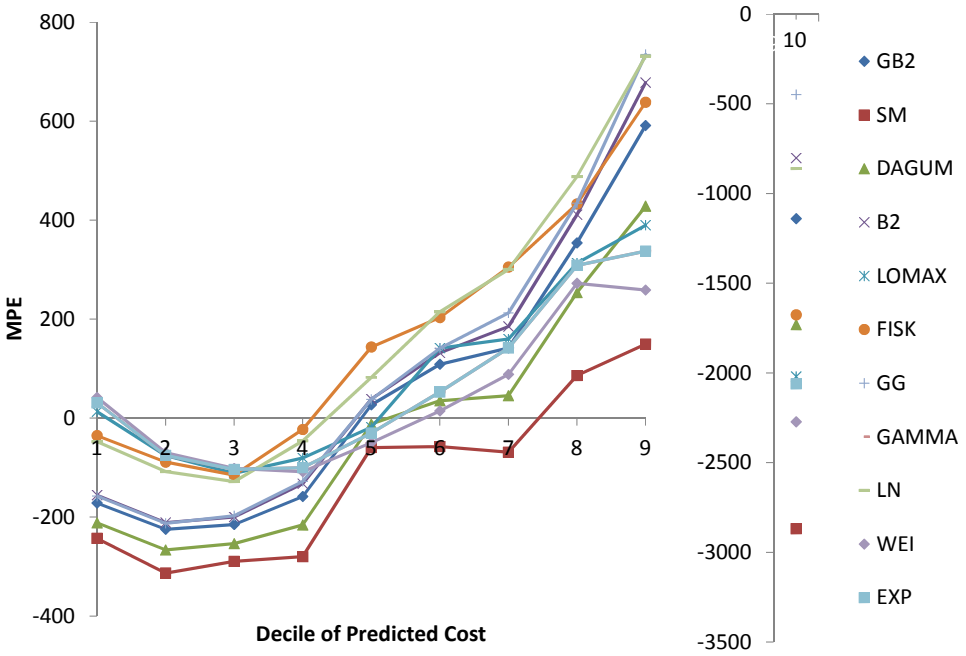


Figure 2.5: Mean prediction error for each model by each decile of predicted level of cost

given that all of these models use the same log link function and are nested or limiting cases of the GB2 distribution. Since the link function is the same for all models, it is unlikely that the observed systematic pattern causes bias in favour of one distribution over another. However, it is still worth noting that there are observable differences between the distributions from this plot. With the exception of the SM and DAGUM distributions, all models over-predict costs in five deciles and under-predict in five. The DAGUM over-predicts in six of the deciles, and SM in eight of the deciles. SM performs the worst in the decile with largest costs: over-predicting, on average, by $£2,868$. The best performing distribution, in terms of over-prediction in decile with highest predicted costs, is GG over-predicting by $£448$ on average.

2.5.2 Validation Set Results

In displaying these results, we show commonly used metrics to evaluate the fitting of means. While a policymaker may only be interested in fitting the costs on average, in which case the informative metric is MPE, they may also be concerned with large errors for individual costs, in which case they may look at RMSE (where larger errors are more influential, since they are squared). For illustrative purposes, we display results for all

metrics, rather than specify a loss function for the policymaker, which would determine the relative importance of each metric. In addition, the economic importance of differences in values for each of these metrics is determined by the policymaker's loss function. It may be that a small loss of accuracy (MAPE) is compensated for by the gains to reduced bias (MPE). Since we do not impose a loss function on these results, we discuss each aspect of the fit of conditional means in turn. It is clear from the results in Table 2.6 that no single distribution dominates in all of the criteria and that there is a trade-off between accuracy, bias and goodness-of-fit.

	MPE (£)	**MAPE** (£)	**RMSE**	$\mathbf{R^2}$	**ADMPE** (£)
GB2	-71.64	1,800.49	4,881.26	0.17859	62.32
SM	-397.47	1,958.42	5,251.70	0.17025	77.48
DAGUM	-195.40	1,856.35	4,984.90	0.17580	51.43
B2	**-8.04**	1,772.24	4,818.32	0.18088	46.10
LOMAX	-130.91	1,808.06	4,997.68	0.18420	53.33
FISK	-25.14	1,782.56	5,028.90	0.17280	45.81
GG	39.52	1,752.94	**4,758.02**	0.18376	45.72
GAMMA	-151.29	1,819.84	4,985.93	0.18681	60.22
LN	60.04	**1,735.93**	4,830.02	0.18265	**42.72**
WEI	-193.53	1,842.73	5,021.77	**0.18708**	63.25
EXP	-151.29	1,819.84	4,985.93	0.18681	60.22

Table 2.6: Results of model performance, when all converged, sample size 5,000

B2 gives the least biased estimates, overestimating by £8 (0.3% of the mean of the population) on average over the replications. In this regard, SM performs the worst – overestimating by 15.2% of the population mean. All of the beta-family of distributions produce, on average, overestimates of the conditional mean, while GG and LN produce underestimates from the gamma-family. LN gives the most accurate results, with GG the second most accurate and SM the least accurate. In the case of LN, the mean absolute prediction error, averaged over replications, is £1,736 (66.5% of the population mean); for SM this is £1,958 (75.0% of the mean of the population). In terms of goodness of fit (RMSE), GG performs the best; B2 also performs well. It is worth highlighting that DAGUM does not perform especially well according to its performance on the tests using the validation set, despite the strong performance in tests relating to its log-likelihood.

Figures 2.6 and 2.7 show the performance metrics for each decile of the actual level of costs of the validation set (again displaying the tenth decile on a separate scale to the

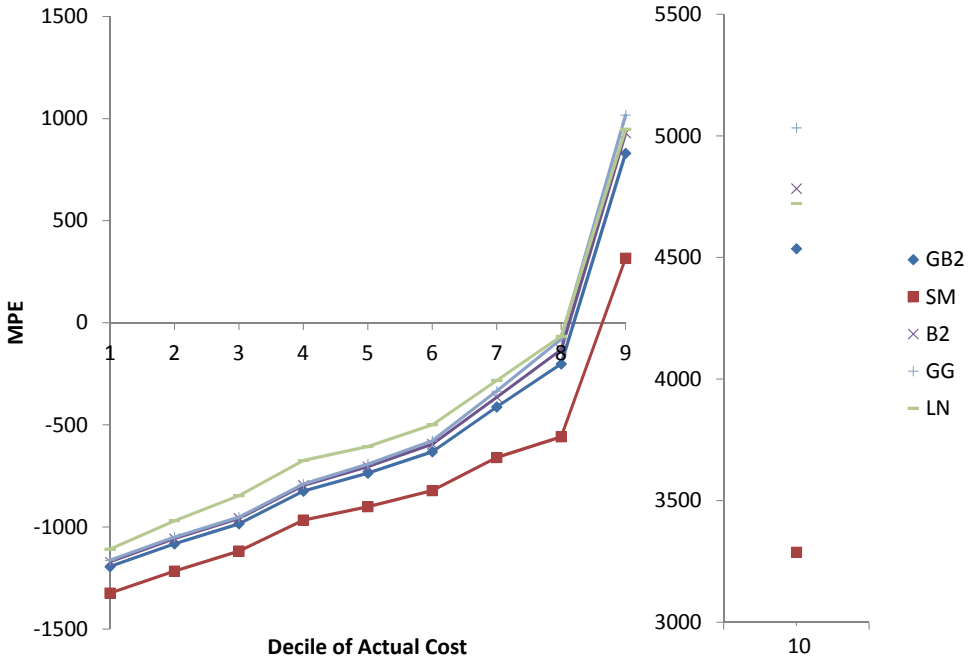


Figure 2.6: Mean prediction error for each model by each decile of level of cost

first nine deciles). This enables us to see how well our models peform on out-of-sample observations by each decile. For the sake of clarity, we display results for only five of the distributions: GB2, B2 (least biased model), SM (most biased and least accurate), GG (best goodness-of-fit) and LN (most accurate). Depending upon their loss function, a policymaker may place greater emphasis on deciles of higher costs than smaller ones. In modelling healthcare costs, for example, the decile containing the largest costs may be considered the most important, and as such models designed to cope with heavy-tailed data are increasingly popular.

Looking at the mean prediction error (Figure 2.6), 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: on average, the highest costs are underpredicted, and the lowest eight deciles overpredicted, by all models. 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 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.

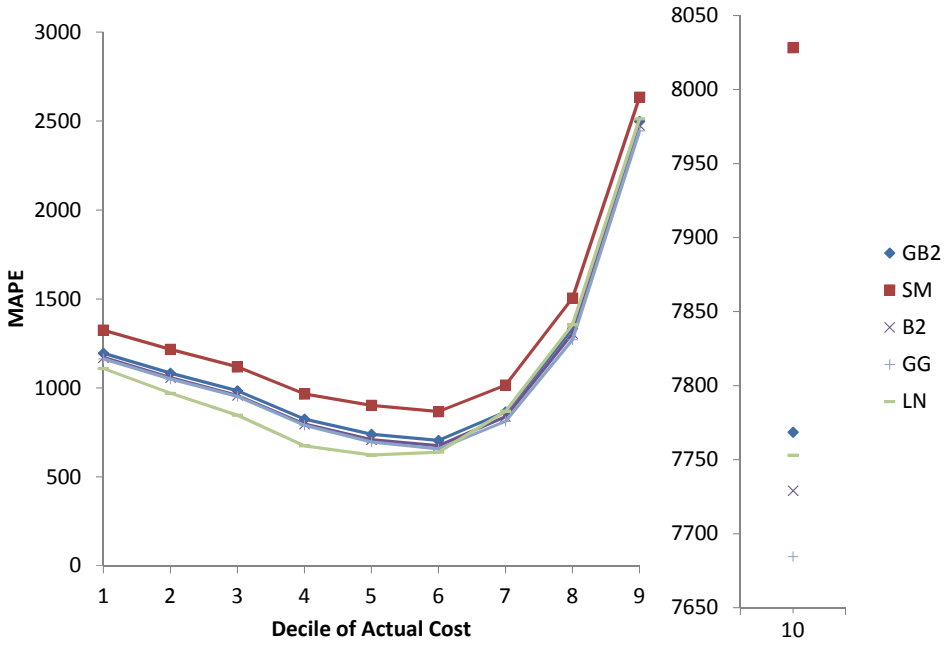


Figure 2.7: Mean absolute prediction error for each model by each decile of level of cost

In terms of fitting the individual expenditures, we present the mean average prediction error for each decile of actual costs (Figure 2.7). 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 forecasting high costs well on average, do a relatively poor job of forecasting individual costs within the highest decile.

Using the cumulative distribution function, we evaluate each model's distributional fit by comparing estimated proportions above a chosen value with actual proportions observed in our data. Here we plot the ratio of estimated to observed proportion above a selected value in order to see how the fit of these tail probabilities varies across the models, for the following levels of cost: £10,000, £15,000, £25,000, £50,000 and £100,000.

From Figure 2.8 it is clear that choice of distribution affects the estimated tail probabilities. In the case of SM, the model overestimates the mass for all levels of costs chosen here. All models of the beta-family overestimate the proportion of data at the highest chosen value of £100,000, while all gamma-family models underestimate this proportion with the exception of the EXP (not shown in Figure 2.8). LN fits the data best for tail probabilities for costs £15,000, £25,000 and £50,000, and GG best for £100,000. In

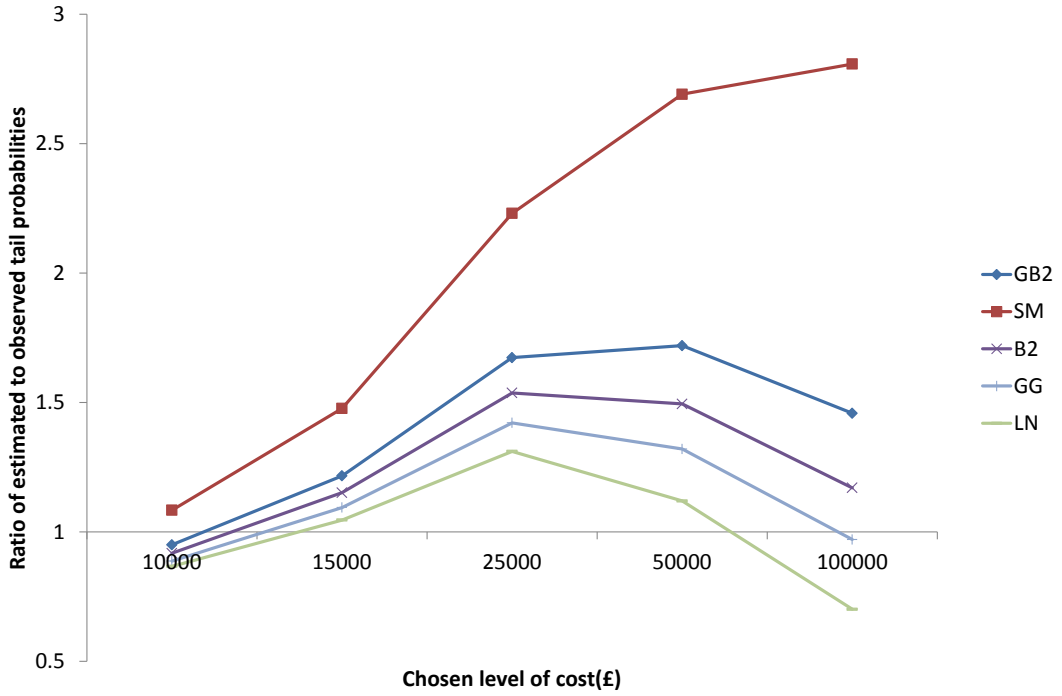


Figure 2.8: Estimated tail probabilities compared to observed actual proportions in data

Figure 2.8, GB2 predicts best the proportion above the value of £10,000, although this was best forecasted by the DAGUM distribution when all models are considered.

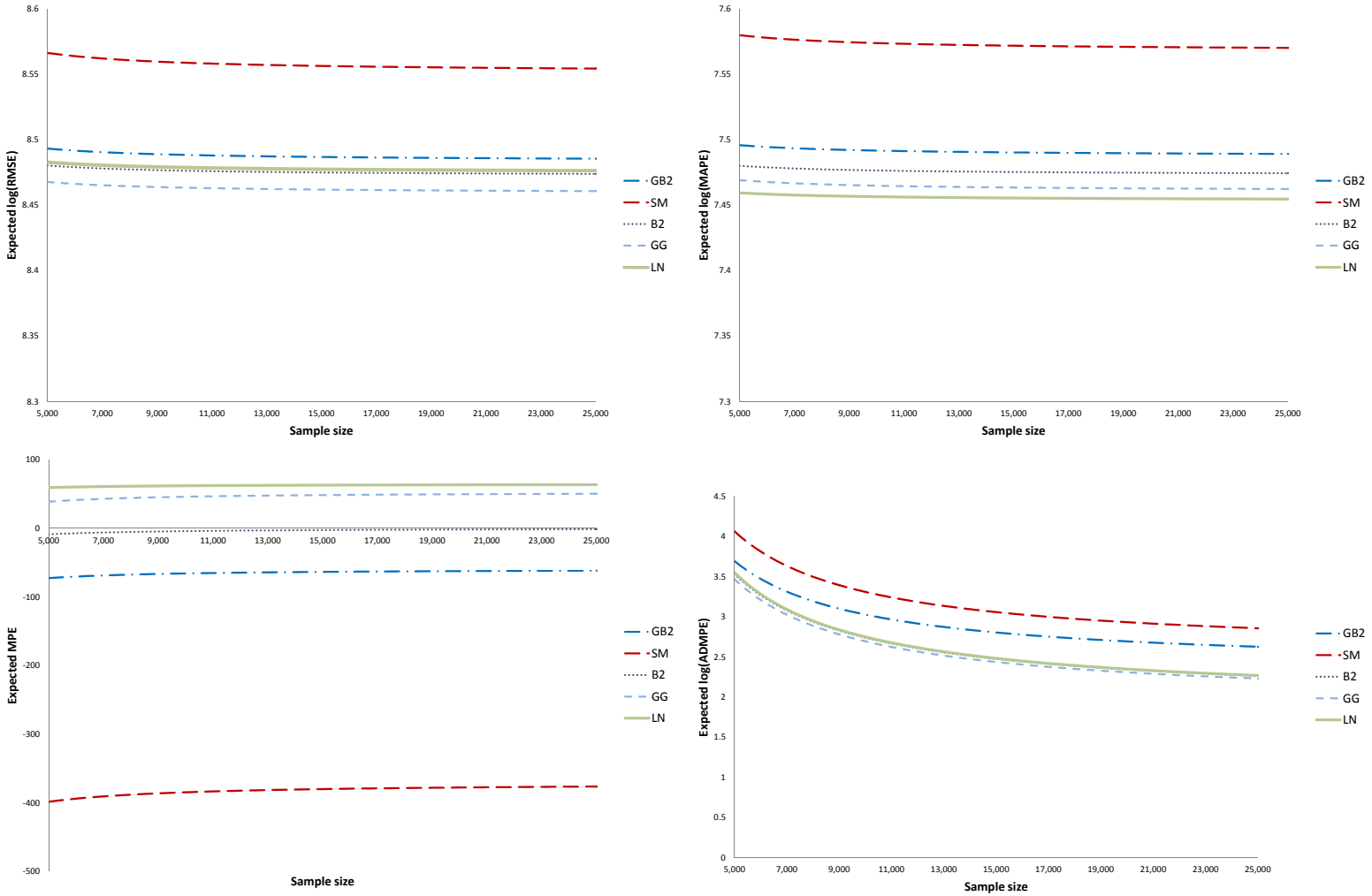


Figure 2.9: Reponse surfaces for log(RMSE), log(MAPE), log(ADMPE), MPE (clockwise from top left) against sample size, constructed evaluating performance on 'validation' set

As outlined in the Methodology section, we estimate these models with different sample sizes. We find that model performance is largely unaffected by sample size, with relative performance in terms of forecasted mean costs changing little (MPE, MAPE, RMSE and ADMPE). Some improvement is observed for all models with LADMPE (log(ADMPE)), suggesting that variation in bias reaches its minimum at a sample size of around 10,000. However, on the whole, these results suggest that these models are as appropriate at smaller samples (5,000) as larger ones. Figure 2.9 illustrates this in the form of response surfaces. With the exception of MPE for B2, all coefficient estimates for the constant term for all metrics of performance were significantly different from zero at the 5 percent level implying that, as the sample size approaches infinity, the metric of performance does not converge to zero.

In Deb and Burgess (2003), it is found that bias falls when models are estimated on an increasing number of observations. This fall in bias happens more slowly with more complex (finite mixture) models than with other estimated models. In our results, the performance of distributions with the most estimated parameters is not different to others, with increasing sample size. We suggest that this is because the estimation of a further two shape parameters (GAMMA to GB2) will cause only two additional degrees of freedom to be lost, compared to the loss of a further 34 degrees of freedom in the estimation of an additional component for a finite mixture model. This is because the latter requires the estimation of another coefficient for each independent variable, for the additional shape parameter and for the probability of class membership.

2.6 Conclusions

We estimate the GB2 distribution on English hospital inpatient data using maximum likelihood estimation, specifying the conditional mean as an exponential function of covariates, and evaluate its (and its nested and limiting cases) performance using a quasi-experimental design. The results suggest that there may be potential for the use of beta-type distributions in forecasting individual healthcare costs. In particular, B2 exhibits less bias than other models, without losing much accuracy. GG performs well in terms of accurately forecasting means, and also best forecasts the tail probability for costs

exceeding £100,000 from the models chosen, while LN forecasts tail probabilities best for other high costs (£15,000, £25,000 and £50,000). Unlike results obtained by Deb and Burgess (2003), the more complicated parametric distributions exhibit little evidence of worse performance due to smaller sample size.

In summary, the increased flexibility offered by the GB2 due to its additional parameters compared to its nested and limiting cases does not result in an improved forecasted fit over its competing models for the data used here. This may not be the case, however, for other healthcare cost data, or other model specifications. Hence, in a spirit similar to Manning et al. (2005), the GB2 could be used as a flexible distribution to allow the analyst to select among the competing distributions nested by it. In our illustration this is particularly useful, given that no single model dominates all criteria which may enter the policymaker's loss function.

Acknowledgements

The authors gratefully acknowledge funding from the Economic and Social Research Council (ESRC) under grant reference RES-060-25-0045. We are deeply grateful to John Mullahy for insightful comments and conversations and to Will Manning for detailed feedback, and to Partha Deb, Jed Frees and Dave Vanness for helpful suggestions. We are grateful 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, in particular John Forbes and Denzil Fiebig. We also thank Michael Ransom and other participants at the 4th Annual Health Econometrics Workshop, New York 2012, and the anonymous referees for their remarks.

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.

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. 2008. The generalized F distribution: An umbrella for parametric survival analysis. *Statistics in Medicine* **27**: 4301–4312.

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, Smith P, Gravelle H, Martin S, Bardsley M, Rice N, Georghiou T, Dusheiko M, Billings J, Lorenzo MD, Sanderson C. 2011. A person based formula for allocating commissioning funds to general practices in england: development of a statistical model. *BMJ* **343**: d6608.

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

Duan N, Manning WG, Morris CN, Newhouse JP. 1983. A comparison of alternative models for the demand for medical care. *Journal of Business & Economic Statistics* **1**: 115–126.

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.

Hosmer DW, Lemeshow S. 1980. Goodness of fit tests for the multiple logistic regression model. *Communications in Statistics - Theory and Methods* **9**: 1043.

Jenkins S. 2009. GB2FIT: stata module to fit generalized beta of the second kind distribution by maximum likelihood. *Statistical software components* **S456823**. 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.

Mandelbrot B. 1963. New methods in statistical economics. *Journal of Political Economy* **71**: 421–440.

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, Sorensen J, Turley PA. 2013. Skewness and kurtosis properties of income distribution models. *Review of Income and Wealth* **59**: 360–374.

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–S108.

Pregibon D. 1980. Goodness of link tests for generalized linear models. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* **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.

Vuong QH. 1989. Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica* **57**: 307–333.

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

2.7 Appendix A

We use the variables shown in Table 2A1 to construct our regression models. They are based on the ICD10 chapters, which are given in Table 2A2.

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 2A1: Classification of morbidity characteristics

ICD10 codes beginning with U were dropped because there were no observations in the 6,164,114 used. Only a small number (3,170) 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 2A2: ICD10 chapter codes

2.8 Appendix B

	Sample size	
Nested model	**50,000**	**100,000**
SM (1 restriction)	100%	100%
DAGUM (1 restriction)	100%	100%
B2 (1 restriction)	100%	100%
LOMAX (2 restrictions)	100%	100%
FISK (2 restrictions)	100%	100%

Table 2B1: Results of tests on nested model restrictions (percentage rejected at 5% significance level)

		Sample size		
Model		**10,000**	**50,000**	**100,000**
GB2	**AIC**	167,948	839,650	1,679,333
	BIC	168,200	839,958	1,679,666
SM	**AIC**	168,038	840,115	1,680,260
	BIC	168,283	840,415	1,680,584
DAGUM	**AIC**	167,959	839,717	1,679,466
	BIC	168,205	840,017	1,679,789
B2	**AIC**	167,954	839,685	1,679,404
	BIC	168,200	839,985	1,679,728
LOMAX	**AIC**	171,619	857,922	1,715,877
	BIC	171,857	858,213	1,716,191
FISK	**AIC**	168,224	841,013	1,682,054
	BIC	168,462	841,304	1,682,368
GG	**AIC**	167,981	839,829	1,679,696
	BIC	168,226	840,129	1,680,019
GAMMA	**AIC**	170,559	852,747	1,705,630
	BIC	170,797	853,038	1,705,944
LN	**AIC**	168,262	841,177	1,682,389
	BIC	168,500	841,468	1,682,703
WEI	**AIC**	171,385	856,974	1,714,108
	BIC	171,623	857,265	1,714,422
EXP	**AIC**	171,789	858,825	1,717,724
	BIC	172,020	859,107	1,718,029

Table 2B2: Values for each model's average AIC and BIC at sample sizes 10,000, 50,000 and 100,000

	Decile of predicted cost									
Model	1	2	3	4	5	6	7	8	9	10
GB2	-165.36	-236.15	-217.35	-179.12	65.46	87.79	151.18	370.57	622.78	-1,114.90
SM	-241.15	-310.65	-297.22	-295.49	-35.46	-60.56	-55.91	98.43	189.08	-2,789.97
DAGUM	-208.64	-277.35	-254.72	-238.95	17.73	28.13	51.03	258.65	450.82	-1,692.81
B2	-148.59	-223.62	-202.55	-157.54	75.62	121.36	194.53	416.25	710.65	-780.92
LOMAX	11.62	-78.19	-115.85	-93.83	-19.31	171.67	156.04	309.74	427.20	-2,035.92
FISK	-38.86	-88.57	-134.80	-26.57	143.05	214.11	294.99	448.97	661.05	-1,665.84
GG	-149.70	-225.25	-202.98	-151.99	72.42	133.39	211.06	456.06	775.33	-415.29
GAMMA	27.83	-65.39	-116.76	-102.36	-41.55	67.11	179.97	267.02	356.25	-2,062.45
LN	-49.64	-103.71	-150.41	-46.01	82.03	235.43	297.27	483.76	777.41	-863.41
WEI	39.61	-65.57	-118.35	-106.75	-56.13	16.83	140.25	220.60	278.74	-2,230.64
EXP	27.83	-65.39	-116.76	-102.36	-41.55	67.11	179.98	267.02	356.25	-2,062.45

Table 2B3: Models' average mean prediction error (£) by decile of predicted cost at sample size 10,000

	Decile of predicted cost									
Model	1	2	3	4	5	6	7	8	9	10
GB2	-156.95	-242.73	-226.12	-184.12	54.57	66.85	188.80	366.06	665.94	-1,110.68
SM	-238.41	-304.85	-313.72	-289.21	-32.44	-78.27	-23.56	98.20	233.75	-2,759.88
DAGUM	-204.83	-282.08	-268.40	-244.46	8.78	9.66	76.91	250.66	492.94	-1,704.27
B2	-139.19	-231.78	-212.98	-161.30	70.58	89.59	228.72	416.73	745.34	-790.08
LOMAX	11.78	-81.69	-116.04	-98.57	-21.16	183.39	139.97	331.27	439.83	-2,061.92
FISK	-37.57	-96.37	-140.17	-27.85	127.55	202.22	296.45	460.12	684.12	-1,681.48
GG	-136.87	-237.02	-213.42	-157.37	73.61	104.46	245.52	453.24	814.91	-410.70
GAMMA	27.02	-68.67	-123.93	-97.79	-37.02	78.69	175.22	265.66	376.97	-2,100.42
LN	-49.24	-108.78	-151.78	-46.89	78.61	228.43	283.82	497.98	800.61	-873.83
WEI	34.87	-67.58	-127.05	-99.83	-55.05	25.25	149.05	217.30	295.61	-2,241.03
EXP	27.02	-68.67	-123.93	-97.79	-37.02	78.69	175.22	265.66	376.97	-2,100.42

Table 2B4: Models' average mean prediction error (£) by decile of predicted cost at sample size 50,000

	Decile of predicted cost									
Model	1	2	3	4	5	6	7	8	9	10
GB2	-154.96	-244.64	-225.48	-184.78	55.22	65.42	193.13	369.10	663.77	-1,121.78
SM	-239.21	-303.40	-317.74	-285.17	-38.40	-65.00	-25.30	101.73	235.31	-2,768.65
DAGUM	-204.09	-282.23	-269.34	-244.74	6.37	13.97	80.08	256.20	490.15	-1,713.99
B2	-137.55	-233.35	-212.47	-161.43	69.36	91.74	232.01	419.79	742.63	-796.99
LOMAX	13.58	-82.83	-120.06	-95.92	-18.25	187.36	139.22	332.83	437.79	-2,068.10
FISK	-36.20	-96.87	-143.87	-24.95	127.51	205.05	296.49	464.95	679.76	-1,690.64
GG	-134.56	-239.89	-211.56	-158.18	73.71	107.39	246.48	456.74	814.38	-416.21
GAMMA	27.41	-68.03	-127.28	-98.60	-31.98	77.68	172.16	274.18	371.08	-2,088.70
LN	-47.65	-111.02	-151.30	-46.53	79.24	233.17	281.07	505.45	794.12	-880.61
WEI	33.62	-65.96	-130.80	-102.49	-50.09	22.44	155.65	219.45	295.47	-2,218.59
EXP	27.41	-68.03	-127.28	-98.60	-31.98	77.68	172.16	274.18	371.08	-2,088.71

Table 2B5: Models' average mean prediction error (£) by decile of predicted cost at sample size 100,000

	MPE (£)	**MAPE** (£)	**RMSE**	$\mathbf{R^2}$	**ADMPE** (£)
GB2	-67.85	1,794.14	4,862.61	0.18018	41.59
SM	-387.08	1,948.28	5,220.23	0.17181	55.52
DAGUM	-193.16	1,850.90	4,966.87	0.17733	37.94
B2	**-5.60**	1,766.92	4,803.31	0.18248	32.98
LOMAX	-132.10	1,803.23	4,977.98	0.18601	38.79
FISK	-26.18	1,778.86	5,011.06	0.17416	30.21
GG	44.43	1,746.55	**4,741.17**	0.18561	32.68
GAMMA	-153.58	1,812.89	4,956.13	0.18986	45.25
LN	60.26	**1,731.94**	4,815.65	0.18413	**29.81**
WEI	-192.24	1,831.96	4,974.51	**0.19119**	48.30
EXP	-153.58	1,812.89	4,956.13	0.18986	45.25

Table 2B6: Results of model performance, when all converged, at sample size 10,000

	MPE (£)	**MAPE** (£)	**RMSE**	$\mathbf{R^2}$	**ADMPE** (£)
GB2	-59.42	1,786.73	4,838.82	0.18164	18.51
SM	-372.74	1,936.75	5,180.76	0.17323	23.34
DAGUM	-188.19	1,844.85	4,942.92	0.17869	17.25
B2	**0.06**	1,760.95	4,784.22	0.18393	14.97
LOMAX	-128.25	1,796.51	4,950.71	0.18771	16.29
FISK	-22.97	1,773.56	4,984.77	0.17556	14.91
GG	52.23	1,739.61	**4,721.36**	0.18733	14.70
GAMMA	-150.98	1,804.06	4,922.69	0.19240	20.17
LN	64.56	**1,726.68**	4,795.68	0.18555	**13.74**
WEI	-187.19	1,819.81	4,928.19	**0.19455**	21.91
EXP	-150.98	1,804.06	4,922.69	0.19240	20.17

Table 2B7: Results of model performance, when all converged, at sample size 50,000

	MPE ($£$)	**MAPE** ($£$)	**RMSE**	$\mathbf{R^2}$	**ADMPE** ($£$)
GB2	-59.50	1,786.16	4,835.72	0.18188	13.47
SM	-371.64	1,935.53	5,175.13	0.17347	17.42
DAGUM	-187.77	1,844.02	4,938.97	0.17893	11.75
B2	**0.40**	1,760.21	4,780.94	0.18419	9.84
LOMAX	-127.87	1,795.51	4,944.40	0.18806	11.54
FISK	-22.86	1,772.97	4,981.22	0.17576	9.88
GG	52.87	1,738.71	**4,717.85**	0.18764	9.62
GAMMA	-149.41	1,801.89	4,910.67	0.19309	14.28
LN	64.76	**1,726.02**	4,792.27	0.18581	**9.08**
WEI	-184.19	1,816.56	4,911.62	**0.19546**	15.80
EXP	-149.41	1,801.89	4,910.67	0.19309	14.28

Table 2B8: Results of model performance, when all converged, at sample size 100,000

	Decile of actual cost									
Model	1	2	3	4	5	6	7	8	9	10
GB2	-1,194.68	-1,082.86	-984.33	-824.72	-736.75	-631.72	-412.20	-201.57	830.17	4,535.73
SM	-1,324.18	-1,216.47	-1,118.42	-966.22	-900.65	-822.11	-659.55	-558.37	316.18	3,287.54
DAGUM	-1,251.33	-1,142.81	-1,046.99	-890.18	-808.12	-712.31	-511.07	-336.73	645.98	4,112.51
B2	-1,170.80	-1,057.34	-957.85	-796.61	-704.87	-595.00	-364.39	-132.28	929.54	4,782.94
LOMAX	-1,201.04	-1,021.40	-865.55	-678.00	-658.94	-581.21	-412.68	-285.27	568.69	3,844.23
FISK	-1,100.82	-966.12	-840.62	-677.48	-614.18	-514.92	-325.78	-157.44	782.77	4,178.59
GG	-1,163.18	-1,050.13	-952.13	-790.16	-692.29	-578.06	-334.82	-79.90	1,017.04	5,032.71
GAMMA	-1,218.69	-1,040.57	-883.03	-686.11	-671.20	-600.51	-432.06	-317.73	533.97	3,820.90
LN	-1,108.67	-969.95	-846.39	-674.30	-606.41	-499.79	-282.35	-65.69	948.12	4,721.60
WEI	-1,245.16	-1,063.27	-901.98	-699.77	-691.29	-628.14	-465.29	-368.78	462.60	3,683.85
EXP	-1,218.69	-1,040.57	-883.03	-686.11	-671.20	-600.51	-432.06	-317.73	533.97	3,820.90

Table 2B9: Models' average mean prediction error (£) by decile of actual cost at sample size 5,000

	Decile of actual cost									
Model	1	2	3	4	5	6	7	8	9	10
GB2	-1,191.00	-1,079.02	-981.23	-821.50	-733.71	-628.60	-408.36	-197.44	833.40	4,542.44
SM	-1,317.46	-1,209.51	-1,112.58	-959.76	-894.40	-815.07	-650.69	-546.96	329.29	3,318.77
DAGUM	-1,248.50	-1,139.88	-1,045.05	-888.12	-806.28	-710.43	-508.54	-334.17	647.21	4,115.04
B2	-1,167.96	-1,054.32	-955.52	-794.12	-702.68	-592.78	-361.63	-129.62	930.91	4,785.45
LOMAX	-1,199.75	-1,021.41	-865.60	-678.53	-658.82	-581.48	-413.55	-287.56	564.54	3,839.09
FISK	-1,100.48	-966.41	-842.13	-678.90	-615.42	-516.48	-326.83	-158.86	780.25	4,178.72
GG	-1,159.84	-1,046.50	-949.07	-786.76	-689.21	-574.67	-330.45	-74.73	1,022.38	5,046.99
GAMMA	-1,217.08	-1,041.51	-882.58	-686.20	-671.02	-601.68	-433.41	-321.54	526.85	3,810.22
LN	-1,107.76	-969.68	-846.88	-675.02	-606.69	-500.16	-282.17	-65.60	947.88	4,724.40
WEI	-1,242.06	-1,064.15	-901.23	-699.14	-690.14	-628.20	-464.68	-369.79	461.23	3,693.81
EXP	-1,217.08	-1,041.50	-882.58	-686.20	-671.02	-601.68	-433.41	-321.54	526.85	3,810.22

Table 2B10: Models' average mean prediction error (£) by decile of actual cost at sample size 10,000

	Decile of actual cost									
Model	1	2	3	4	5	6	7	8	9	10
GB2	-1,190.09	-1,076.76	-976.28	-817.79	-730.00	-623.79	-403.59	-189.26	845.75	4,581.28
SM	-1,313.72	-1,204.27	-1,104.36	-952.98	-887.09	-806.47	-641.43	-532.73	349.56	3,378.74
DAGUM	-1,249.43	-1,139.62	-1,042.24	-886.81	-804.91	-708.23	-506.81	-329.83	654.63	4,144.42
B2	-1,168.18	-1,053.19	-951.71	-791.53	-700.35	-589.54	-358.86	-124.40	939.04	4,813.29
LOMAX	-1,201.00	-1,021.29	-862.67	-677.80	-657.09	-579.83	-411.89	-283.82	571.10	3,859.90
FISK	-1,103.28	-968.41	-842.01	-680.21	-616.31	-516.95	-327.24	-156.46	786.23	4,210.42
GG	-1,159.42	-1,044.50	-944.13	-782.84	-685.73	-570.06	-326.02	-67.31	1,033.66	5,082.69
GAMMA	-1,217.95	-1,040.36	-878.27	-684.45	-668.06	-599.52	-431.42	-318.77	530.38	3,816.72
LN	-1,109.54	-970.55	-845.55	-675.15	-606.31	-499.19	-281.00	-61.56	955.92	4,754.37
WEI	-1,241.46	-1,061.80	-895.93	-696.51	-685.78	-624.67	-460.84	-364.80	468.38	3,709.81
EXP	-1,217.95	-1,040.35	-878.26	-684.45	-668.06	-599.52	-431.42	-318.77	530.39	3,816.71

Table 2B11: Models' average mean prediction error (£) by decile of actual cost at sample size 50,000

	Decile of actual cost									
Model	1	2	3	4	5	6	7	8	9	10
GB2	-1,190.81	-1,077.40	-976.46	-817.45	-730.32	-624.01	-404.10	-189.89	845.39	4,583.75
SM	-1,314.01	-1,204.44	-1,104.01	-952.03	-886.86	-805.96	-641.10	-532.09	351.04	3,385.75
DAGUM	-1,249.98	-1,140.09	-1,042.20	-886.20	-805.01	-708.14	-506.96	-329.89	655.09	4,148.80
B2	-1,168.76	-1,053.69	-951.74	-791.04	-700.48	-589.54	-359.06	-124.58	939.36	4,817.54
LOMAX	-1,201.80	-1,021.98	-863.11	-677.97	-657.58	-580.16	-412.36	-284.52	571.40	3,867.56
FISK	-1,103.78	-968.92	-842.23	-680.06	-616.61	-517.09	-327.65	-156.98	786.20	4,214.05
GG	-1,159.99	-1,044.94	-944.10	-782.27	-685.77	-569.98	-326.04	-67.20	1,034.43	5,088.67
GAMMA	-1,219.59	-1,041.62	-878.90	-684.58	-668.63	-599.72	-431.52	-318.61	532.87	3,834.43
LN	-1,110.17	-971.15	-845.81	-675.03	-606.62	-499.39	-281.38	-62.02	956.10	4,758.98
WEI	-1,243.19	-1,063.16	-896.54	-696.50	-686.21	-624.45	-460.23	-363.35	473.44	3,736.64
EXP	-1,219.59	-1,041.62	-878.90	-684.58	-668.63	-599.72	-431.52	-318.61	532.87	3,834.43

Table 2B12: Models' average mean prediction error (£) by decile of actual cost at sample size 100,000

	Decile of actual cost									
Model	1	2	3	4	5	6	7	8	9	10
GB2	1,194.68	1,082.86	984.33	824.73	739.83	705.41	865.00	1,334.82	2,498.19	7,768.57
SM	1,324.18	1,216.47	1,118.42	966.23	901.70	867.39	1,016.02	1,504.92	2,634.20	8,028.31
DAGUM	1,251.33	1,142.81	1,046.99	890.19	809.77	769.11	913.42	1,383.18	2,529.15	7,821.25
B2	1,170.80	1,057.34	957.85	796.62	708.71	675.75	838.85	1,305.62	2,475.31	7,728.98
LOMAX	1,201.04	1,021.40	865.56	678.58	691.61	737.37	1,003.46	1,498.31	2,581.42	7,791.23
FISK	1,100.82	966.12	840.62	677.60	633.76	664.33	920.77	1,437.57	2,616.28	7,960.11
GG	1,163.18	1,050.13	952.13	790.17	696.06	657.93	812.52	1,271.42	2,444.79	7,684.57
GAMMA	1,218.69	1,040.57	883.08	687.15	708.74	758.92	1,021.83	1,517.09	2,587.67	7,764.05
LN	1,108.67	969.95	846.39	674.41	623.39	638.64	869.31	1,354.81	2,512.71	7,753.06
WEI	1,245.16	1,063.27	902.09	701.44	731.73	786.66	1,050.06	1,548.78	2,607.71	7,779.48
EXP	1,218.69	1,040.57	883.08	687.15	708.74	758.92	1,021.83	1,517.09	2,587.67	7,764.05

Table 2B13: Models' average mean absolute prediction error (£) by decile of actual cost at sample size 5,000

	Decile of actual cost									
Model	1	2	3	4	5	6	7	8	9	10
GB2	1,191.00	1,079.02	981.23	821.51	736.53	701.72	861.06	1,329.77	2,490.17	7,742.94
SM	1,317.46	1,209.51	1,112.58	959.77	895.24	860.14	1,008.60	1,495.50	2,622.30	7,995.38
DAGUM	1,248.50	1,139.88	1,045.05	888.13	807.63	766.52	909.93	1,378.49	2,521.49	7,797.06
B2	1,167.96	1,054.32	955.52	794.13	706.30	672.82	835.61	1,301.40	2,468.28	7,706.37
LOMAX	1,199.75	1,021.41	865.61	678.97	689.75	735.16	1,000.76	1,494.46	2,573.97	7,761.92
FISK	1,100.48	966.41	842.13	679.00	633.85	663.30	918.22	1,433.61	2,608.45	7,935.57
GG	1,159.84	1,046.50	949.07	786.77	692.77	653.89	808.09	1,265.71	2,436.39	7,659.96
GAMMA	1,217.08	1,041.51	882.59	687.04	706.73	756.67	1,018.67	1,511.95	2,576.49	7,719.56
LN	1,107.76	969.68	846.88	675.11	622.60	636.81	866.17	1,350.53	2,505.47	7,730.53
WEI	1,242.06	1,064.15	901.28	700.42	728.32	782.42	1,043.82	1,539.23	2,590.65	7,716.30
EXP	1,217.08	1,041.50	882.59	687.04	706.73	756.67	1,018.67	1,511.95	2,576.49	7,719.56

Table 2B14: Models' average mean absolute prediction error (£) by decile of actual cost at sample size 10,000

	Decile of actual cost									
Model	1	2	3	4	5	6	7	8	9	10
GB2	1,190.09	1,076.76	976.28	817.80	732.51	696.40	855.56	1,321.05	2,479.34	7,714.84
SM	1,313.72	1,204.27	1,104.36	952.98	887.79	851.30	1,000.42	1,483.45	2,607.72	7,954.95
DAGUM	1,249.43	1,139.62	1,042.24	886.82	805.95	763.13	905.62	1,370.65	2,510.68	7,767.84
B2	1,168.18	1,053.19	951.71	791.54	703.67	668.77	831.27	1,294.07	2,458.72	7,681.67
LOMAX	1,201.00	1,021.29	862.67	678.10	686.19	731.12	996.03	1,486.79	2,563.29	7,727.85
FISK	1,103.28	968.41	842.01	680.30	633.01	660.98	914.08	1,425.66	2,596.77	7,903.40
GG	1,159.42	1,044.50	944.13	782.85	689.04	648.79	802.85	1,257.42	2,425.99	7,634.27
GAMMA	1,217.95	1,040.36	878.27	684.99	701.23	750.85	1,012.85	1,503.37	2,563.70	7,676.08
LN	1,109.54	970.55	845.55	675.23	620.72	633.57	861.32	1,342.60	2,495.20	7,704.44
WEI	1,241.46	1,061.80	895.94	697.24	720.69	774.28	1,035.33	1,527.44	2,573.70	7,659.06
EXP	1,217.95	1,040.35	878.27	684.99	701.23	750.85	1,012.85	1,503.37	2,563.70	7,676.08

Table 2B15: Models' average mean absolute prediction error (£) by decile of actual cost at sample size 50,000

	Decile of actual cost									
Model	1	2	3	4	5	6	7	8	9	10
GB2	1,190.81	1,077.40	976.46	817.45	732.82	696.43	855.49	1,320.09	2,477.72	7,710.18
SM	1,314.01	1,204.44	1,104.01	952.04	887.54	850.71	999.91	1,481.92	2,605.51	7,948.66
DAGUM	1,249.98	1,140.09	1,042.20	886.21	806.02	762.89	905.36	1,369.45	2,508.82	7,762.61
B2	1,168.76	1,053.69	951.74	791.05	703.81	668.62	831.01	1,292.89	2,456.92	7,676.84
LOMAX	1,201.80	1,021.98	863.11	678.27	686.41	730.95	995.48	1,485.62	2,560.92	7,719.77
FISK	1,103.78	968.92	842.23	680.15	633.11	660.80	913.80	1,424.81	2,595.35	7,898.95
GG	1,159.99	1,044.94	944.10	782.27	689.10	648.55	802.44	1,256.02	2,423.91	7,629.00
GAMMA	1,219.59	1,041.62	878.91	685.10	701.37	750.18	1,011.41	1,500.55	2,558.57	7,660.58
LN	1,110.17	971.15	845.81	675.11	620.88	633.42	860.96	1,341.55	2,493.49	7,699.55
WEI	1,243.19	1,063.16	896.55	697.19	720.49	773.00	1,032.91	1,523.10	2,566.47	7,638.26
EXP	1,219.59	1,041.62	878.91	685.10	701.37	750.18	1,011.41	1,500.55	2,558.57	7,660.58

Table 2B16: Models' average mean absolute prediction error (£) by decile of actual cost at sample size 100,000

	Chosen threshold values				
Model	10,000	15,000	25,000	50,000	100,000
GB2	0.95	1.22	1.67	1.72	1.46
SM	1.08	1.48	2.23	2.69	2.81
DAGUM	1.01	1.33	1.91	2.12	1.99
B2	0.92	1.15	1.54	1.49	1.17
LOMAX	1.13	1.42	1.89	1.75	1.20
FISK	0.94	1.20	1.63	1.62	1.28
GG	0.89	1.09	1.42	1.32	0.97
GAMMA	1.02	1.25	1.57	1.31	0.77
LN	0.87	1.05	1.31	1.12	0.70
WEI	1.10	1.36	1.74	1.50	0.93
EXP	1.13	1.41	1.85	1.67	1.10

Table 2B17: Estimated to observed tail probabilities at sample size 5,000

	Chosen threshold values				
Model	10,000	15,000	25,000	50,000	100,000
GB2	0.95	1.21	1.67	1.70	1.43
SM	1.08	1.47	2.22	2.66	2.76
DAGUM	1.01	1.33	1.90	2.11	1.97
B2	0.92	1.15	1.53	1.48	1.15
LOMAX	1.13	1.42	1.89	1.75	1.19
FISK	0.94	1.20	1.63	1.61	1.27
GG	0.88	1.09	1.41	1.30	0.94
GAMMA	1.03	1.25	1.57	1.30	0.75
LN	0.87	1.05	1.31	1.11	0.69
WEI	1.10	1.36	1.74	1.49	0.90
EXP	1.13	1.41	1.85	1.66	1.08

Table 2B18: Estimated to observed tail probabilities at sample size 10,000

	Chosen threshold values				
Model	10,000	15,000	25,000	50,000	100,000
GB2	0.95	1.21	1.65	1.67	1.39
SM	1.08	1.46	2.19	2.62	2.70
DAGUM	1.00	1.32	1.89	2.09	1.95
B2	0.91	1.14	1.51	1.46	1.12
LOMAX	1.13	1.42	1.88	1.74	1.17
FISK	0.94	1.20	1.62	1.60	1.25
GG	0.88	1.08	1.39	1.26	0.90
GAMMA	1.03	1.25	1.56	1.28	0.73
LN	0.86	1.04	1.30	1.10	0.68
WEI	1.10	1.36	1.73	1.47	0.88
EXP	1.13	1.41	1.84	1.64	1.05

Table 2B19: Estimated to observed tail probabilities at sample size 50,000

	Chosen threshold values				
Model	10,000	15,000	25,000	50,000	100,000
GB2	0.95	1.21	1.65	1.67	1.39
SM	1.08	1.46	2.19	2.62	2.70
DAGUM	1.00	1.32	1.89	2.08	1.94
B2	0.91	1.14	1.51	1.45	1.12
LOMAX	1.13	1.42	1.88	1.73	1.16
FISK	0.94	1.20	1.62	1.59	1.25
GG	0.88	1.08	1.38	1.26	0.90
GAMMA	1.02	1.24	1.55	1.27	0.72
LN	0.86	1.04	1.29	1.09	0.67
WEI	1.10	1.35	1.72	1.46	0.86
EXP	1.13	1.41	1.84	1.63	1.04

Table 2B20: Estimated to observed tail probabilities at sample size 100,000

	Response surface regressions			
Regression coefficient	**MPE**	**log(MAPE)**	**log(RMSE)**	**log(ADMPE)**
α	-59.01 (2.17)	7.49 (0.00)	8.48 (0.00)	2.36 (0.08)
β	-67,877.12 (41,670.39)	41.87 (10.44)	48.47 (12.11)	6,672.10 (849.97)

Table 2B21: Regression coefficients for GB2 response surface regressions

	Response surface regressions			
Regression coefficient	**MPE**	**log(MAPE)**	**log(RMSE)**	**log(ADMPE)**
α	-370.76 (2.77)	7.57 (0.00)	8.55 (0.00)	2.55 (0.08)
β	-139,152.2 (51,537.13)	60.77 (14.17)	74.95 (17.69)	7,547.27 (731.01)

Table 2B22: Regression coefficients for SM response surface regressions

	Response surface regressions			
Regression coefficient	**MPE**	**log(MAPE)**	**log(RMSE)**	**log(ADMPE)**
α	-187.71 (2.39)	7.52 (0.00)	8.50 (0.00)	2.17 (0.09)
β	-41,486.18 (34099.92)	34.54 (8.92)	47.39 (12.74)	7,614.33 (779.22)

Table 2B23: Regression coefficients for DAGUM response surface regressions

	Response surface regressions			
Regression coefficient	**MPE**	**log(MAPE)**	**log(RMSE)**	**log(ADMPE)**
α	0.52 (1.67)	7.47 (0.00)	8.47 (0.00)	1.95 (0.10)
β	-46,299.87 (30,371.41)	35.63 (7.38)	40.20 (9.49)	7,924.93 (848.69)

Table 2B24: Regression coefficients for B2 response surface regressions

	Response surface regressions			
Regression coefficient	**MPE**	**log(MAPE)**	**log(RMSE)**	**log(ADMPE)**
α	-128.33 (1.91)	7.49 (0.00)	8.51 (0.00)	2.09 (0.09)
β	-17,618.43 (34,299.08)	36.08 (9.61)	53.58 (15.62)	8,554.69 (708.65)

Table 2B25: Regression coefficients for LOMAX response surface regressions

	Response surface regressions			
Regression coefficient	**MPE**	**log(MAPE)**	**log(RMSE)**	**log(ADMPE)**
α	-23.15 (1.61)	7.48 (0.00)	8.51 (0.00)	1.95 (0.09)
β	-13,794.05 (29,379.64)	28.29 (8.01)	49.11 (13.30)	8,111.43 (736.21)

Table 2B26: Regression coefficients for FISK response surface regressions

	Response surface regressions			
Regression coefficient	**MPE**	**log(MAPE)**	**log(RMSE)**	**log(ADMPE)**
α	53.20 (1.66)	7.46 (0.00)	8.46 (0.00)	1.92 (0.09)
β	-72,061.14 (30,246.83)	42.70 (7.18)	43.89 (8.41)	7,728.32 (903.79)

Table 2B27: Regression coefficients for GG response surface regressions

	Response surface regressions			
Regression coefficient	**MPE**	**log(MAPE)**	**log(RMSE)**	**log(ADMPE)**
α	-150.62 (2.28)	7.50 (0.00)	8.50 (0.00)	2.40 (0.08)
β	-8,535.90 (40,424.21)	50.19 (11.72)	73.68 (19.42)	7,014.64 (787.65)

Table 2B28: Regression coefficients for GAMMA response surface regressions

	Response surface regressions			
Regression coefficient	**MPE**	**log(MAPE)**	**log(RMSE)**	**log(ADMPE)**
α	64.61 (1.51)	7.45 (0.00)	8.47 (0.00)	1.95 (0.08)
β	-26,741.8 (27,319.42)	30.07 (6.68)	40.57 (9.52)	8,006.19 (669.79)

Table 2B29: Regression coefficients for LN response surface regressions

	Response surface regressions			
Regression coefficient	**MPE**	**log(MAPE)**	**log(RMSE)**	**log(ADMPE)**
α	-185.56 (2.45)	7.50 (0.00)	8.50 (0.00)	2.48 (0.08)
β	-45,269.77 (42,870.22)	72.22 (13.20)	107.22 (23.71)	6,635.87 (756.65)

Table 2B30: Regression coefficients for WEI response surface regressions

	Response surface regressions			
Regression coefficient	**MPE**	**log(MAPE)**	**log(RMSE)**	**log(ADMPE)**
α	-150.62 (2.28)	7.50 (0.00)	8.50 (0.00)	2.40 (0.08)
β	-8,535.71 (40,424.23)	50.19 (11.72)	73.68 (19.42)	7,014.82 (787.57)

Table 2B31: Regression coefficients for EXP response surface regressions

Chapter 3

Comparison of Developments

A quasi-Monte Carlo comparison of developments in parametric and semi-parametric regression methods for heavy tailed and non-normal data: with an application to healthcare costs

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

[a] *Department of Economics and Related Studies, University of York, York, YO10 5DD, UK*
[b] *Centre for Health Economics, University of York, York, YO10 5DD, UK*
[c] *Oxford Outcomes, 688 W. Hastings Street, Suite 450, Vancouver, British Columbia, V6B 1P1, Canada*

Summary

We conduct a quasi-Monte Carlo comparison of the recent developments in parametric and semi-parametric regression methods for healthcare costs, both against each other and against standard practice. The population of English NHS hospital inpatient episodes for the financial year 2007-2008 (summed for each patient: 6,164,114 observations in total) is randomly divided into two equally sized sub-populations to form an estimation set and a validation set. Evaluating out-of-sample using the validation set, a conditional density approximation estimator shows considerable promise in forecasting conditional means, performing best for accuracy of forecasting and amongst the best four (of sixteen compared) for bias and goodness-of-fit. The best performing model for bias is linear regression with square root transformed dependent variable, while a generalised linear model with square root link function and Poisson distribution performs best in terms of goodness-of-fit. Commonly used models utilising a log link are shown to perform badly relative to other models considered in our comparison.
JEL classification: C1; C5

Key words: Health econometrics; healthcare costs; heavy tails; quasi-Monte Carlo

*Corresponding author: *E-mail address:* james.lomas@york.ac.uk

3.1 Introduction

The distribution of healthcare costs provides many challenges to the applied researcher: values are non-negative (often with many observations with costs of zero), heteroskedastic, positively skewed and leptokurtic. While these, or similar, challenges are found within many areas of empirical economics, the large interest in modelling healthcare costs has driven the development of an expanding array of estimation approaches and provides a natural context to compare methods for handling heavy-tailed and non-normal distributions. Econometric models of healthcare costs include applications to risk adjustment in insurance schemes (Van de Ven and Ellis, 2000); in devolving budgets to healthcare providers (e.g. Dixon et al., 2011); in studies calculating attributable healthcare costs to specific health factors or conditions (Johnson et al., 2003; Cawley and Meyerhoefer, 2012) and in identifying treatment costs in health technology assessments (Hoch et al., 2002).

In attempting to capture the complex distribution of healthcare costs, two broad modelling approaches have been pursued. The first consists of flexible parametric models – distributions such as the three-parameter generalised gamma and the four-parameter generalised beta of the second kind. This approach is attractive because of the range of distributions that these models encompass, whereas models with fewer parameters are inherently more restrictive, especially in regard to the assumptions they impose upon higher moments of the distribution (e.g. skewness and kurtosis). The second is the use of semi-parametric models including extended estimating equations, finite mixture models and conditional density approximation estimators. The extended estimating equations model (EEE) adopts the generalised linear models framework and allows for the link and distribution functions to be estimated from data, rather than specified *a priori*. Finite mixture models introduce heterogeneity (both observed and unobserved) through mixtures of distributions. Conditional density approximation estimators are implemented by dividing the empirical distribution into discrete intervals and then decomposing the conditional density function into 'discrete hazard rates'. Despite the burgeoning availability of healthcare costs data via administrative records, together with an increased necessity for policymakers to understand the determinants of healthcare costs and more, it is surprising that no previous study compares comprehensively the models belonging to these

two strands of literature. In this paper we compare these approaches both to each other and against standard practice: linear regression on levels, and on square root and log transformations, of costs and generalised linear models (GLM).

Traditional Monte Carlo simulation approaches would not be appropriate for such an extensive comparison, as we are interested in a very large number of permutations of assumptions underlying the distribution of the outcome variable. In addition, such studies are prone to affording advantage to certain models arising from the chosen distributional assumptions used for generating data. Instead, using a large administrative database consisting of the population of English NHS hospital inpatient users for the year 2007-2008 (6,164,114 unique patients), we adopt a quasi-Monte Carlo approach where regression models are estimated on observations from one sub-population and evaluated on the remaining sub-population. This enables us to evaluate the regression methods in a rigorous and consistent manner – whilst ensuring results are not driven either by overfitting to rare but influential observations, or traditional Monte Carlo distributional assumptions – and are generalisable to hospital inpatient services.

This paper compares and contrasts systematically these recent developments[1] in semi-parametric and fully parametric modelling both against each other and against standard practice. No comprehensive empirical comparison of these methods is currently present in existing literature, and given the number of choices available for modelling heavy-tailed, non-normal data, this study makes an important contribution towards forming a ranking of possible approaches (for a similar study comparing propensity score methods, see Huber et al. (2013)).[2] The focus of this paper is the performance of these models in terms of predicting the conditional mean, given its importance in informing policy in healthcare and its prominence in comparisons between econometric methods in healthcare cost regressions.[3] Given our focus, we analyse bias, accuracy and goodness of fit of forecasted conditional means. We find that no model performs best across all metrics of evaluation. Commonly

[1] More strictly speaking, recent developments that have featured in a Monte Carlo, cross-validation or quasi-Monte Carlo empirical comparative study. An example of a promising method that is not compared is the extension to GLM proposed by Holly et al. (2011) – the fourth order pseudo maximum likelihood method – which has been applied to healthcare costs in Holly (2009).

[2] Mihaylova et al. (2011) provide an excellent review of statistical methods for the analysis of healthcare cost data with an emphasis on data collected alongside randomised trials.

[3] If the policymaker has a sufficiently large budget, Arrow and Lind (1970) argue that the policymaker should focus on mean outcomes. Other features of the distribution may be of interest (Vanness and Mullahy, 2007), especially when the policymaker has a smaller budget to allocate to healthcare.

used approaches – linear regression on levels of costs, linear regression on log-transformed costs, the use of gamma GLM with log link, and the use of the log-normal distribution – are not among the four best performing approaches with any of our chosen metrics.[4] Our results indicate that models estimated with a square root link perform much better than those with log or linear link functions. We find that linear regression with a square root-transformed dependent variable is the best performing model in terms of bias; the conditional density approximation estimator (using multinomial logit) for accuracy; and the Poisson GLM with square root link best in terms of goodness of fit.

3.2 Previous comparative studies

A number of studies have compared the performance of regression-based approaches to modelling healthcare cost data, where model performance is assessed on either actual costs (that is, costs with an unknown true distribution) (Deb and Burgess, 2003; Veazie et al., 2003; Buntin and Zaslavsky, 2004; Basu et al., 2006; Hill and Miller, 2010; Jones et al., 2014) or simulated costs from an assumed distribution (Basu et al., 2004; Gilleskie and Mroz, 2004; Manning et al., 2005). Using actual costs preserves the true empirical distribution of cost data, and all of its complexities, while simulating costs provides a benchmark using the known parameters of the assumed distribution (classic Monte Carlo) against which models can be compared.

Studies based on the classic Monte Carlo design are therefore ideally suited to assessing whether or not regression methods can fit data when specific assumptions, and permutations thereof, are imposed or relaxed. The complexities of the observed distribution of healthcare costs are such that a comprehensive comparison of modelling approaches would require an infeasibly large number of permutations of distributional assumptions used to generate data to make a classic Monte Carlo simulation worthwhile. Choosing a subset of the possible permutations of assumptions is prone to cause bias the results in favour of certain methods. A reliance on actual data, as an alternative approach, requires large datasets so that forecasting is evaluated on sufficient observations to credibly reflect all of the idiosyncratic features of cost data. With this approach, however, it is difficult to

[4] Linear regression on levels of costs performs well in terms of bias, the fifth best among models compared, but is the worst in terms of accuracy.

assess exactly which aspect of the distribution of healthcare costs is problematic for each method under comparison.

3.2.1 Studies using cross-validation approaches

With improvements in computational capacity, there has recently been a number of papers using large datasets to perform quasi-Monte Carlo comparisons across regression models for healthcare costs. Quasi-Monte Carlo comparisons divide the data into two groups, with samples repeatedly drawn from one group and models estimated, while the other group is used to evaluate out-of-sample performance (using the coefficients from the estimated models).

Deb and Burgess (2003) examine a number of models to predict total healthcare expenditures using a quasi-Monte Carlo approach with data from US Department of Veterans Affairs (VA) comprised of approximately 3 million individual records. From within these observations a sub-group of 1.5 million individual records is used as an 'estimation' group and another sub-group of 1 million records formed a 'prediction' group. They examine the predictive performance of models across different sizes of sample drawn from the 'estimation' group. For each sample drawn, model predictive performance is assessed on the full set of observations in the 'prediction' group according to mean prediction error (MPE), root mean squared error (RMSE), mean absolute prediction error (MAPE) and absolute deviations in mean prediction error (ADMPE). Using this methodology they are able to show that models based on a gamma density have better performance in forecasting individual costs than standard linear regression, with the most accurate individual forecasts coming from a finite mixture model with two gamma density components. In terms of bias, the use of linear regression (on levels and square root transformed levels of costs) performs best. The authors also note that the performance of finite mixture models in forecasting individual costs improves with increasing sample size, with MAPE between 10-15% lower than linear regression from sample sizes as large as 20,000 observations. Their results highlight a trade-off between bias and precision, and the need for caution surrounding the use of finite mixture models at smaller sample sizes.

Jones et al. (2014) focus exclusively on parametric models and suggest the use of the generalised beta of the second kind as an appropriate distribution for healthcare costs.

Their quasi-Monte Carlo design compares this distribution together with its nested and limiting cases, including the generalised gamma. Using data from Hospital Episode Statistics (HES) split into 'estimation' and 'validation' sets, they find little evidence of performance of models varying with sample size, but find variation between models in their ability to forecast mean costs, with generalised gamma the most accurate and beta of the second kind the least biased.

Hill and Miller (2010) and Buntin and Zaslavsky (2004) also use cross-validation techniques so that models are estimated on samples of data and evaluated on the remaining observations. Samples for estimation and the remaining data for evaluation differ across replications such that, unlike a quasi-Monte Carlo design, individuals may fall into either the estimation sample or the validation sample at each replication. This approach is less data intensive and providing sufficient replications should produce sufficient information in the evaluation exercise to judge model performance. The approaches are similar in that they both replicate the sampling process to ensure there is no 'lucky split', and guard against overfitting by evaluating out of sample.

Hill and Miller (2010) use the first eight waves of the Medical Expenditure Panel Survey (MEPS) dataset (from 1996-1997 to 2003-2004) to compare linear regression on untransformed and log transformed dependent variables, as well as Poisson and gamma GLMs with log link, EEE and a generalised gamma model. They examine four outcomes: total and prescription expenditures for privately insured adults (28,579 and 22,011 observations respectively) and elderly adults (12,547 and 11,671 observations respectively). For each outcome, 1,024 half samples were created for estimation and validation. Models using a log link are found to perform well in only one of these: total expenditures for privately insured, nonelderly adults; with this outcome the gamma GLM and generalised gamma model also perform well (in terms of MPE and MAPE). They show that the flexible link function of EEE improved goodness of fit, without inducing overfitting, in all four outcomes. In this way, Hill and Miller (2010) represents the first paper to compare common practice with the semi-parametric EEE model and the non-nested, fully parametric, generalised gamma model.

Buntin and Zaslavsky (2004) examine eight alternative estimators, comparing the performance of models with transformed dependent variables and GLMs (with log link). The

authors use data from the 1996 Medicare Current Beneficiary Survey (MCBS), taking 10,134 observations in total. This is split in half to form an estimation group and a validation group and repeated 100 times in total. They find that predictive performance is improved with careful consideration of the nature of heteroskedasticity. A GLM with variance proportional to the mean and using two smearing factors in a transformed dependent variable model are both found to be good choices for their application in terms of lower MAPE and mean squared forecast error (MSFE).

In Veazie et al. (2003) 500 half samples are drawn repeatedly from a dataset consisting of 8,495 observations from MCBS (risk adjusters from 1993 and expenditures from 1994). In addition, they compare models estimated on the years 1992-1993 (7,450 observations), and evaluated out-of-sample on the 1993-1994 observations, with the full samples bootstrapped 500 times to derive results. They find that with linear regression, a square root transformed dependent variable can reduce MAPE, but not necessarily MPE compared to using the level of costs.

Finally, Basu et al. (2006) compare EEE to linear regression with a log-transformed dependent variable, as well as GLM with log link and gamma variance, using data from Medstat's MarketScan database (final sample of 7,428 observations). Performance is mainly assessed in-sample, where the EEE performs well in terms of MPE across deciles of covariates. Split-sampling is used to perform tests of overfitting (Copas, 1983), with the authors finding little evidence of overfitting by the EEE approach compared to other approaches.

3.2.2 Recent developments in semi-parametric and fully parametric modelling

Figure 3.1 outlines the literature comparing regression models for healthcare costs as described above. As shown, there is no study that comprehensively and systematically evaluates all recent developments in approaches. In addition, any synthesis of the existing literature would be inconclusive in terms of which method is most appropriate for an application. Amongst the semi-parametric methods, EEE has never been directly compared in a rigorous evaluation against any of the finite mixture models. They have both separately been compared against standard practice (transformed dependent variable regression and GLM) in Basu et al. (2006); Hill and Miller (2010) and Deb and Burgess (2003), for EEE

and finite mixture models respectively. The conditional density approximation estimator, as yet, has not been compared with other healthcare cost regression models using actual data, although evidence from Monte Carlo studies suggests it to be a versatile approach (compared with standard practice methods) (Gilleskie and Mroz, 2004). Jones et al. (2014) introduce the use of the flexible parametric generalised beta of the second kind distribution with healthcare cost regressions and compare this against the generalised gamma which is a limiting case of the former. Given an increasing interest in modelling healthcare costs for resource allocation, risk adjustment and identifying attributable treatment costs, together with the burgeoning availability of data through administrative records, a comprehensive and systematic comparison of available approaches would appear timely. The results of this comparison will have resonance beyond healthcare costs and should be of interest to empirical applications to other right skewed, leptokurtic or heteroskedastic distributions such as income and wages.

	Studies using Monte Carlo			Studies using cross-validation				Studies using quasi-Monte Carlo		
	Basu et al. (2004)	Gilleskie and Mroz (2004)	Manning et al (2005)	Veazie et al (2003)	Buntin and Zaslavsky (2004)	Basu et al (2006)	Hill and Miller (2010)	Deb and Burgess (2003)	Jones et al (2013)	This paper
linear regression				■	■		■	■		■
linear regression (log)			■		■	■	■	■		■
linear regression (square root)				■	■			■		■
log-normal	■				■				■	■
gaussian GLM					■					(a)
Poisson					■		■			■
gamma	■	■	■		■	■	■	■	■	■
extended estimating equations						■	■			■
Weibull	■		■						■	(b)
generalised gamma			■				■		■	■
GB2									■	■
finite mixture of gammas								■		■
conditional density estimator		■								■

Figure 3.1: Models included in recent published comparitive work

(a) Not commonly used and problematic in estimation for our data in preliminary work.

(b) A special case of generalised gamma and generalised beta of the second kind which are included in our analysis.

3.3 Specification of models

We compare 16 different models applicable to healthcare cost data. Each makes different assumptions about the distribution of the outcome (cost) variable. Each regression uses the same vector of covariates X_i, although the precise way in which they affect the distribution varies across models. All models specify at least one linear index of covariates $X_i'\beta$. In addition, linear regression methods with transformed outcome require assumptions surrounding the form of heteroskedasticity (modelled as a function of X_i), in order to retransform predictions onto the natural cost scale (Duan, 1983). Within the GLM family, we explicitly model the mean and variance functions as some transformation of the linear predictor (Blough et al., 1999). Fully parametric distributions, such as the gamma- and beta-family of models, require an assumption about the form of the entire distribution. In this paper, a single parameter is estimated as a function of the linear index. Finite mixture models allow for multiple densities, each a function of the covariates in linear form. For conditional density approximation estimator models, the empirical distribution of costs is divided into intervals, and functions of the independent variables predict the probability of lying within each interval.

Beginning with linear regression, we estimate three models using ordinary least squares: the first is on the level of costs, the second and third use a log and square root transformed dependent variable respectively (log transformation is more commonly used in the literature (Jones, 2011)). With these approaches, predictions are generated on a transformed scale, and it is necessary to calculate an adjustment in order to retransform predictions to their natural cost scale. This is done by applying a smearing factor, which varies according to covariates in the presence of heteroskedasticity (Duan, 1983).

Given the complications in retransformation in the presence of heteroskedasticity, researchers more frequently use methods that estimate on the natural cost scale and explicitly model the variance as a function of covariates. The dominant approach that achieves these aims is the use of GLM (Blough et al., 1999). There are two components to GLM: the first is a link function that relates the index of covariates to the conditional mean, and the second is a distribution function that describes the variance as a function of the conditional mean. These are estimated simultaneously, using pseudo- or quasi-maximum

likelihood, leading to estimates that are consistent providing the mean function is correctly specified. Typically, the link function in applied work takes the form of a log or square root function. In this paper we consider two types of distribution function, each a power function of the conditional mean. In the Poisson case, the variance is proportional to the conditional mean function of covariates and in the gamma case the variance is proportional to the conditional mean squared. Two of the combinations of link functions and distribution families are associated with commonly used distributions. In particular, the GLM with log link and gamma variance is commonly applied to healthcare costs, and the GLM with a log link and Poisson variance is associated with the Poisson model (see discussion in Mullahy, 1997).

3.3.1 Flexible parametric models

Within the GLM and OLS approaches, much focus is placed on heteroskedacity and the form that it takes. Recent developments in fully parametric modelling have been made where the modeling of higher moments, skewness and kurtosis, is tackled explicitly. With this approach, the researcher estimates the entire distribution using maximum likelihood, which requires that the distribution is correctly specified for consistent results. If the distribution is correctly specified, then estimates are efficient.

Generalised gamma

We estimate two models from within the gamma-family, which have typically been used for durations, but also have precedent in the healthcare costs literature (Manning et al., 2005): the log-normal and generalised gamma distributions. Each of these is estimated, using maximum likelihood, with a scale parameter specified as an exponential function of covariates, denoted $\exp(X_i'\beta)$. The probability density function and conditional mean for the generalised gamma distribution are given below:

$$f(y_i|X_i) = \frac{\kappa\left(\kappa^{-2}\left(\frac{y_i}{\exp\left(X_i'\beta\right)}\right)^{\kappa/\sigma}\right)^{\kappa^{-2}} exp\left(-\kappa^{-2}\left(\frac{y_i}{\exp\left(X_i'\beta\right)}\right)^{\kappa/\sigma}\right)}{\sigma y_i \Gamma\left(\kappa^{-2}\right)} \tag{3.1}$$

$$E(y_i|X_i) = \left(\exp\left(X_i'\beta\right)\right)\left(\kappa^{2\sigma/\kappa}\right)\frac{\Gamma\left(\kappa^{-2}+\frac{\sigma}{\kappa}\right)}{\Gamma\left(\kappa^{-2}\right)} \tag{3.2}$$

where σ is a scale parameter, κ is a shape parameter and $\Gamma(.)$ is the gamma function

When $\kappa \to 0$ the generalised gamma distribution approaches the limiting case of the log-normal distribution, for which the probability density function and conditional mean are:

$$f(y_i|X_i) = \frac{1}{\sigma y_i\sqrt{2\pi}}\exp\left(\frac{-\left(\ln y_i - X_i'\beta\right)^2}{2\sigma^2}\right) \tag{3.3}$$

$$E(y_i|X_i) = \left(\exp\left(X_i'\beta\right)\right)\exp\left(\frac{\sigma^2}{2}\right) \tag{3.4}$$

Generalised beta of the second kind

We also include the generalised beta of the second kind, which has yet to be compared with a broad range of regression models.[5] Beta-type models, as gamma-type models, require assumptions about the form of the entire distribution. Until recently, they have been used largely in actuarial applications, as well as for the modelling of incomes (Cummins et al., 1990; Bordley et al., 1997). However, they have been suggested for use with healthcare costs due to their ability to model heavy tails, for example in Mullahy (2009), and have been used with healthcare costs in Jones et al. (2014). We include the generalised beta of the second kind, since all beta-type (and gamma-type) distributions are nested or limiting cases of this distribution. It therefore offers the greatest flexibility in terms of modelling healthcare costs amongst the duration models used here: see for example the implied restrictions on skewness and kurtosis (McDonald et al., 2013). The probability density function and conditional mean are:

$$f(y_i) = \frac{a y_i^{ap-1}}{b(X_i)^{ap}B(p,q)[1+(\frac{y_i}{b(X_i)})^a]^{(p+q)}} \tag{3.5}$$

$$E(y_i|X_i) = b(X_i)\left[\frac{\Gamma(p+\frac{1}{a})\Gamma(q-\frac{1}{a})}{\Gamma(p)\Gamma(q)}\right] \tag{3.6}$$

[5] In Jones et al. (2014), beta-type models are limited to comparison with gamma-type distributions.

where a is a scale parameter, p and q are shape parameters and $B(p,q) = \Gamma(p)\Gamma(q)/\Gamma(p+q)$ is the beta function.

We parameterise the generalised beta of the second kind with the scale parameter b as two different functions of covariates: a log link and a square root link.

3.3.2 Semi-parametric methods

Extended estimating equations

A flexible extension of GLM is proposed by Basu and Rathouz (2005) and Basu et al. (2006), known as the extended estimating equations (EEE). It approximates the most appropriate link using a Box-Cox function, where $\lambda = 0$ implies a log link and $\lambda = 0.5$ implies a square root link:

$$E(y_i|X_i) = (\lambda X_i'\beta + 1)^{\frac{1}{\lambda}} \tag{3.7}$$

as well as a general power function to define the variance with constant of proportionality θ_1 and power θ_2:

$$var(y_i|X_i) = \theta_1(E(y_i|X_i))^{\theta_2} \tag{3.8}$$

Suppose that the distribution of the outcome variable is unknown, but has mean and variance nested within (3.7) and (3.8). An incorrectly specified GLM mean function[6] yields biased and inconsistent estimates, while estimates from EEE should be unbiased, providing the specification of regressors is correct. A well-specified mean function combined with an incorrectly specified distribution form will be inefficient compared to EEE. If the distribution is known to be a specific GLM form, the EEE is less efficient than the appropriate GLM, but both are unbiased.

Finite mixture models

Finite mixture models have been employed in health economics in order to allow for heterogeneity both in response to observed covariates and in terms of unobserved latent

[6] In common usage GLM mean functions are limited to standard forms such as log and square root link function.

classes (Deb and Trivedi, 1997). Heterogeneity is modelled through a number of components, denoted C, each of which can take a different specification of covariates (and shape parameters, where specified), written as $f_j(y_i|X_i)$, and where there is a parameter for the probability of belonging to each component, π_j. The general form of the probability density function of finite mixture models is given as:

$$f(y_i|X_i) = \sum_{j}^{C} \pi_j f_j(y_i|X_i) \tag{3.9}$$

We use two gamma distribution components in our comparison.[7] In one of the models used, we allow for log links in both components (3.10), and in the other we allow for a square root link (3.11). In both, the probability of class membership is treated as constant for all individuals and a shape parameter, α_j, is estimated for each component.

$$f_j(y_i|X_i) = \frac{y_i^{\alpha_j}}{y_i \Gamma(\alpha_j) \exp(X_i'\beta_j)^{\alpha_j}} \exp\left(-\left(\frac{y_i}{\exp(X_i'\beta_j)}\right)\right) \tag{3.10}$$

$$f_j(y_i|X_i) = \frac{y_i^{\alpha_j}}{y_i \Gamma(\alpha_j) (X_i'\beta_j)^{2\alpha_j}} \exp\left(-\left(\frac{y_i}{(X_i'\beta_j)^2}\right)\right) \tag{3.11}$$

The conditional mean is given for the log link specification and for the square root link by (3.12) and (3.13) respectively:

$$E(y_i|X_i) = \sum_{j}^{C} \pi_j \alpha_j \exp(X_i'\beta_j) \tag{3.12}$$

$$E(y_i|X_i) = \sum_{j}^{C} \pi_j \alpha_j (X_i'\beta_j)^2 \tag{3.13}$$

Unlike the models in the previous section, this approach can allow for a multi-modal distribution of costs. In this way, finite mixture models represent a flexible extension of parametric models (Deb and Burgess, 2003). Using increasing numbers of components, it is theoretically possible to fit any distribution, although in practice researchers tend to use few components (two or three) and achieve good approximation to the distribution of

[7] Preliminary work showed that models with a greater number of components lead to problems with convergence in estimation. Empirical studies such as Deb and Trivedi (1997) provide support for the two components specification for healthcare use.

interest (Heckman, 2001).

Conditional density approximation estimators

Finally, we use two additional models that are applications of the conditional density approximation estimator outlined in Gilleskie and Mroz (2004). Their method is an extension of the two-part model frequently used to deal with zero costs, in that the range of outcome variable is divided into Q parts (or intervals), where the mean (of observations to be used in estimation) within interval j ($j = 1, ..., Q$) is $\overline{y}_j$ and the lower and upper threshold values are y_{j-1} and y_j, respectively[8]. The probability of an observation falling into interval j can be written as (3.14):

$$p_{ij}(X_i) = P(y_{j-1} \leq y_i < y_j | X_i) = \int_{y_{j-1}}^{y_j} f(y_i|X_i)dy_i \tag{3.14}$$

The density function is then approximated by Q 'discrete hazard rates', defined as the probability of lying in interval j conditional on not lying in intervals $1, ..., j-1$ and written as $\lambda(j, X_i)$ as shown in (3.15):

$$\lambda(j, X_i) = P(y_{j-1} \leq y_i < y_j | X_i, y_i \geq y_{j-1}) = \frac{\int_{y_{j-1}}^{y_j} f(y_i|X_i)dy_i}{1 - \int_{y_0}^{y_{j-1}} f(y_i|X_i)dy_i} \tag{3.15}$$

The effect of covariates can vary smoothly, or discontinuously, across intervals depending upon how the model is specified: with the most flexible case using a separate model for each interval's hazard rate. We assume that only the probability of lying within an interval depends upon covariates, and that the mean value of the outcome variable, for a given interval, does not vary with covariates. The conditional mean function is therefore obtained using (3.16):

$$E(y_i|X_i) = \sum_{j=1}^{Q} p_{ij}(X_i)\overline{y}_j \tag{3.16}$$

One of the main benefits of this approach is the flexibility afforded with respect to the intervals that are used. There is flexibility in terms of the number of intervals, and where the boundaries between them are placed, as well as the degree to which the 'dis-

[8] y_0 is equal to the lowest observed cost and y_Q is equal to the highest observed cost.

crete hazard rates' are estimated separately for each interval. Within our illustration, we use 15 equally sized intervals across all samples.[9] In practice, a researcher would experiment with different intervals and compare model performance in order to decide upon the specification. Having decided upon the intervals to be used, we use a multinomial logit specification and an ordered logit specification to model the probabilities of lying within each interval.[10] The multinomial logit specification is similar to running a separate logit model for each 'discrete hazard rate', whereas the ordered logit specification is analogous to allowing the 'discrete hazard rate' to vary discontinuously for each interval but with no discontinuity in the effects of covariates[11].

$$p_{ij}(X_i) = \frac{\exp\left(X_i'\beta_j\right)}{\sum_{l=1}^{Q}\exp\left(X_i'\beta_l\right)} \tag{3.17}$$

where $\beta_1 = 0$ to normalise for estimation purposes

$$p_{ij}(X_i) = \frac{\exp\left(\psi_j - X_i'\beta\right)}{1+\exp\left(\psi_j - X_i'\beta\right)} - p_{ij-1} \tag{3.18}$$

where ψ_j represents the estimated threshold value for each category from the ordered logit model, $p_{i0} = 0$ and $p_{i15} = 1 - p_{i14}$, so we only estimate 14 threshold values (in our application $Q = 15$)

Conditional means from these models are calculated as in (3.16), where the probabilities, p_{ij}, are calculated using (3.17) for the multinomial logit specification and (3.18) for the ordered logit specification.

3.4 Data and Choice of Variables

Our study uses individual-level data from the English Hospital Episode Statistics (HES) (for the financial year 2007-2008).[12] This dataset contains information on all

[9] Gilleskie and Mroz (2004) in their application to healthcare costs find that between 10 and 20 intervals results in a good aproximation, based on an adjusted log-likelihood to guard against overfitting, and we found 15 intervals to result in good convergence performance in preliminary work.

[10] This differs from the less parametric single logit specification adopted in Gilleskie and Mroz (2004), which is more computationally demanding, and instead uses an approach similar to Han and Hausman (1990).

[11] Gilleskie and Mroz (2004) also allow the data to determine how flexibly to estimate the 'discrete hazard rates', see paper for more details.

[12] In our dataset, episodes are grouped into spells, which can be thought of as discrete admissions for a patient.

inpatient episodes, outpatient visits and A&E attendances for all patients admitted to English NHS hospitals (Dixon et al., 2011). For our study, we exclude spells which were primarily mental or maternity healthcare, as well as private sector spells.[13] HES is a large administrative dataset collected by the NHS Information Centre[14], with our dataset comprising 6,164,114 separate observations, representing the population of hospital inpatient healthcare users for the year 2007-2008. Since data is taken from administrative records, we only have information on users of inpatient NHS services, and therefore can only model strictly positive costs.[15]

The cost variable used throughout is individual patient annual NHS hospital cost for all spells finishing in the financial year 2007-2008. In order to cost utilisation of inpatient NHS facilities, tariffs from 2008-2009[16] were applied to the most expensive episode within the spell of an inpatient stay (following standard practice for costing NHS activity). Then, for each patient, all spells occuring within the financial year were summed. The data are summarised in Table 3.1 and Figure 3.2.

The challenges of modelling cost data are clearly observed in Table 3.1 and in Figure 3.2[17]: the observed costs are heavily right-hand skewed (even after log transformation), with the mean far in excess of the median, and are highly leptokurtic (although roughly mesokurtic following log transformation). Whilst transforming the data clearly reduces skewness, neither of these transformations result in a completely symmetric distribution, implying that a flexible link function could be useful. The distribution may also be multimodal, or at least noisy with many spikes, which can most clearly be seen in Figure 3.2 on the histogram of the log transformed costs.

We construct a linear index of covariates and divide the data into quantiles according to

[13] This dataset was compiled as part of a wider project considering the allocation of NHS resources to primary care providers. Since a lot of mental healthcare is undertaken in the community and with specialist providers, and hence not recorded in HES, the data is incomplete, 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, and accordingly for the setting of healthcare budgets are determined using alternative mechanisms.

[14] Now named the Health and Social Care Information Centre.

[15] Zeros are typically handled by a two-part specification and the main challenge is to capture the long and heavy tail of the distribution rather than the zeros.

[16] Reference costs for 2005-2006, which were the basis for the tariffs from 2008-2009, were used when 2008-2009 tariffs were unavailable.

[17] Costs above $£30,000$ were excluded, for this figure only, to make the graphs clearer.

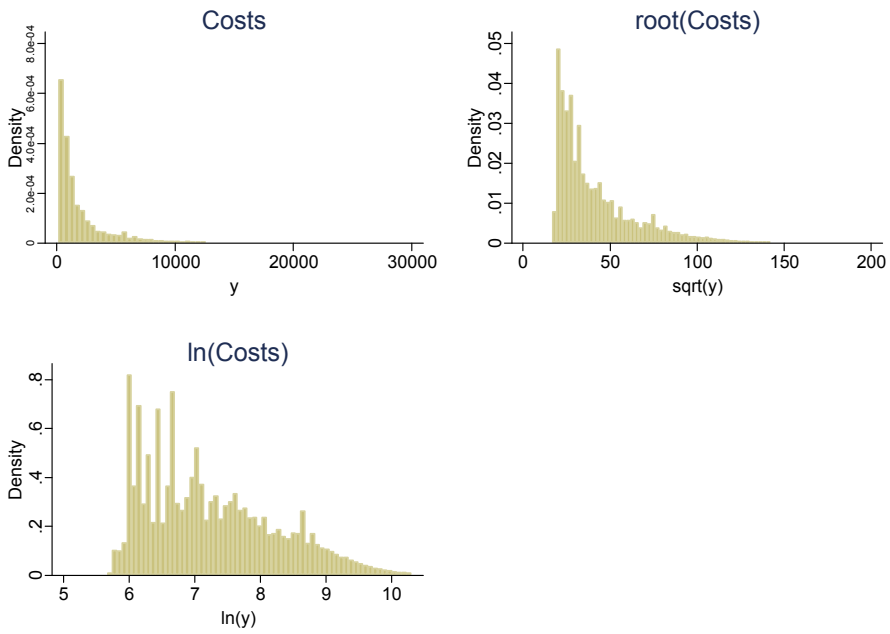


Figure 3.2: Histogram plots of costs

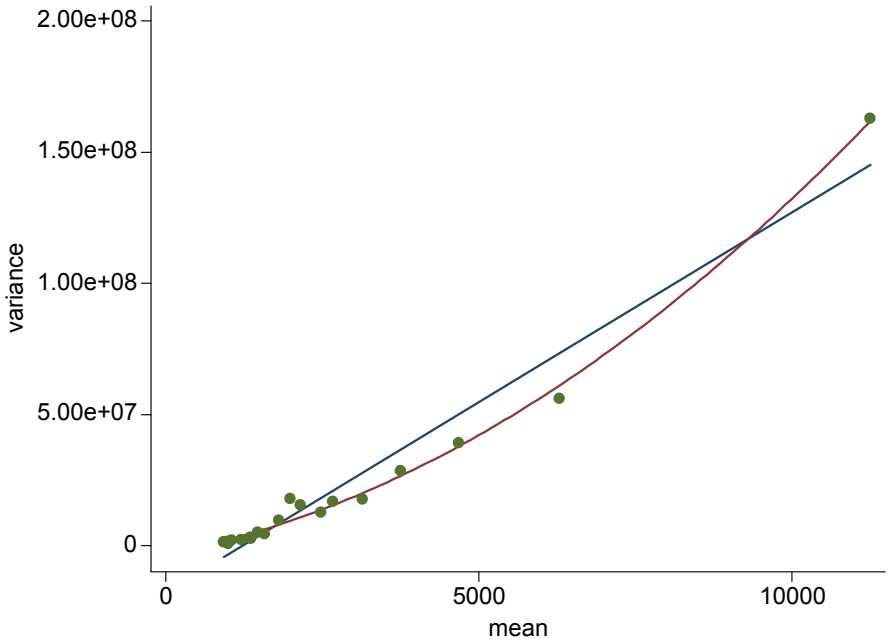


Figure 3.3: Variance against mean for each of the 20 quantiles of the linear index of covariates

Note:

The data were divided into twenty subsets using the deciles of a simple linear predictor for healthcare costs using the set of regressors introduced later. Figure 3.3 plots the means and variances of actual healthcare costs for each of these subsets, with fitted linear and quadratic trends.

	Level	Square root	Logarithm
N	6,164,114		
Mean	£2,610	43.18	7.25
Median	£1,126	33.56	7.03
Standard deviation	£5,088	27.30	1.00
Skewness	13.03	2.84	0.74
Kurtosis	363.18	19.62	2.99
Maximum	£604,701	777.63	13.31
99th percentile	£19,015	137.89	13.31
95th percentile	£8,956	94.64	9.10
90th percentile	£6,017	77.57	8.70
75th percentile	£2,722	52.17	7.91
25th percentile	£610	24.70	6.41
10th percentile	£446	21.12	6.10
5th percentile	£407	20.16	6.01
1st percentile	£347	18.63	5.85
Minimum	£217	14.73	5.38

Table 3.1: Descriptive statistics for hospital costs

this, to analyse conditional (on X) distributions of the outcome variable.[18] First, we plot the variances of each quantile against their means (Figure 3.3). This gives us a sense both of the nature of heteroskedasticity and of feasible assumptions relating these aspects of the distribution. From Figure 3.3, we can see that there is evidence against homoskedasticity (where there would be no visible trend), and evidence for some relationship between the variance and the mean.

We also carry out a similar analysis for higher moments of the distribution, plotting the kurtosis of each quantile against their skewness. Parametric distributions impose restrictions upon possible skewness and kurtosis: one-parameter distributions are restricted to a single point (e.g. normal distribution imposes a skewness of 0 and a kurtosis of 3), two-parameter distributions allow for a locus of points to be estimated, and distributions with three or more parameters allow for spaces of possible skewness and kurtosis combinations. Figure 3.4 shows that the data is non-normal and provides motivation for flexible methods since they appear better-able to model the higher moments of the conditional distributions of the outcome variables analysed here.[19]

[18] This is done by regressing the outcome variable on the set of covariates we include in our regression models using OLS.

[19] A similar analysis can be found in Pentsak (2007). Note also that the lower bound of the Pearson Type IV distribution, used in Holly and Pentsak (2006), is equal to the upper bound for the beta of the second kind distribution (also known as Pearson Type VI).

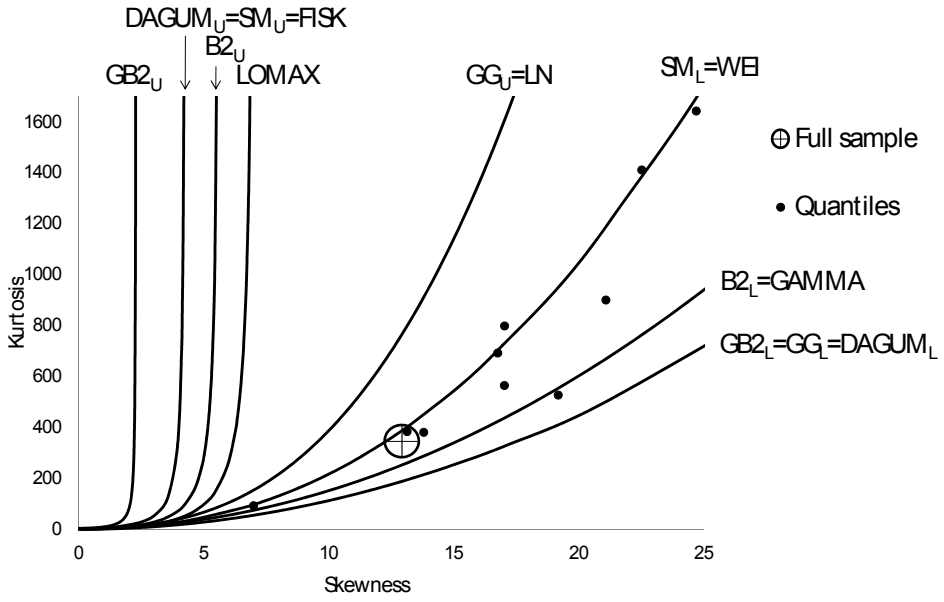


Figure 3.4: Kurtosis against skewness for each of the 10 quantiles of the linear index of covariates, adapted from McDonald et al. (2013)

Note:

The dots shown on Figure 3.4 were generated as follows: the data were divided into ten subsets using the deciles of a simple linear predictor for healthcare costs using the set of regressors used in this paper. Figure 3.4 plots the skewness and kurtosis coefficients of actual healthcare costs for each of these subsets, the skewness and kurtosis coefficient of the full estimation sub-population (represented by the larger circle with cross) and theoretically possible skewness-kurtosis spaces and loci for parametric distributions considered in the literature.

All of the models in the quasi-Monte Carlo comparison use a specified vector of covariates, and have at least one linear index of these. This vector mirrors the practice in the literature regarding comparing econometric methods for healthcare costs, allowing models to control for age (as well as age squared and age cubed), gender (interacted fully with age terms), and morbidity characteristics (from ICD classifications).[20] Each of the 24 morbidity markers indicates the presence or absence, coded 1 and 0 respectively, of one or more spells with any diagnosis within the relevant subset of ICD10 chapters, during the financial year 2007-2008 (see Appendix A). We do not use a fully interacted specification, since morbidity is modelled with a separate intercept for presence of each type of diagnosis (and not interacted with age or gender). However, we do allow for interactions between age and its higher orders and gender. This means that we are left with a specification close to those used in the comparative literature as well as a parsimonious version of the set of covariates used to model costs in Person-Based Resource Allocation in England, as in, for example, Dixon et al. (2011). In addition, making the specification less complicated aids computation and results in fewer models failing to converge.

3.5 Methodology

3.5.1 Quasi-Monte Carlo Design

By using the HES data, we have access to a large amount of observations representing the whole population of English NHS inpatient costs. To exploit this, we use a quasi-Monte Carlo design similar to Deb and Burgess (2003).[21] The population of observations (6,164,114) is randomly divided into two equally sized sub-populations: an 'estimation' set (3,082,057) and a 'validation' set (3,082,057).[22] 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 models are estimated on the samples and performance then evaluated on both the sample drawn from the 'estimation' set and the full 'validation' set. Figure 3.5

[20] Morbidity information is available through the HES dataset, adapted from the ICD10 chapters (WHO, 2007) – see Appendix A for further details.

[21] Using a split-sample to evaluate models has precedent in the comparative literature on healthcare costs, see Duan et al. (1983); Manning et al. (1987).

[22] Given the size of the dataset, any sub-optimality resulting from the proportions allocated to each set is likely to be minimal. To ensure the results are replicable, we set a fixed seed for splitting the dataset and for randomly drawing samples.

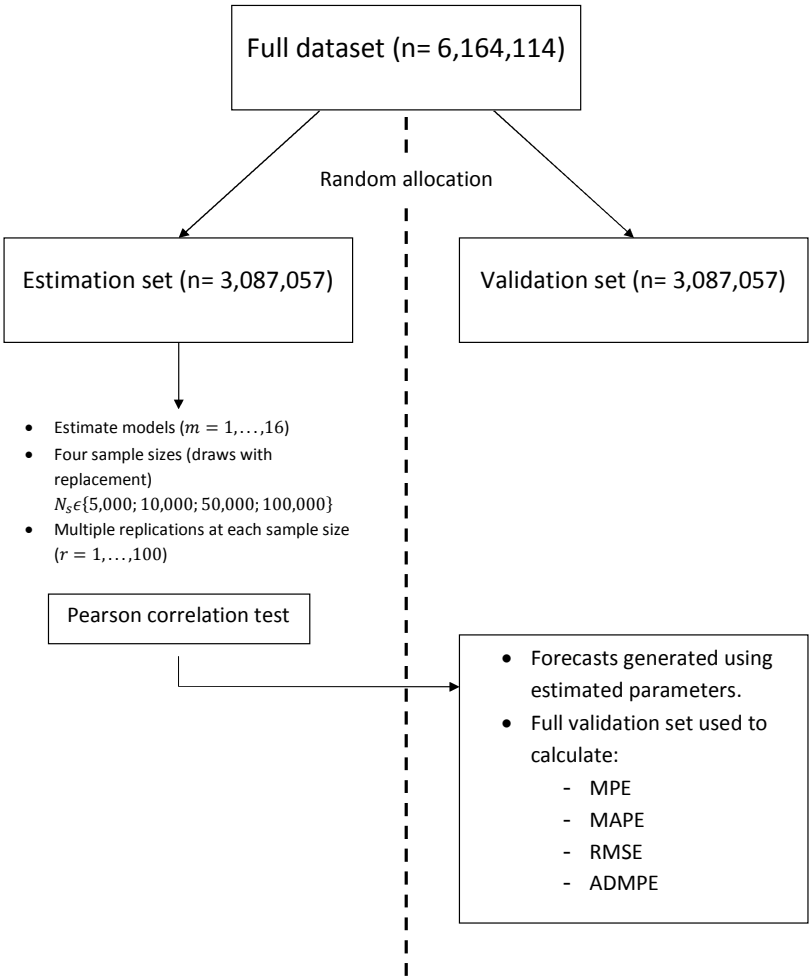


Figure 3.5: Diagram setting out study design

illustrates our study design in the form of a diagram: note the subscript m denotes the model used, N_s the sample size used, and r the replication number.

In order to execute this quasi-experimental design, we automate the model selection process for each approach: for instance, with the conditional density approximation estimator, we specify a number of bins to be estimated, a priori, rather than undergoing the investigative process outlined in Gilleskie and Mroz (2004). Similarly, all models have been automated to some extent, since we set *a priori*: the specification of regressors (all models), the parameters that vary with covariates (generalised gamma and generalised beta of the second kind), and the number of mixtures to model (finite mixture models). Our specification of regressors was based on preliminary work, which showed alternative

specifications to give similar results, but with worse convergence performance.[23]

3.5.2 Evaluation of Model Performance

Estimation Sample

Researchers modelling healthcare costs will typically carry out multiple tests to establish the reliability of their model specification. These tests are carried out in sample, and help to inform the selection of models that will then be used for predictive purposes. They are commonly used to build the specification of the 'right hand side' of the regression: the covariates used and interactions between them. In addition, researchers working with healthcare costs use these tests to establish the appropriate link function between covariates and expected conditional mean, and other assumptions about functional form. We include results from the Pearson correlation coefficient test, which is simple to carry out and has intuitive appeal.[24] In order to carry out the Pearson correlation coefficient test, residuals (computed on the raw cost scale) are regressed against predicted values of cost. If the slope coefficient on the predicted costs is significant, then this implies a detectable linear relationship between the residuals and the covariates, and so evidence of model misspecification.

Validation Set

We use our models to estimate forecasted mean healthcare costs over the year for individuals ($\hat{y}_i^V = E(\widehat{y_i^V|X_i^V})$[25], V denotes that the observation is from the 'validation' set) and evaluate performance on metrics designed to reflect the bias (mean prediction error, MPE), accuracy (mean absolute prediction error, MAPE) and goodness of fit (root mean square error, RMSE) of these forecasts. MPE can be thought of as measuring the bias of predictions at an aggregate level, where positive and negative errors can cancel each other out, while MAPE is a measure of the accuracy of individual predictions. RMSE is

[23] For example, one alternative specification featured a count of the number of morbidities instead a vector of morbidity markers.

[24] We also carried out Pregibon link, Ramsey RESET and modified Hosmer-Lemeshow tests in preliminary work although only results from the Pearson correlation coefficient tests are included, since they were found to display the same pattern more clearly (with the other tests there was smaller variation in rejection rates across the different models).

[25] This is computed using coefficients from models estimated on the 'estimation' set, e.g. for linear regression $E(\widehat{y_i^V|X_i^V}) = \hat{\alpha}^E + \hat{\beta}^E X_i^V$

similar to MAPE in that positive and negative errors do not cancel out, however larger errors count for disproportionately more, since they are squared. In addition, we evaluate the variability of bias across replications (absolute deviations of mean prediction error, ADMPE). These are all evaluated on the full 'validation' set. Formulae for calculating these metrics are provided below, where m denotes the model used, s the sample size used, and r the replication.

$$MPE_{msr} = \frac{\sum (y_i - \hat{y}_i)}{N_s} \tag{3.19}$$

$$MAPE_{msr} = \frac{\sum |y_i - \hat{y}_i|}{N_s} \tag{3.20}$$

$$RMSE_{msr} = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{N_s}} \tag{3.21}$$

$$ADMPE_{msr} = \left| MPE_{msr} - \frac{\sum_{r=1}^{R} MPE_{msr}}{R} \right| \tag{3.22}$$

Only replications where all 16 models are successfully estimated on the sample are included for evaluation, and model performance according to each criterion is calculated as an average over all included replications, e.g. $MPE_{ms} = \frac{\sum_{r=1}^{R} MPE_{msr}}{R}$.[26]

In order to get a greater insight into the performance of different distributions, we evaluate forecasted conditional means at different values of the covariates. In practice this is done by patitioning the fitted values of costs into deciles. We assess MPE and MAPE for deciles of predicted costs, since there is concern that models perform with varying success at different points in the distribution. Models designed for heavy-tails, for instance, might be expected to perform better in predicting the biggest costs. This also represents a desire to fit the distribution of costs for different groups of observations according to their observed covariates.

We combine the results that we obtain from different sample sizes (N_s), and attempt

[26] All models estimated successfully every time, except for CDEM and EEE. CDEM could not be estimated on two of the 100 replicates with samples of 5,000 observations. EEE could not be estimated on four, four, six and four of the 100 replicates with sample sizes of 5,000, 10,000, 50,000 and 100,000 observations, respectively.

to find a pattern in the way in which models perform as sample size varies. To do this we construct response surfaces (as in, for example Deb and Burgess (2003)). These are polynomial approximations to the relationship between the statistics of interest 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).

$$MPE_{msr} = \alpha_m^{MPE} + \beta_m^{MPE}\frac{1}{N_s} + u_{msr}^{MPE} \tag{3.23}$$

We specify the relationship between MPE and the inverse of the sample size, reflecting that we expect reduced bias as the number of observations increases. In particular, the value of α_m^{MPE} represents the value of MPE to which the model approaches asymptotically with increasing sample size: testing whether or not this is statistically significant from zero gives an indication of whether the estimator is consistent. Here, u_{msr}^{MPE} represents the error term from the regression. 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:

$$\ln\left(MAPE_{msr}\right) = \alpha_m^{LMAPE} + \beta_m^{LMAPE}\frac{1}{N_s} + u_{msr}^{LMAPE} \tag{3.24}$$

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

3.6 Results and Discussion

To begin with, we consider the results from the smallest samples that we draw from the 'estimation' set (5,000 observations). Results from larger samples are analysed by way of the response surfaces which we present later. Table 3.2 is a key for the labels we use for each model in discussion of the results.

3.6.1 Estimation Sample Results

We first conduct tests of misspecification across the models used. Researchers use these

OLS	linear regression
LOGOLSHET	transformed linear regression (log), heteroskedastic smearing factor
SQRTOLSHET	transformed linear regression ($\sqrt{\cdot}$), heteroskedastic smearing factor
GLMLOGP	generalised linear model, log link, Poisson-type family
GLMLOGG	generalised linear model, log link, gamma-type family
GLMSQRTP	generalised linear model, $\sqrt{}$-link, Poisson-type family
GLMSQRTG	generalised linear model, $\sqrt{}$-link, gamma-type family
LOGNORM	log-normal
GG	generalised gamma
GB2LOG	generalised beta of the second kind, log link
GB2SQRT	generalised beta of the second kind, $\sqrt{}$-link
FMMLOGG	two-component finite mixture of gamma densities, log link
FMMSQRTG	two-component finite mixture of gamma densities, $\sqrt{}$-link
EEE	extended estimating equations
CDEM	conditional density approximation estimator (multinomial logit)
CDEO	conditional density approximation estimator (ordered logit)

Table 3.2: Key for model labels

tests to inform the specification of regressors, and the appropriateness of distributional assumptions, in particular the link function. Since we use the same regressors in all models, our tests are used to inform choices of distributional assumptions. The Pearson correlation coefficient test is able to detect if there is a a linear association between the estimated residuals and estimated conditional means, where the null hypothesis is no association. A lack of this kind of association suggests evidence against misspecification. It is also possible, however, that the relationship between the error and covariates is non-linear which this test cannot detect. Linear regression estimated using OLS, by construction, generates residuals orthogonal to predicted costs, and so the Pearson test cannot be applied to this model.

Table 3.3 shows that, according to this test, there is less evidence of misspecification when the model is estimated using a square root link function compared to other possible link functions, when all other distributional assumptions are the same. This is also the case in the GLM family of models, where the link and distribution functions can be flexibly estimated using EEE, with results indicating that there is less evidence of misspecification with GLMSQRTP and GLMSQRTG than the flexible case (on average across replications with sample size 5,000, the estimated λ coefficient in EEE was 0.28 with standard deviation of 0.07, indicating a link function between logarithmic and square root). Whilst EEE should be better-specified on the scale of estimation (following, effectively, the transformation of the dependent variable), the re-transformation may lead to increased evidence of misspecification on the scale of interest (levels of costs). Introducing more flexibility

Model	Pearson
OLS	N/A
LOGOLSHET	99%
SQRTOLSHET	0%
GLMLOGP	11%
GLMLOGG	99%
GLMSQRTP	0%
GLMSQRTG	13%
LOGNORM	95%
GG	89%
GB2LOG	96%
GB2SQRT	85%
FMMLOGG	85%
FMMSQRTG	82%
EEE	48%
CDEM	7%
CDEO	1%

Table 3.3: % of tests rejected at 5% significance level, when all converged, 94 converged replications, sample size 5,000

in terms of the whole distribution, generally, appears to have mixed effects upon results from this test. In the case of LOGNORM and GLMLOGG which are special cases of GG, there is the least evidence of misspecification from the most complicated distribution amongst the three. There is also evidence of less misspecification with FMMLOGG compared to GLMLOGG, which it nests. Conversely, GG and LOGNORM are special cases of GB2LOG, for which there is the most evidence of misspecification among these three models. Looking at the rejection rates above for FMMSQRTG and GLMSQRTG, there is more evidence of misspecification in the more flexible case. Finally, the results from CDEM and CDEO are promising, with little evidence of misspecification compared to other models tested. This may be because there is no retransformation process onto the scale of interest for these models.

3.6.2 Validation Set Results

All tests in the above section are carried out on the estimation sample. Given the practical implementation of the models considered here, a researcher may be more interested in how models perform in forecasting costs out-of-sample. Results based on the estimation sample may arise from overfitting the data. Therefore, our main focus is the forecasting

performance out-of-sample, that is evaluation on the 'validation' set.

We look first at performance of model predictions on the whole 'validation' set. Then we consider how well the models forecast for different levels of covariates throughout the distribution, by analysing performance by decile of predicted costs. Finally, we analyse the out-of-sample performance with increasing sample size by constructing response surfaces.

	Bias	Accuracy	Goodness of fit
	MPE (£)	**MAPE** (£)	**RMSE**
OLS	-1.56	1833.49	4475.49
LOGOLSHET	-140.53	1816.63	4960.08
SQRTOLSHET	**0.11**	1725.95	**4432.94**
GLMLOGP	**-1.44**	1748.43	4557.19
GLMLOGG	-147.33	1818.06	4984.86
GLMSQRTP	**0.26**	1704.77	**4426.24**
GLMSQRTG	46.71	**1689.28**	**4454.25**
LOGNORM	64.25	1734.10	4825.51
GG	44.60	1750.79	4754.22
GB2LOG	-63.96	1796.91	4873.13
GB2SQRT	134.84	**1686.48**	4483.35
FMMLOGG	-3.19	1758.06	4782.69
FMMSQRTG	121.80	**1690.28**	4477.10
EEE	-42.31	1727.28	4508.03
CDEM	**0.89**	**1683.40**	**4444.85**
CDEO	-10.13	1725.53	4474.84

Table 3.4: Results of model performance, when all converged, sample size 5,000; averaged across 94 replications

Looking at the results in Table 3.4, where the four best performing models in each category (MPE, MAPE and RMSE) are emboldened, it is clear that some of the most commonly used models – OLS, LOGOLSHET, GLMLOGG, and LOGNORM – do not perform well on any metric. CDEM is among the models with top four performance in every category illustrating the potential advantages of this approach for analysts concerned with any of bias, accuracy or goodness of fit. Generally speaking, the results also indicate that a square root link function is the most appropriate of those featured.

In terms of bias, models which are mean-preserving in sample also perform well out-of-sample in these results. This is evidenced by the strong performance of OLS, GLMLOGP and GLMSQRTP, with absolute levels of mean prediction error of £1.56, £1.44 and £0.26 respectively. All models with a square root link function underpredict costs on average, whereas some log link function models underpredict (LOGNORM and GG) and others

overpredict on average (LOGOLSHET, GLMLOGP, GB2LOG and FMMLOGG). SQRTOLSHET and CDEM perform best and third best respectively, and worst performing is GLMLOGG, which overpredicts by £147.33 on average (5.64% of the population mean).

With respect to accuracy and goodness of fit, a clear message from the results is that the best performing link function is the square root. The ordering of the other link functions varies. For accuracy the flexible link function of EEE is next best, followed by log link function and then OLS. For goodness of fit OLS is second best, followed by EEE while the log link is the worst. There is variation in performance amongst different models with the same link function, which we discuss next when considering the gains to increased flexibility. In addition, CDEM performs very well according to these criteria.

First we consider the gains to using a mixture of gamma distributions, over the nested single gamma distribution models. Looking at the results for the GLMLOGG and FMMLOGG, the mixture improves forecasting performance in terms of bias, accuracy and goodness of fit. This is also observed in results from other sample sizes (see Appendix B). As discussed earlier, the gains to this increased flexibility are insufficient for results from FMMLOGG to perform better than relatively simple models using a square root link function (e.g. GLMSQRTP). Comparing results from GLMSQRTG with FMMSQRTG is more complicated, at sample size 5,000, as seen in Table 3.4, we observe GLMSQRTG to perform better than FMMSQRTG in all metrics. FMMSQRTG performs better than GLMSQRTG, at larger samples, in terms of accuracy – with FMMSQRTG the best performing model of all 16 compared – but the nested single distribution case (GB2SQRTG) performs better, at all sample sizes, in terms of bias and goodness of fit (see Appendix B).

Greater flexibility amongst the fully parametric models has an ambiguous effect on performance of forecasting means. GG is a limiting case of GB2 and its performance is better across all metrics. Conversely, LOGNORM, a special case of GG and GB2, performs best of the three in terms of accuracy, the worst in terms of bias, and second in terms of goodness-of-fit. Using GG or GB2LOG improves performance over special case GLMLOGG based on MPE, MAPE and RMSE. Once again, the best of these four models performs worse than certain models with a square root link function. Comparing GLMSQRTG and GB2SQRT, we can see that there is not a great deal gained from introducing more parameters, since performance is worse for GB2SQRT than GLMSQRTG except in

the cases of accuracy at sample sizes 5,000 and 10,000 (the difference is small at all sample sizes analysed).

Crucially, these results only consider performance based on the mean, while some of these models are capable of providing information on higher moments and on other features of the conditional distribution such as tail probabilities.[27] We construct graphs of bias and accuracy by decile of predicted costs. This can be thought of as analysing the fit of models for the mean of distributions of costs conditional on observed variables, since each decile of predicted costs represents a group of observations with certain values of covariates. In previous analysis, we have considered all observations as equal, but it is possible that a policymaker prioritises the prediction error of certain observations over others. There is considerable interest in modelling the outcomes for high-cost patients, since these can be responsible for large proportions of overall costs. The highest costs are likely to be found in the highest decile of predicted costs.

[27]This is a significant qualitative advantage of parametric models over models such as linear regression, where the models have been used to predict probabilities of lying beyond a threshold value, e.g. tail probabilities, see Jones et al. (2014) who find that the GG and LOGNORM distribution perform best for the threshold values they choose.

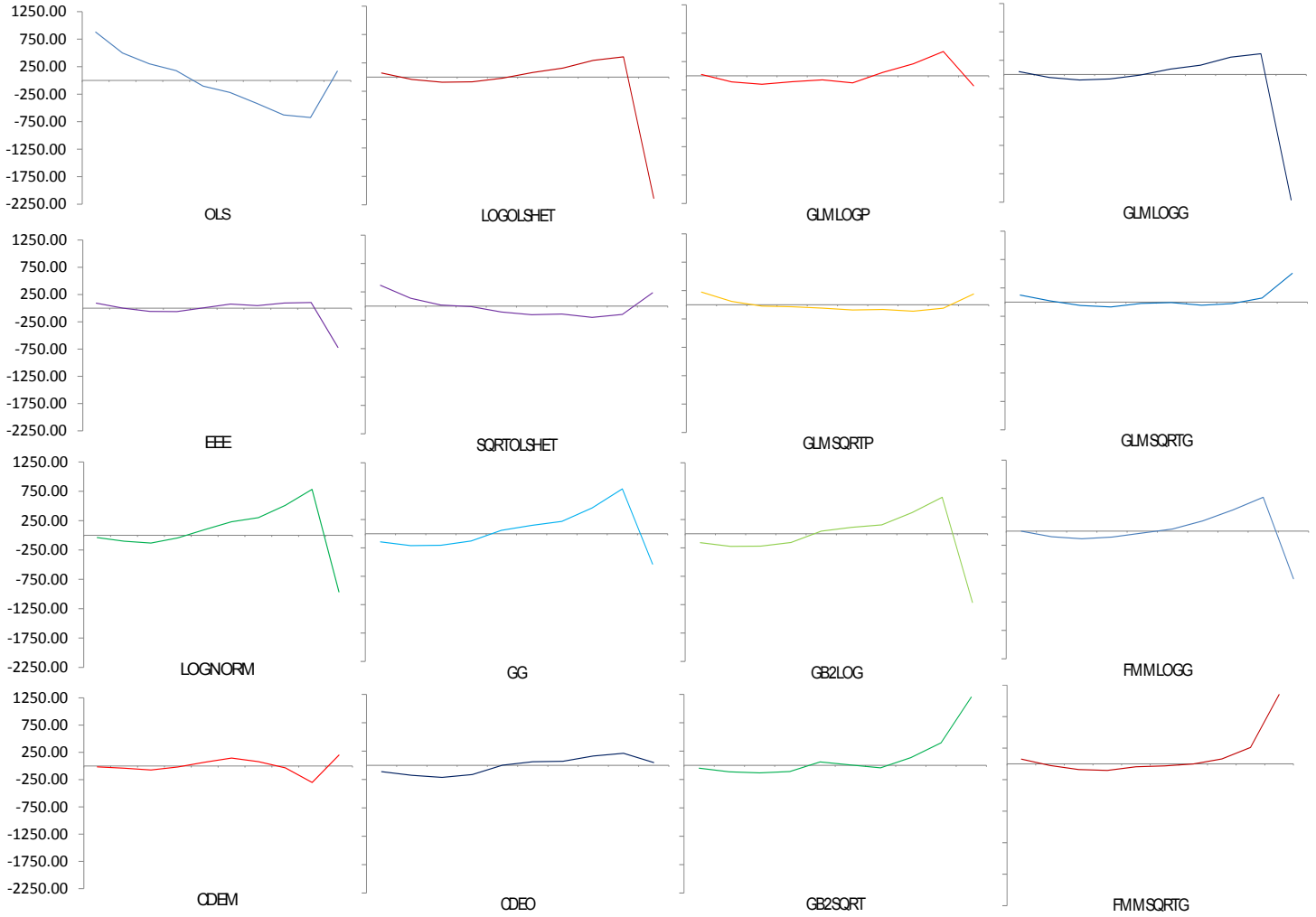


Figure 3.6: MPE by decile of fitted costs

Models with the same link function follow a largely similar pattern. Those, for example, with square root link functions underpredict in the decile of highest predicted costs, whereas log link models overpredict in the last decile. Results with other link functions – OLS, EEE, CDEM and CDEO – all have different patterns. Generally speaking, the first decile of predicted costs from square root models are on average underpredictions (only GB2SQRT overpredicts in the smallest decile), which combined with the underpredicted last decile gives them a 'u-shaped' line. The performance of each model varies across the deciles. SQRTOLSHET has a 'u-shaped' line, and while it performs best in predicting costs on average across all deciles, the performance for certain groups may be worse than other models. For example, CDEM performs slightly worse across all ten deciles, but has a smaller range of over- and underpredictions. In terms of the highest decile of predicted costs, the model with the lowest MPE is CDEO, underestimating on average £48.96. Generally this decile tends to be the largest absolute MPE for models, with values as large as an average overprediction of £2211.47 in the case of GLMLOGG.

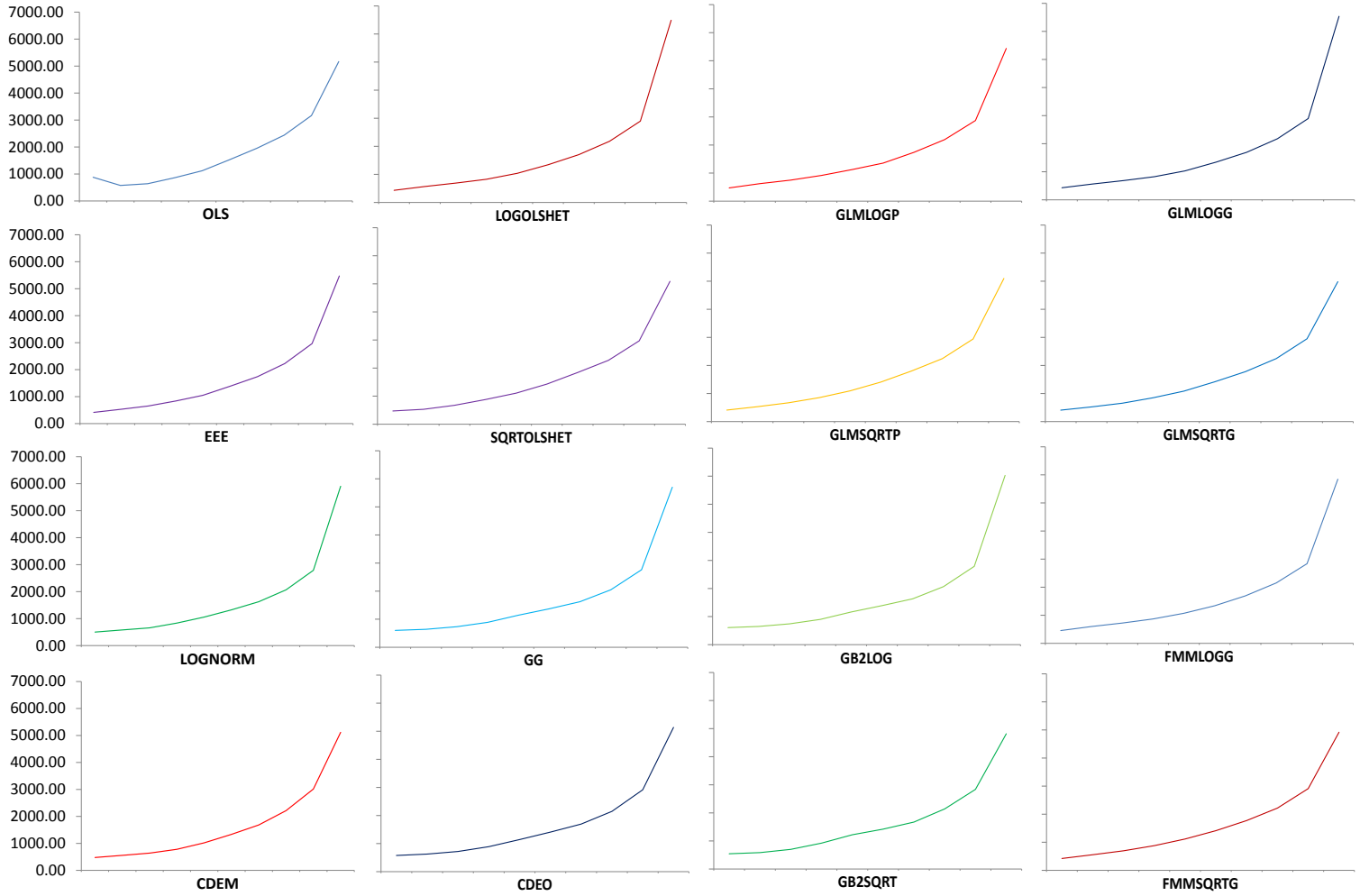


Figure 3.7: MAPE by decile of fitted costs

When looking at MAPE by decile of predicted cost, it is striking that the pattern across models is very similar. In all models, except OLS, the MAPE is higher in deciles with larger predicted costs. The most inaccurate models in the highest decile are those with a log link function, followed by EEE, then OLS, the conditional density approximation estimators with the most accurate being models with a square root link function. Generally, it appears that models that predict larger costs overall are the least accurate in the highest decile, implying that models which estimate the largest range of predicted conditional means will not necessarily perform best in forecasting mean costs for patients most likely to be high cost patients (those with lots of observed morbidity). GLMLOGG overpredicts on average over the whole validation subset by £147.33, has the largest overprediction in the highest decile, and is also the least accurate in this decile with MAPE of £6536.24, over twice the population mean cost.

Figure 3.8 displays the response surfaces constructed to analyse how each model's performance varied with increasing sample size for the subset of best performing models (those emboldened in Table 3.4). We have already touched upon this earlier when looking at results regarding accuracy between related distributions. The performance of most estimated models varies little as sample sizes increase above 5,000. There is some evidence of the variability of MPE (measured using ADMPE) reducing as sample size increases, although this happens at a similar rate across all models. Largely, though, the response surfaces for each model are parallel indicating that relative performance of models changes little. Further, the fact that they are flat represents evidence of performance not changing for each model with increasing sample size. The exception to this is that the performance of FMMSQRTG varies with increasing sample size: its accuracy improves, and its bias worsens. This suggests that this model behaves differently with samples as small as 5,000 observations, possibly because of the number of parameters that are required. On the whole, though, from samples of 5,000 observations or more, there is little evidence that more flexible models require more observations than less flexible ones.

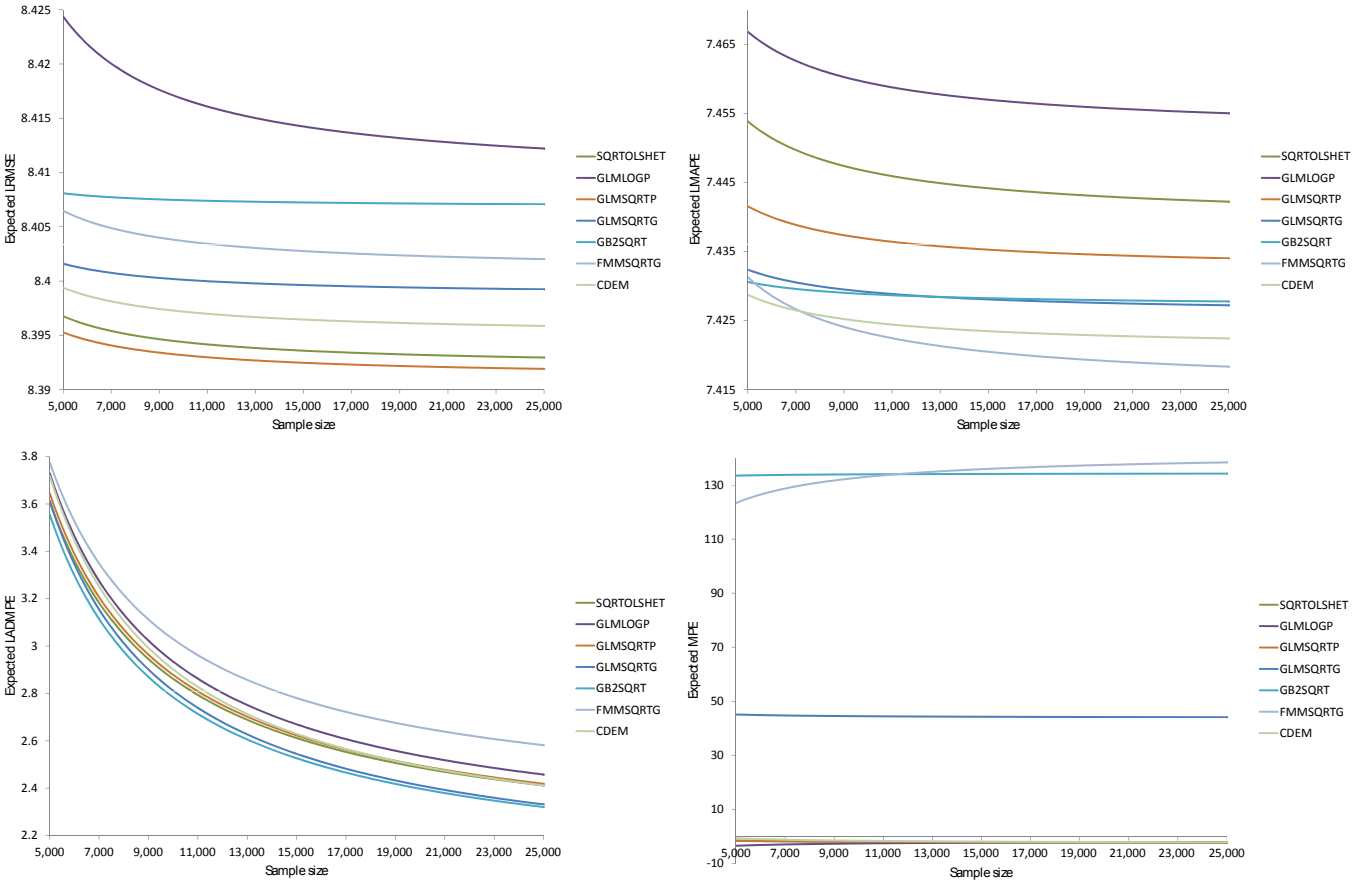


Figure 3.8: Reponse surfaces for log(RMSE), log(MAPE), MPE, log(ADMPE) (clockwise from top left) against sample size, constructed evaluating performance on 'validation' set

3.7 Conclusions

We have systematically evaluated the state of the art in regression models for healthcare costs, using administrative English hospital inpatient data, employing a quasi-Monte Carlo design to ensure rigour and drawing conclusions based on out-of-sample forecasting. We have compared recently-adopted semi- and fully parametric regression methods that have never before been evaluated against one another, as well as comparing with regression methods that are now considered standard practice in modelling healthcare cost data.

Our results echo other studies, in that there is no single model that dominates in all respects: SQRTOLSHET is the best performing model in terms of bias, CDEM for accuracy, and in terms of goodness of fit the best performer is GLMSQRTP. Therefore the policymaker has to weigh up these factors in arriving at their preferred model, based upon their loss function over prediction errors. It is worth noting, however, that CDEM performs amongst the best four models for all three metrics. Another striking result is that four models commonly employed in regression methods for healthcare costs do not perform amongst the best four of any of the three metrics (OLS, LOGOLSHET, GLMLOGG and LOGNORM). Our analysis by decile shows the way in which models are sensitive to the choice of link function, with square root link functions underpredicting in the decile of highest predicted costs, and log link models overpredicting in the last decile. Finally, the response surfaces indicate that, on the whole, the more recent developments do not suffer because of the use of smaller sample sizes (from 5,000 observations).

Acknowledgements

The authors gratefully acknowledge funding from the Economic and Social Research Council (ESRC) under grant reference RES-060-25-0045. We would like to thank members of the Health, Econometrics and Data Group (HEDG), at the University of York, for useful discussions. This paper has benefited from feedback from numerous presentations at the University of Leeds (AUHE Symposium), Brunel University, the University of Oxford, the University of Padua (Italian Health Econometrics Workshop) and the University of Copenhagen. In particular, we would like to thank Alberto Holly, Boby Mihaylova and Sandy Tubeuf for fruitful discussions of the work. We are also grateful to John Mullahy, Will Manning, Partha Deb and Anirban Basu for their ongoing support and generosity in providing feedback on our work.

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.

Basu A, Rathouz PJ. 2005. Estimating marginal and incremental effects on health outcomes using flexible link and variance function models. *Biostatistics* **6**: 93–109.

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

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.

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.

Cawley J, Meyerhoefer C. 2012. The medical care costs of obesity: An instrumental variables approach. *Journal of Health Economics* **31**: 219 – 230.

Copas JB. 1983. Regression, prediction and shrinkage. *Journal of the Royal Statistical Society. Series B (Methodological)* **45**: pp. 311–354.

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**.

Deb P, Trivedi PK. 1997. Demand for medical care by the elderly: A finite mixture approach. *Journal of Applied Econometrics* **12**: 313–336.

Dixon J, Smith P, Gravelle H, Martin S, Bardsley M, Rice N, Georghiou T, Dusheiko M, Billings J, Lorenzo MD, Sanderson C. 2011. A person based formula for allocating commissioning funds to general practices in england: development of a statistical model. *BMJ* **343**: d6608.

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

Duan N, Manning WG, Morris CN, Newhouse JP. 1983. A comparison of alternative models for the demand for medical care. *Journal of Business & Economic Statistics* **1**: 115–126.

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

Han A, Hausman JA. 1990. Flexible parametric estimation of duration and competing risk models. *Journal of Applied Econometrics* **5**: 1–28.

Heckman JJ. 2001. Micro data, heterogeneity, and the evaluation of public policy: Nobel lecture. *Journal of Political Economy* **109**: 673–748.

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.

Hoch JS, Briggs AH, Willan AR. 2002. Something old, something new, something borrowed, something blue: a framework for the marriage of health econometrics and cost-effectiveness analysis. *Health Economics* **11**: 415–430.

Holly A. 2009. Modeling risk using fourth order pseudo maximum likelihood methods. Institute of Health Economics and Management (IEMS), University of Lausanne, Switzerland.

Holly A, Monfort A, Rockinger M. 2011. Fourth order pseudo maximum likelihood methods. *Journal of Econometrics* **162**: 278–293.

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.

Huber M, Lechner M, Wunsch C. 2013. The performance of estimators based on the propensity score. *Journal of Econometrics* **175**: 1–21.

Johnson E, Dominici F, Griswold M, L Zeger S. 2003. Disease cases and their medical costs attributable to smoking: an analysis of the national medical expenditure survey. *Journal of Econometrics* **112**: 135–151.

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

Jones AM, Lomas J, Rice N. 2014. Applying beta-type size distributions to healthcare cost regressions. *Journal of Applied Econometrics* **29**: 649–670.

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

Manning WG, Duan N, Rogers W. 1987. Monte Carlo evidence on the choice between sample selection and two-part models. *Journal of Econometrics* **35**: 59 – 82.

McDonald JB, Sorensen J, Turley PA. 2013. Skewness and kurtosis properties of income distribution models. *Review of Income and Wealth* **59**: 360–374.

Mihaylova B, Briggs A, O'Hagan A, Thompson SG. 2011. Review of statistical methods for analysing healthcare resources and costs. *Health Economics* **20**: 897–916.

Mullahy J. 1997. Heterogeneity, excess zeros, and the structure of count data models. *Journal of Applied Econometrics* **12**: 337–350.

Mullahy J. 2009. Econometric modeling of health care costs and expenditures: a survey of analytical issues and related policy considerations. *Medical Care* **47**: S104–S108.

Pentsak Y. 2007. Addressing skewness and kurtosis in health care econometrics. PhD Thesis, University of Lausanne.

Van de Ven WP, Ellis RP. 2000. Risk adjustment in competitive health plan markets. In Culyer AJ, Newhouse JP (eds.) *Handbook of Health Economics*, volume 1 of *Handbook of Health Economics*, chapter 14. Elsevier, 755–845.

Vanness DJ, Mullahy J. 2007. Perspectives on mean-based evaluation of health care. In Jones AM (ed.) *The Elgar Companion to Health Economics*. Elgar Original Reference.

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

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

3.8 Appendix A

We use the variables shown in Table 3A1 to construct our regression models. They are based on the ICD10 chapters, which are given in Table 3A2.

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 3A1: Classification of morbidity characteristics

ICD10 codes beginning with U were dropped because there were no observations in the 6,164,114 used. Only a small number (3,170) 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 3A2: ICD10 chapter codes

3.9 Appendix B

	Sample size		
Model	**10,000**	**50,000**	**100,000**
OLS	-	-	-
LOGOLSHET	100%	100%	100%
SQRTOLSHET	0%	44%	100%
GLMLOGP	46%	100%	100%
GLMLOGG	100%	100%	100%
GLMSQRTP	0%	13%	79%
GLMSQRTG	42%	100%	100%
LOGNORM	99%	100	100%
GG	99%	100%	100%
GB2LOG	100%	100%	100%
GB2SQRT	100%	100%	100%
FMMLOGG	98%	100%	100%
FMMSQRTG	97%	100%	100%
EEE	69%	100%	100%
CDEM	48%	100%	100%
CDEO	6%	89%	99%

Table 3B1: Results Pearson correlation coefficient tests (percentage rejected at 5% significance level)

	MPE (£)	**MAPE** (£)	**RMSE**
OLS	-6.19	1815.31	4460.42
LOGOLSHET	-149.73	1811.87	4938.27
SQRTOLSHET	-5.63	1715.94	4421.70
GLMLOGP	-6.77	1738.14	4522.67
GLMLOGG	-155.09	1813.78	4960.21
GLMSQRTP	-5.51	1699.16	4416.51
GLMSQRTG	41.14	1685.78	4447.57
LOGNORM	59.19	1732.46	4817.63
GG	43.26	1747.08	4743.14
GB2LOG	-68.96	1794.72	4865.08
GB2SQRT	131.03	1684.73	4480.02
FMMLOGG	-1.94	1747.21	4728.73
FMMSQRTG	136.06	1672.63	4461.85
EEE	-39.21	1716.71	4483.11
CDEM	-4.80	1677.85	4433.55
CDEO	-15.66	1724.60	4471.24

Table 3B2: Results of model performance, when all converged, sample size 10,000

	MPE (£)	**MAPE** (£)	**RMSE**
OLS	-1.81	1796.18	4449.72
LOGOLSHET	-147.59	1802.19	4906.22
SQRTOLSHET	-1.67	1703.75	4413.87
GLMLOGP	-1.27	1725.75	4495.21
GLMLOGG	-151.88	1804.52	4924.32
GLMSQRTP	-1.64	1690.68	4409.46
GLMSQRTG	44.49	1679.73	4442.24
LOGNORM	64.37	1726.74	4795.49
GG	52.19	1739.61	4721.18
GB2LOG	-59.70	1786.85	4838.97
GB2SQRT	135.16	1681.21	4478.19
FMMLOGG	-1.78	1739.02	4707.86
FMMSQRTG	139.06	1663.82	4453.58
EEE	-27.50	1703.97	4461.69
CDEM	-1.77	1671.65	4427.06
CDEO	-12.33	1721.03	4468.20

Table 3B3: Results of model performance, when all converged, sample size 50,000

	MPE (£)	**MAPE** (£)	**RMSE**
OLS	-1.38	1793.94	4448.34
LOGOLSHET	-145.55	1799.78	4894.66
SQRTOLSHET	-1.13	1702.17	4412.80
GLMLOGP	-0.23	1723.93	4491.04
GLMLOGG	-149.34	1801.91	4911.21
GLMSQRTP	-1.08	1689.60	4408.53
GLMSQRTG	45.22	1679.03	4441.85
LOGNORM	64.92	1725.98	4792.42
GG	52.96	1738.69	4718.07
GB2LOG	-59.08	1785.98	4835.64
GB2SQRT	135.08	1680.93	4477.82
FMMLOGG	0.10	1737.17	4702.38
FMMSQRTG	140.53	1662.90	4453.24
EEE	-23.27	1700.90	4456.40
CDEM	-1.26	1670.77	4426.08
CDEO	-11.73	1720.51	4467.82

Table 3B4: Results of model performance, when all converged, sample size 100,000

	Decile of predicted cost									
Model	1	2	3	4	5	6	7	8	9	10
OLS	878.41	498.24	301.60	177.47	-103.47	-218.87	-417.56	-627.31	-674.45	169.25
LOGOLSHET	72.19	-41.42	-90.06	-82.74	-18.47	81.12	158.80	294.32	356.78	-2135.90
SQRTOLSHET	366.18	137.92	18.07	-9.74	-104.61	-152.77	-140.83	-199.66	-147.44	233.38
GLMLOGP	23.56	-107.43	-149.27	-105.00	-71.83	-124.65	60.05	209.83	427.23	-176.98
GLMLOGG	50.81	-52.36	-96.95	-79.00	-12.05	94.98	162.95	304.74	364.94	-2211.47
GLMSQRTP	223.58	58.94	-24.00	-35.27	-59.78	-92.82	-82.67	-114.25	-60.64	189.14
GLMSQRTG	124.08	20.66	-59.16	-84.72	-22.72	-9.06	-54.07	-27.41	71.68	507.66
LOGNORM	-38.43	-99.80	-131.68	-43.98	97.28	230.40	300.73	510.15	782.57	-964.60
GG	-143.01	-210.11	-203.73	-128.22	59.81	146.80	217.61	454.21	789.59	-536.75
GB2LOG	-157.20	-222.93	-219.12	-154.65	45.15	111.74	155.70	371.60	642.29	-1212.01
GB2SQRT	-50.69	-114.56	-132.83	-109.30	59.73	6.65	-44.19	134.96	396.12	1202.57
FMMLOGG	-1.93	-98.99	-135.11	-105.46	-34.17	36.82	179.35	373.03	596.01	-841.44
FMMSQRTG	75.97	-25.76	-89.96	-103.93	-45.77	-31.70	-1.11	78.45	260.96	1100.74
EEE	91.79	0.36	-59.09	-62.56	6.34	75.45	46.98	94.32	103.54	-720.37
CDEM	-16.33	-40.35	-73.86	-18.31	69.14	144.93	79.88	-34.04	-302.29	200.10
CDEO	-111.98	-176.27	-213.31	-162.65	3.14	63.10	71.40	163.83	212.48	48.96

Table 3B5: Models' average mean prediction error (£) by decile of predicted cost at sample size 5,000

	Decile of predicted cost									
Model	1	2	3	4	5	6	7	8	9	10
OLS	821.88	469.46	293.66	173.65	-134.52	-228.16	-434.91	-634.77	-668.78	279.44
LOGOLSHET	59.60	-52.01	-106.49	-95.15	-30.01	67.37	156.75	280.35	346.95	-2124.82
SQRTOLSHET	333.52	114.76	-5.18	-16.99	-118.91	-169.77	-144.20	-200.53	-145.95	296.31
GLMLOGP	-7.46	-138.57	-177.80	-109.04	-69.16	-162.94	55.85	200.82	417.45	-76.94
GLMLOGG	40.74	-61.02	-109.44	-89.01	-26.68	83.26	160.16	288.70	354.38	-2192.10
GLMSQRTP	213.91	46.63	-39.07	-36.10	-76.04	-109.29	-90.19	-119.35	-71.38	225.44
GLMSQRTG	116.40	12.89	-73.70	-100.94	-47.57	12.06	-87.90	-33.00	71.34	541.61
LOGNORM	-44.86	-103.90	-143.50	-43.17	85.94	235.45	284.50	508.01	782.45	-968.86
GG	-144.37	-226.13	-201.47	-144.25	68.16	134.43	215.00	461.44	792.23	-522.18
GB2LOG	-159.65	-237.28	-216.89	-171.23	52.58	96.44	153.83	375.92	640.54	-1223.68
GB2SQRT	-55.62	-123.94	-128.14	-126.20	71.97	-5.73	-57.67	130.66	400.15	1204.90
FMMLOGG	-15.44	-115.07	-151.67	-104.54	-49.12	10.19	187.04	366.43	605.81	-753.02
FMMSQRTG	66.90	-35.99	-105.13	-132.67	-68.87	-24.54	-10.48	84.48	326.12	1260.66
EEE	85.26	-9.14	-74.40	-82.26	-25.88	88.86	17.44	97.27	116.52	-605.90
CDEM	-34.26	-43.05	-87.72	-22.54	58.40	143.35	68.13	-49.86	-328.54	248.16
CDEO	-119.90	-182.20	-225.87	-162.25	-17.17	60.67	64.90	165.18	215.17	44.94

Table 3B6: Models' average mean prediction error (£) by decile of predicted cost at sample size 10,000

	Decile of predicted cost									
Model	1	2	3	4	5	6	7	8	9	10
OLS	768.06	444.90	261.71	211.65	-162.65	-224.78	-436.43	-629.07	-636.96	384.46
LOGOLSHET	40.66	-63.18	-120.65	-98.74	-27.11	71.95	154.10	278.92	354.17	-2066.07
SQRTOLSHET	296.11	104.74	-34.85	5.55	-141.65	-177.56	-129.95	-175.76	-131.53	367.73
GLMLOGP	-37.97	-170.27	-203.77	-117.15	-50.17	-178.15	68.46	183.92	429.62	62.80
GLMLOGG	26.92	-70.31	-119.89	-90.86	-25.44	83.54	158.30	285.66	358.34	-2125.11
GLMSQRTP	198.14	37.23	-51.29	-24.78	-93.41	-109.47	-92.28	-101.31	-66.36	286.77
GLMSQRTG	104.24	7.93	-83.75	-100.93	-64.08	42.70	-117.01	-23.52	82.48	596.56
LOGNORM	-54.04	-111.30	-144.41	-41.15	81.71	236.05	269.16	515.50	789.18	-896.86
GG	-141.90	-238.99	-202.57	-154.60	77.50	113.44	230.59	469.88	804.80	-435.95
GB2LOG	-160.65	-245.80	-215.95	-181.97	61.39	72.92	174.47	382.37	657.11	-1140.59
GB2SQRT	-60.75	-130.80	-119.73	-129.98	82.93	-11.39	-68.85	138.36	413.63	1238.06
FMMLOGG	-34.22	-125.95	-161.88	-93.53	-49.81	-1.79	213.11	362.77	611.62	-738.07
FMMSQRTG	56.60	-38.19	-119.91	-132.29	-76.54	-11.73	-25.19	87.49	338.81	1311.47
EEE	76.23	-17.18	-84.60	-87.23	-43.18	103.90	-12.48	106.24	136.25	-453.13
CDEM	-47.07	-43.60	-101.08	-24.31	48.17	145.44	57.68	-56.54	-339.01	342.65
CDEO	-130.29	-190.45	-234.35	-156.66	-29.08	45.46	73.90	169.07	233.67	95.50

Table 3B7: Models' average mean prediction error (£) by decile of predicted cost at sample size 50,000

	Decile of predicted cost									
Model	1	2	3	4	5	6	7	8	9	10
OLS	761.33	441.25	263.68	211.50	-166.68	-225.62	-434.71	-628.76	-633.12	396.37
LOGOLSHET	38.30	-62.53	-125.03	-100.31	-27.79	72.07	150.17	280.03	355.76	-2036.25
SQRTOLSHET	290.94	106.72	-45.14	6.63	-146.40	-177.43	-131.87	-166.39	-127.88	378.95
GLMLOGP	-41.05	-173.23	-207.74	-119.94	-47.60	-178.91	70.02	178.29	432.38	85.42
GLMLOGG	25.07	-69.95	-123.63	-93.28	-24.34	81.28	155.48	285.85	360.55	-2090.44
GLMSQRTP	196.89	35.30	-54.10	-23.58	-98.07	-116.64	-90.84	-95.65	-64.14	299.60
GLMSQRTG	102.64	7.06	-85.09	-104.71	-66.86	47.18	-125.79	-18.43	85.42	610.59
LOGNORM	-54.69	-112.28	-143.09	-42.46	81.70	237.09	264.89	516.00	789.82	-887.70
GG	-140.56	-241.55	-200.69	-157.87	76.96	109.26	233.94	468.15	806.01	-423.72
GB2LOG	-159.90	-247.54	-214.15	-185.09	59.92	65.76	180.76	380.80	659.11	-1130.20
GB2SQRT	-61.84	-131.53	-119.32	-130.27	81.62	-11.40	-72.45	139.53	414.41	1241.97
FMMLOGG	-35.47	-126.45	-164.61	-88.71	-57.36	-9.50	218.23	360.09	615.49	-710.66
FMMSQRTG	54.01	-39.31	-124.91	-136.39	-76.82	-10.32	-31.55	92.74	344.52	1333.18
EEE	76.77	-19.28	-85.65	-91.64	-45.13	95.20	-15.34	106.00	141.78	-395.60
CDEM	-48.70	-44.60	-101.13	-24.65	46.88	145.68	55.88	-58.64	-339.60	356.41
CDEO	-131.91	-189.53	-236.78	-153.91	-33.74	44.92	73.75	171.04	236.17	102.71

Table 3B8: Models' average mean prediction error (£) by decile of predicted cost at sample size 100,000

	Decile of predicted cost									
Model	1	2	3	4	5	6	7	8	9	10
OLS	879.77	575.57	639.97	864.80	1120.42	1530.58	1953.91	2437.88	3165.22	5167.22
LOGOLSHET	432.16	567.71	690.35	825.97	1038.71	1341.88	1702.04	2181.12	2896.97	6490.37
SQRTOLSHET	472.78	530.79	673.13	878.29	1104.90	1427.97	1839.97	2275.21	2969.18	5088.17
GLMLOGP	472.76	624.62	749.98	916.79	1127.15	1354.09	1737.56	2192.54	2871.79	5438.02
GLMLOGG	431.11	564.87	686.22	824.57	1032.54	1341.37	1692.92	2175.93	2895.86	6536.24
GLMSQRTP	410.47	531.24	669.20	853.52	1093.50	1405.85	1801.24	2242.47	2942.28	5098.81
GLMSQRTG	407.94	522.60	655.95	849.47	1084.61	1415.75	1778.61	2244.43	2950.61	4983.76
LOGNORM	500.51	580.15	657.62	833.68	1056.19	1325.01	1625.88	2065.70	2789.91	5907.09
GG	601.19	640.44	733.21	888.59	1143.62	1368.59	1622.82	2046.71	2761.63	5701.97
GB2LOG	607.97	648.02	740.91	899.20	1163.18	1389.65	1633.55	2067.98	2790.67	6028.71
GB2SQRT	544.30	584.86	701.04	922.90	1220.87	1420.87	1672.45	2142.12	2837.97	4818.24
FMMLOGG	455.63	598.45	722.83	869.04	1070.73	1338.32	1691.78	2150.84	2837.05	5846.93
FMMSQRTG	422.12	556.24	696.00	879.97	1116.32	1413.42	1772.12	2217.32	2910.98	4919.30
EEE	415.25	531.62	651.85	836.66	1049.22	1386.31	1739.97	2225.07	2966.02	5471.84
CDEM	476.96	555.78	639.96	780.26	1019.69	1335.12	1683.51	2213.41	3013.70	5116.58
CDEO	576.63	626.78	717.86	889.49	1142.60	1409.13	1694.02	2146.64	2915.21	5137.76

Table 3B9: Models' average mean absolute prediction error (£) by decile of predicted cost at sample size 5,000

	Decile of predicted cost									
Model	1	2	3	4	5	6	7	8	9	10
OLS	823.15	555.90	631.90	860.48	1101.13	1540.49	1950.52	2427.09	3139.43	5123.35
LOGOLSHET	431.30	567.50	688.57	816.60	1031.93	1331.88	1709.15	2170.82	2891.03	6480.89
SQRTOLSHET	451.16	528.18	664.82	878.72	1089.47	1422.35	1840.52	2266.38	2951.12	5067.67
GLMLOGP	477.64	625.26	746.17	906.93	1131.35	1337.55	1734.70	2182.37	2859.00	5381.39
GLMLOGG	431.00	565.77	684.20	819.89	1023.22	1333.50	1701.53	2167.19	2890.53	6521.88
GLMSQRTP	405.98	531.14	661.24	856.03	1079.96	1400.72	1803.69	2236.63	2930.65	5086.55
GLMSQRTG	406.69	523.08	654.88	846.33	1064.11	1435.75	1765.19	2241.50	2942.36	4978.86
LOGNORM	502.29	582.93	648.20	841.26	1049.65	1324.72	1618.63	2064.72	2789.15	5903.89
GG	607.39	633.99	727.07	887.22	1142.93	1353.45	1624.78	2046.48	2756.94	5691.36
GB2LOG	614.49	642.18	734.34	900.43	1159.46	1376.05	1637.34	2068.12	2787.27	6028.33
GB2SQRT	549.51	580.31	696.52	918.68	1229.69	1411.57	1668.49	2140.93	2835.99	4816.42
FMMLOGG	460.53	598.19	720.21	864.41	1059.81	1314.83	1694.76	2139.13	2823.88	5797.27
FMMSQRTG	424.65	554.36	697.88	868.73	1084.75	1416.72	1760.39	2185.57	2866.85	4867.40
EEE	413.78	534.75	653.24	833.36	1025.46	1399.82	1727.13	2217.48	2948.43	5414.48
CDEM	478.12	554.96	638.92	778.62	1017.52	1334.62	1678.98	2206.38	3002.68	5088.50
CDEO	581.02	628.72	709.23	898.91	1127.09	1411.07	1695.33	2146.62	2915.49	5133.30

Table 3B10: Models' average mean absolute prediction error (£) by decile of predicted cost at sample size 10,000

	Decile of predicted cost									
Model	1	2	3	4	5	6	7	8	9	10
OLS	768.85	541.26	610.06	881.75	1069.10	1551.55	1933.73	2409.88	3113.74	5082.23
LOGOLSHET	429.58	561.83	688.53	807.53	1024.78	1329.72	1704.85	2164.61	2880.53	6430.87
SQRTOLSHET	422.32	539.92	640.41	893.53	1058.08	1413.50	1840.22	2257.59	2931.13	5041.73
GLMLOGP	480.08	618.62	756.96	878.99	1138.11	1328.83	1746.09	2149.72	2853.08	5307.87
GLMLOGG	431.81	560.11	684.30	814.80	1015.15	1331.36	1698.38	2162.99	2879.88	6467.43
GLMSQRTP	396.93	534.22	650.66	866.29	1055.36	1395.82	1798.02	2233.11	2914.62	5062.77
GLMSQRTG	405.92	521.15	647.28	852.27	1039.32	1459.18	1744.01	2236.95	2932.55	4959.75
LOGNORM	500.44	584.59	645.16	844.44	1053.92	1318.84	1609.03	2065.50	2783.57	5862.68
GG	611.60	629.46	722.95	877.98	1144.29	1330.65	1638.31	2042.62	2750.67	5648.39
GB2LOG	619.63	635.77	732.23	894.72	1155.89	1352.95	1654.53	2062.14	2782.81	5978.79
GB2SQRT	553.94	575.79	690.26	919.19	1235.06	1399.53	1658.93	2139.57	2830.54	4810.14
FMMLOGG	461.64	596.10	713.92	864.25	1043.18	1293.90	1704.38	2120.48	2810.22	5783.08
FMMSQRTG	426.90	554.34	687.22	870.68	1059.59	1428.87	1739.51	2170.59	2851.74	4849.70
EEE	412.88	535.05	649.33	837.32	1004.28	1413.33	1708.59	2212.19	2930.73	5337.10
CDEM	474.76	556.18	640.39	776.69	1017.12	1337.97	1671.41	2201.47	2990.46	5050.98
CDEO	584.29	627.26	708.12	905.02	1116.05	1402.84	1704.46	2140.35	2913.49	5109.27

Table 3B11: Models' average mean absolute prediction error (£) by decile of predicted cost at sample size 50,000

	Decile of predicted cost									
Model	1	2	3	4	5	6	7	8	9	10
OLS	761.93	538.22	613.63	882.50	1061.12	1553.78	1932.49	2407.91	3111.01	5077.11
LOGOLSHET	430.25	562.48	687.36	805.11	1025.77	1329.85	1703.25	2165.65	2877.75	6411.24
SQRTOLSHET	418.79	544.34	632.39	898.00	1053.61	1411.66	1839.24	2258.54	2928.42	5037.66
GLMLOGP	481.31	616.64	760.43	872.17	1138.80	1329.53	1749.36	2141.91	2853.14	5296.89
GLMLOGG	432.92	560.49	683.34	812.21	1017.80	1329.84	1697.54	2163.70	2877.59	6444.67
GLMSQRTP	396.72	535.84	647.29	869.59	1051.08	1391.81	1798.97	2234.50	2912.49	5058.61
GLMSQRTG	407.62	518.79	648.16	851.77	1037.56	1462.26	1739.84	2237.83	2931.44	4956.07
LOGNORM	499.05	585.58	645.14	842.99	1054.22	1319.81	1607.29	2065.35	2782.88	5858.31
GG	613.13	629.43	722.41	875.78	1142.98	1327.14	1643.17	2040.20	2749.96	5643.64
GB2LOG	621.73	634.84	732.10	893.66	1152.43	1347.92	1661.59	2060.34	2782.50	5973.69
GB2SQRT	555.68	575.06	688.65	919.95	1234.31	1399.84	1656.56	2140.01	2829.95	4810.15
FMMLOGG	462.06	595.23	714.82	867.91	1040.55	1288.38	1707.44	2118.45	2808.24	5769.40
FMMSQRTG	426.74	555.77	686.13	872.24	1057.40	1432.20	1734.65	2168.91	2850.22	4845.77
EEE	413.77	533.85	651.76	836.46	1006.16	1408.20	1710.56	2210.26	2926.58	5312.42
CDEM	474.63	554.41	641.63	776.70	1016.61	1339.16	1670.01	2201.30	2988.91	5045.26
CDEO	584.83	627.99	705.53	907.38	1110.80	1404.08	1705.17	2140.17	2913.84	5106.01

Table 3B12: Models' average mean absolute prediction error (£) by decile of predicted cost at sample size 100,000

	Response surface regressions			
Regression coefficient	**MPE**	**LMAPE**	**LRMSE**	**LADMPE**
α	-2.46 (1.87)	7.49 (0.00)	8.40 (0.00)	2.11 (0.09)
β	-3524.10 (32871.71)	113.56 (15.16)	31.89 (1.98)	7432.02 (773.24)

Table 3B13: Regression coefficients for OLS response surface regressions

	Response surface regressions			
Regression coefficient	**MPE**	**LMAPE**	**LRMSE**	**LADMPE**
α	-148.13 (2.31)	7.50 (0.00)	8.50 (0.00)	2.29 (0.09)
β	27441.85 (40826.75)	47.30 (11.91)	64.37 (18.59)	8147.63 (717.13)

Table 3B14: Regression coefficients for LOGOLSHET response surface regressions

	Response surface regressions			
Regression coefficient	**MPE**	**LMAPE**	**LRMSE**	**LADMPE**
α	-2.44 (1.88)	7.44 (0.00)	8.39 (0.00)	2.11 (0.09)
β	4161.85 (33067.30)	72.66 (10.43)	23.83 (1.29)	7496.99 (828.20)

Table 3B15: Regression coefficients for SQRTOLSHET response surface regressions

	Response surface regressions			
Regression coefficient	**MPE**	**LMAPE**	**LRMSE**	**LADMPE**
α	-1.75 (1.95)	7.45 (0.00)	8.41 (0.00)	2.14 (0.09)
β	-8381.67 (35203.29)	73.77 (9.66)	76.03 (7.89)	7947.14 (721.92)

Table 3B16: Regression coefficients for GLMLOGP response surface regressions

	Response surface regressions			
Regression coefficient	**MPE**	**LMAPE**	**LRMSE**	**LADMPE**
α	-151.94 (2.37)	7.50 (0.00)	8.50 (0.00)	2.37 (0.08)
β	12368.49 (41957.46)	44.87 (12.24)	72.04 (19.88)	7680.98 (735.76)

Table 3B17: Regression coefficients for GLMLOGG response surface regressions

	Response surface regressions			
Regression coefficient	**MPE**	**LMAPE**	**LRMSE**	**LADMPE**
α	-2.40 (1.87)	7.43 (0.00)	8.39 (0.00)	2.11 (0.09)
β	4768.78 (33017.63)	47.11 (8.07)	21.02 (1.08)	7661.16 (707.81)

Table 3B18: Regression coefficients for GLMSQRTP response surface regressions

	Response surface regressions			
Regression coefficient	**MPE**	**LMAPE**	**LRMSE**	**LADMPE**
α	43.86 (1.74)	7.43 (0.00)	8.40 (0.00)	2.01 (0.09)
β	6300.86 (30958.31)	32.27 (6.43)	14.78 (1.35)	8004.18 (702.89)

Table 3B19: Regression coefficients for GLMSQRTG response surface regressions

	Response surface regressions			
Regression coefficient	**MPE**	**LMAPE**	**LRMSE**	**LADMPE**
α	63.69 (1.55)	7.45 (0.00)	8.47 (0.00)	1.92 (0.08)
β	-6348.01 (27826.75)	24.95 (6.92)	36.18 (9.83)	8009.81 (687.22)

Table 3B20: Regression coefficients for LOGNORM response surface regressions

	Response surface regressions			
Regression coefficient	**MPE**	**LMAPE**	**LRMSE**	**LADMPE**
α	52.17 (1.70)	7.46 (0.00)	8.46 (0.00)	1.89 (0.10)
β	-47610.97 (30514.46)	36.78 (7.41)	40.14 (8.65)	8123.00 (830.41)

Table 3B21: Regression coefficients for GG response surface regressions

	Response surface regressions			
Regression coefficient	**MPE**	**LMAPE**	**LRMSE**	**LADMPE**
α	-60.42	7.49	8.48	2.29
	(2.21)	(0.00)	(0.00)	(0.09)
β	-30592.30	32.25	40.50	7699.85
	(41448.16)	(10.48)	(12.11)	(733.74)

Table 3B22: Regression coefficients for GB2LOG response surface regressions

	Response surface regressions			
Regression coefficient	**MPE**	**LMAPE**	**LRMSE**	**LADMPE**
α	134.39	7.43	8.41	2.01
	(1.62)	(0.00)	(0.00)	(0.08)
β	-4550.79	17.68	6.36	7719.11
	(29380.16)	(5.35)	(1.41)	(688.62)

Table 3B23: Regression coefficients for GB2SQRT response surface regressions

	Response surface regressions			
Regression coefficient	**MPE**	**LMAPE**	**LRMSE**	**LADMPE**
α	-0.61	7.46	8.45	2.44
	(2.70)	(0.00)	(0.00)	(0.08)
β	-13166.18	60.05	83.16	8279.92
	(50627.97)	(14.09)	(19.16)	(757.08)

Table 3B24: Regression coefficients for FMMLOGG response surface regressions

	Response surface regressions			
Regression coefficient	**MPE**	**LMAPE**	**LRMSE**	**LADMPE**
α	142.12	7.42	8.40	2.28
	(2.85)	(0.00)	(0.00)	(0.07)
β	-93880.38	80.92	27.85	7461.19
	(67740.82)	(20.79)	(6.12)	(779.31)

Table 3B25: Regression coefficients for FMMSQRTG response surface regressions

	Response surface regressions			
Regression coefficient	**MPE**	**LMAPE**	**LRMSE**	**LADMPE**
α	-25.08	7.44	8.40	2.36
	(2.35)	(0.00)	(0.00)	(0.08)
β	-96992.48	78.57	58.99	7962.63
	(44199.43)	(12.04)	(7.25)	(726.73)

Table 3B26: Regression coefficients for EEE response surface regressions

	Response surface regressions			
Regression coefficient	**MPE**	**LMAPE**	**LRMSE**	**LADMPE**
α	-2.52	7.42	8.39	2.08
	(1.86)	(0.00)	(0.00)	(0.08)
β	9424.60	39.46	22.01	8165.22
	(32387.59)	(6.22)	(1.16)	(685.22)

Table 3B27: Regression coefficients for CDEM response surface regressions

	Response surface regressions			
Regression coefficient	**MPE**	**LMAPE**	**LRMSE**	**LADMPE**
α	-13.02 (1.91)	7.45 (0.00)	8.40 (0.00)	2.16 (0.09)
β	6579.60 (33749.36)	15.36 (6.44)	8.26 (1.00)	7808.60 (741.31)

Table 3B28: Regression coefficients for CDEO response surface regressions

Chapter 4

Estimating the Full Distribution

Healthcare Cost Regressions: Going Beyond the Mean to Estimate the Full Distribution

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

Summary

Understanding the data generating process behind healthcare costs remains a key empirical issue. Although much research to date has focused on the prediction of the conditional mean cost, this can potentially miss important features of the full conditional distribution such as tail probabilities. We conduct a quasi-Monte Carlo experiment using English NHS inpatient data to compare 14 approaches to modelling the distribution of healthcare costs: nine of which are parametric, and have commonly been used to fit healthcare costs, and five others designed specifically to construct a counterfactual distribution. Our results indicate that no one method is clearly dominant and that there is a trade-off between bias and precision of tail probability forecasts. We find that distributional methods demonstrate significant potential, particularly with larger sample sizes where the variability of predictions is reduced. Parametric distributions such as log-normal, generalised gamma and generalised beta of the second kind are found to estimate tail probabilities with high precision, but with varying bias depending upon the cost threshold being considered.

JEL classification: C1; C5

Key words: Healthcare costs; heavy tails; counterfactual distributions; quasi-Monte Carlo

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

4.1 Introduction

Econometric models of healthcare costs have many uses: to estimate key parameters for populating decision models in cost-effectiveness analyses (Hoch et al., 2002); to adjust for healthcare need in resource allocation formulae in publically funded healthcare systems (Dixon et al., 2011); to undertake risk adjustment in insurance systems (Van de Ven and Ellis, 2000) and to assess the effect on resource use of observable lifestyle characteristics such as smoking and obesity (Johnson et al., 2003; Cawley and Meyerhoefer, 2012; Mora et al., 2014). The distribution of healthcare costs poses substantial challenges for econometric modelling. Healthcare costs are non-negative, highly asymmetric and leptokurtic, and often exhibit a large mass point at zero. The relationships between covariates and costs are likely to be non-linear. Basu and Manning (2009) provide a useful discussion of these issues. The relevance and complexity of modelling healthcare costs has led to the development of a wide range of econometric approaches, and a description of these can be found in Jones (2011).

Much of the focus in comparisons of regression methods for the analysis of healthcare cost data has centered on predictions of the conditional mean of the distribution, $E(y|X)$ (Deb and Burgess, 2003; Veazie et al., 2003; Basu et al., 2004; Buntin and Zaslavsky, 2004; Gilleskie and Mroz, 2004; Manning et al., 2005; Basu et al., 2006; Hill and Miller, 2010; Jones, 2011; Jones et al., 2013, 2014). Applied researchers commonly model cost data using generalised linear models (GLMs) (Blough et al., 1999). This framework offers a relatively simple way to incorporate non-linearities in the relationship between the conditional mean and observed covariates. Furthermore, GLMs allow for heteroskedasticity through a choice of a 'distribution' which specifies the conditional variance as a function of the conditional mean. GLMs use pseudo-maximum likelihood estimation where the researcher is required only to specify the form of the mean and the variance. Unlike maximum likelihood estimation, where consistency requires that the whole likelihood function is correctly specified, pseudo-maximum likelihood is consistent so long as the mean is correctly specified with the choice of 'distribution' affecting the efficiency of estimates. Whilst the GLM framework has attractive properties for researchers concerned only with $E(y|X)$, there are important limitations with this method. GLMs have been found to perform badly with heavy-tailed

data (Manning and Mullahy, 2001), and they implicitly impose restrictions on the entire distribution. For example, whatever distribution is adopted, the skewness must be directly proportional to the coefficient of variation and the kurtosis is linearly related to the square of the coefficient of variation (Holly, 2009). Whilst they may be well placed to estimate $E(y|X)$ and $Var(y|X)$, they cannot produce estimates of $F(y|X)$ or $P(y > k|X)$.

While the mean is an important feature of a distribution, which is essential when the analysis is concerned with the expected total cost, it is generally not the only aspect that is of interest to policymakers (Vanness and Mullahy, 2007). Analysis based solely on the mean misses out potentially important information in other parts of the distribution (Bitler et al., 2006). As a result, a growing literature in econometrics has developed techniques to model the entire distribution, $F(y|X)$, thus 'going beyond the mean' (Fortin et al., 2011). In health economics there is a particular emphasis on identifying individuals or characteristics of individuals that lead to very large costs and there is a demand for empirical strategies to "target the high-end parameters of particular interest" including tail probabilities, $P(y > k)$ (Mullahy, 2009).

In this paper we conduct a quasi-Monte Carlo experiment to compare the fit of the entire conditional distribution of healthcare costs using competing approaches proposed in the economics literature. We therefore consider approaches which offer greater flexibility in terms of their potential applications by estimating $F(y|X)$, imposing fewer restrictions on skewness and kurtosis and allowing for a greater range of estimated effects of a covariate.

We first consider developments in the use of flexible parametric distributions for modelling healthcare costs (Manning et al., 2005; Jones et al., 2014), which have been applied to healthcare costs principally in order to overcome the challenge posed by heavy-tailed data. Unlike the GLM framework, these models impose a functional form for the entire distribution with estimation by maximum likelihood. As a result, an estimate of $f(y|X)$ is produced, which can then be used to calculate $E(y|X)$, $Var(y|X)$[1] and $P(y > k|X)$ as required. By using flexible distributions, the restrictions on skewness and kurtosis can be relaxed somewhat (McDonald et al., 2013), which is likely to lead to a better fit of the full distribution according to measures based on log-likelihood (Jones et al., 2014).

[1]Note that population moments may not be defined for all ranges of parameter estimates (Mullahy, 2009).

A related development is the use of finite mixture models (FMM), which allow the distribution to be estimated as a weighted sum of distribution components (Deb and Trivedi, 1997; Deb and Burgess, 2003). These are also estimated using maximum likelihood, but are often referred to as semi-parametric, since the number of components could, in principle, be increased to approximate any distribution. In this paper we group FMM with the fully parametric distributions given the similarities to these approaches, especially since we use a fixed number of components.

Other developments regarding the estimation of $f(y|X)$ for healthcare costs are less parametric, typically involving dividing the outcome variable into discrete intervals and estimating parameters for each of these intervals. Gilleskie and Mroz (2004) propose using a conditional density approximation estimator for healthcare costs to calculate $E(y|X)$ and other moments, with the density function approximated by a set discrete hazard rates. To implement this, Jones et al. (2013) use an approach based on Han and Hausman (1990), where $F(y|X)$ is estimated by creating a categorical variable that denotes the cost interval into which each observation falls, and running an ordered logit with this as the dependent variable. This implementation is slightly different from what is proposed by Gilleskie and Mroz (2004), but has the advantage of being conceived in order to fit $F(y|X)$ and ties into a related literature on semi-parametric estimators for conditional distributions (Han and Hausman, 1990; Foresi and Peracchi, 1995; Chernozhukov et al., 2013). While the ordered logit specification used in the Han and Hausman (1990) method allows for flexible estimation of the thresholds in the latent scale, methods such as Foresi and Peracchi (1995) instead estimate a series of separate logit models.

More recently, Chernozhukov et al. (2013) propose that a continuum of logits should be estimated (one for each unique value of the outcome variable) to allow for an even greater range of estimates for the effect of a covariate. In an application to Dutch health expenditures, de Meijer et al. (2013) use the Chernozhukov et al. (2013) method to decompose changes in the distribution of health expenditures between two periods. The authors find that the effect of covariates varies across the distribution of health expenditures, which would have been missed if analysis had focused solely on the mean. They also find that pharmaceutical costs are growing mainly at the top of the distribution due to structural effects, whereas growth in hospital care costs is observed more in the mid-

dle of the distribution and can be explained by changes in the observed determinants of expenditure.

The methods described above seek to estimate the full distribution, by modelling $F(y|X)$ for different values of y (interval thresholds) and imposing varying degrees of flexibility on the covariate effects for these. An alternative is to construct $F(y|X)$ through the inverse of the distribution function, the quantile function $q_\tau(X)$.[2] We consider two methods which estimate a range of quantiles separately as functions of the covariates to allow for flexibility as to the estimated effects of each regressor across the full range of the distribution. The first was proposed by Machado and Mata (2005) and Melly (2005) and uses a series of quantile regressions to estimate the full range of quantiles across the distribution (hereafter MM method). Quantile regressions have been used where the outcome variable was healthcare costs for analysing the varying effects of race at different points of the distribution (Cook and Manning, 2009). However we were unable to find any applications of the MM method to construct a complete estimate of $F(y|X)$ with healthcare costs as the outcome variable, although the applications in the original papers were to wages, which share similar distributional characteristics. Quantile functions can alternatively be estimated using recentred-influence-function (RIF) regression (Firpo et al., 2009), where the outcome variable is first transformed according to the recentred-influence-function and then regression used to model the effects of covariates.

This paper provides a systematic comparison of parametric and distributional methods[3] for fitting the full distribution of healthcare costs using real data in a quasi-Monte Carlo experiment. As such, it is novel in two ways: firstly, it provides a methodology for comparing the distributional fit of models which are neither special cases nor estimated using the same procedure, and secondly it is the first paper to compare competing econometric approaches for modelling the distribution of healthcare costs. We find that distributional methods demonstrate significant potential in modelling tail probabilities, particularly with larger sample sizes where the variability of predictions is reduced. Parametric distributions such as log-normal, generalised gamma and generalised beta of the second kind are found to estimate tail probabilities with high precision, but with varying

[2] $\tau \in (0, 1)$ denotes the quantile being considered.
[3] This term was used in Fortin et al. (2011).

bias depending upon the cost threshold being considered.

The study design is described in the next section, followed by a detailed description of the methods compared. We then discuss the results, and place these in the context of related research, and remark upon some of the limitations of our study and possible extensions for future work.

4.2 Methodology and Data

4.2.1 Overview

Rather than comparing competing approaches for estimating $E(y|X)$, which is the focus of most empirical work in this area (Mullahy, 2009), we assess performance in terms of tail probabilities, $P(y > k)$, for varying levels of k to assess the fit of the entire distribution, $F(y|X)$. We compare a number of different regression methods, each with a different number of estimated parameters. Since more complex methods may capture idiosyncratic characteristics of the data as well as the systematic relationships between the dependent and explanatory variables, there is a concern that better fit will not necessarily be replicated when the model is applied to new data (Bilger and Manning, 2014). To guard against this affecting our results, we use a quasi-Monte Carlo design where models are fitted to a sample drawn from an 'estimation' set and performance is evaluated on a 'validation' set. This means that methods are assessed when being applied to new data.[4] Each method is used to produce an estimate of the whole distribution $F(y|X)$, which can then be used to produce a counterfactual distribution given the covariates in the 'validation' set. The counterfactual distribution could be constructed for certain X values, such as patients aged over 65 years old, or female patients only. In this paper we construct the counterfactual distribution for all X values. We evaluate performance based on forecasting tail probabilities, $P(y > k)$.[5]

4.2.2 Data

Our data comes from the English administrative dataset, Hospital Episode Statistics

[4]There are substantial precedents for using split-sample methods to evaluate different regression methods for healthcare costs, for example Duan et al. (1983); Manning et al. (1987).

[5]The values of k are not used in estimating the distribution $F(y|X)$.

(HES)[6], for the financial year 2007-2008. We have excluded spells which were primarily mental or maternity healthcare and all spells taking place within private sector hospitals.[7] The remaining spells constitute the population of all inpatient episodes, outpatient visits and A&E attendances that were completed within 2007-2008 for all patients who were admitted to English NHS hospitals (where treatment was not primarily mental or maternity healthcare). Spells are costed using tariffs from 2008-2009[8] by applying the relevant tariff to the most expensive episode within the spell (where a spell can be thought of as a discrete admission).[9] Our analysis is undertaken at the patient level and so we sum the costs in all spells for each patient to create the dependent variable, giving us 6,164,114 observations in total. The empirical density and cumulative distribution of the outcome variable can be seen in Figure 4.1 and descriptive statistics are found in Table 4.1.[10]

N	$6,164,114$	
Mean	$£2,610$	
Median	$£1,126$	
Standard deviation	$£5,088$	
Skewness	13.03	
Kurtosis	363.18	
Minimum	$£217$	
Maximum	$£604,701$	
	% observations	% of total costs
$> £500$	82.96%	97.20%
$> £1,000$	55.89%	89.80%
$> £2,500$	27.02%	72.35%
$> £5,000$	13.83%	54.65%
$> £7,500$	6.92%	38.67%
$> £10,000$	4.09%	29.35%

Table 4.1: Descriptive statistics for hospital costs

In order to tie in with existing literature on comparisons of econometric methods for healthcare costs, we use a set of morbidity characteristics which we keep constant for

[6]HES is maintained by the NHS Information Centre, now known as the Health and Social Care Information Centre.

[7]This dataset was compiled as part of a wider project considering the allocation of NHS resources for secondary care services. Since a lot of mental healthcare is undertaken in the community and with specialist providers, and hence not recorded in HES, the data is incomplete. In addition, 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 'needs' (morbidity) characteristics, and accordingly for the setting of healthcare budgets are determined using alternative mechanisms.

[8]Reference costs for 2005-2006, which were the basis for the tariffs from 2008-2009, were used when 2008-2009 tariffs were unavailable.

[9]This follows standard practice for costing NHS activity.

[10]Costs above $£10,000$ are excluded in these plots to make illustration clearer.

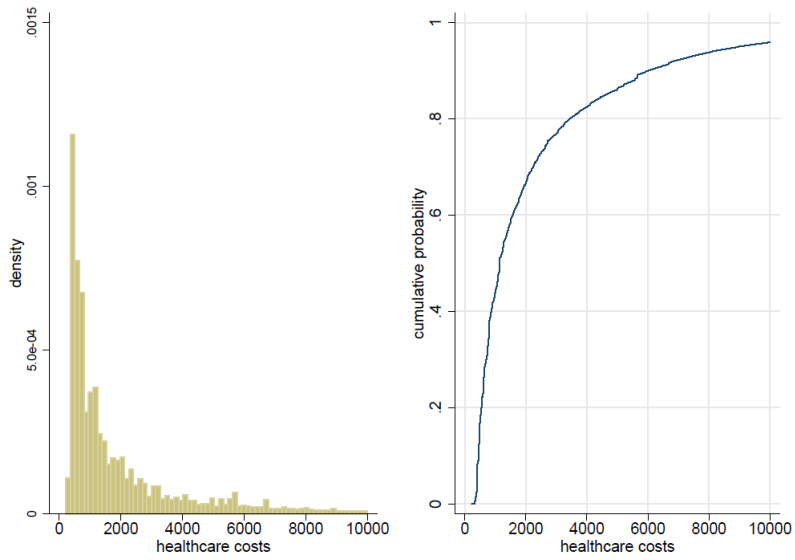


Figure 4.1: Empirical density and cumulative distribution of healthcare costs

each regression method. In addition, we control for age and sex using an interacted, cubic specification, which leaves us with a set of regressors similar to a simplified resource allocation formula where health expenditures are modelled as a function of need (proxied using detailed socio-demographic and morbidity information) (Dixon et al., 2011). In total we use 24 morbidity markers, adapted from the ICD10 chapters (WHO, 2007), which are coded as one if one or more spells occur with any diagnosis within the relevant subset of ICD10 chapters (during the financial year 2007-2008) and zero otherwise.

To give some illustration of the features of the data conditional upon these covariates we construct an index using these regressors and divide the data from the 'estimation' set into five quantiles (quintiles) according to the value of the index.[11] For each quintile we display the empirical distribution of log-costs[12] in Figure 4.2, and in particular pick out those that exceed $\ln(£10,000)$. It is clear from Figure 4.2 that the conditional distributions of log-costs (and thus costs) vary dramatically by quintile of covariates in terms of their shape, range and number of high cost patients, with 17% of observations with annual costs greater than $£10,000$ in the most morbid patients, compared to a population average of 4.09% (and 0.14% in the least morbid quintile). An analysis looking only at the mean of each quintile would overlook these features of the data.

[11] This is constructed by regressing cost against the regressors using OLS and taking the predicted cost.

[12] A log-transformation is used to make the whole distribution easier to illustrate and $P(y > k) = P(\ln(y) > \ln(k))$ since it is a monotonic transformation.

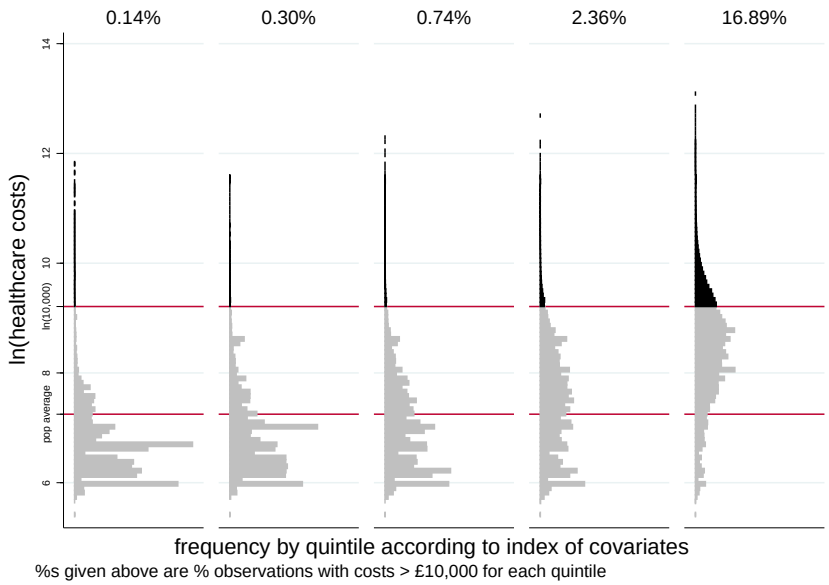


Figure 4.2: Empirical distribution of log-costs for each of the 5 quintiles of the linear index of covariates

We also carry out a similar analysis, this time using untransformed costs and dividing the 'estimation' set into 10 quantiles (deciles) of the linear index of covariates, where we plot the kurtosis of each decile against its skewness. Parametric distributions impose restrictions upon possible skewness and kurtosis: one-parameter distributions are restricted to a single point (e.g. the normal distribution imposes a skewness of 0 and a kurtosis of 3), two-parameter distributions allow for a locus of points to be estimated, and distributions with three or more parameters allow for spaces of possible skewness and kurtosis combinations. Figure 4.3[13] shows that the data is non-normal and provides motivation for flexible methods, since they appear better able to model the higher moments of the conditional distributions of the outcome variable analysed here.[14] We do not represented the other approaches used in this paper in this Figure, since the skewness and kurtosis space is not defined for these approaches. This is because they discretise the distribution or estimate several models, or both, and the effects on implied skewness and kurtosis is unclear.

[13]Key for abbreviations: GB2 – generalised beta of the second kind, SM – Singh-Maddala, B2 – beta of the second kind, GG – generalised gamma, LN – log-normal, WEI – Weibull.

[14]A similar analysis can be found in Pentsak (2007). Note also that the lower bound of the Pearson Type IV distribution, used in Holly and Pentsak (2006), is equal to the upper bound for the beta of the second kind distribution (also known as Pearson Type VI).

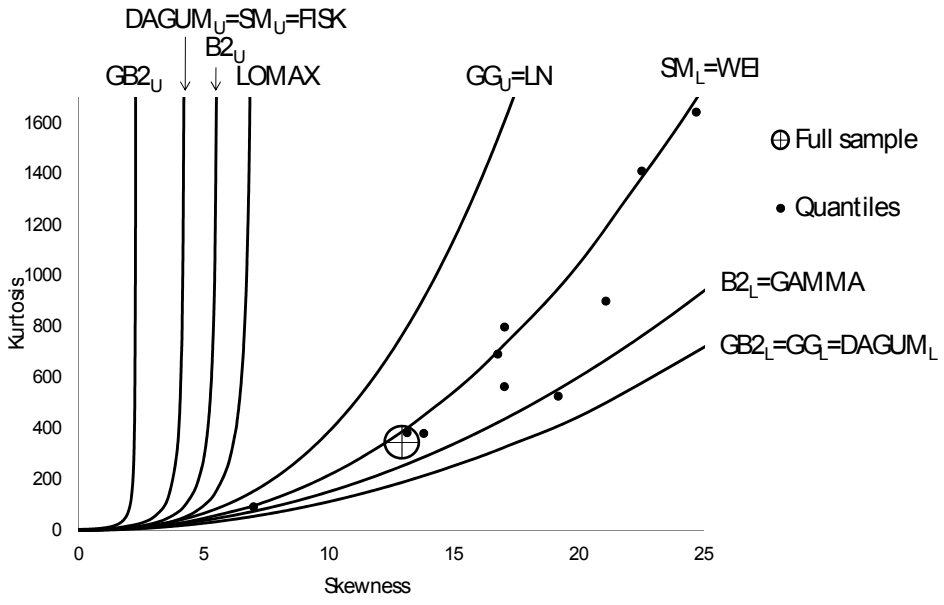


Figure 4.3: Kurtosis against skewness for each of the 10 deciles of the linear index of covariates

Note: Taken from Jones et al. (2014) and adapted from McDonald et al. (2013). The dots shown on Figure 4.3 were generated as follows: the data were divided into ten subsets using the deciles of a simple linear predictor for healthcare costs using the set of regressors used in this paper. Figure 4.3 plots the skewness and kurtosis coefficients of actual healthcare costs for each of these subsets, the skewness and kurtosis coefficient of the full estimation sub-population (represented by the larger circle with cross) and theoretically possible skewness-kurtosis spaces and loci for parametric distributions considered in the literature.

4.2.3 Quasi-Monte Carlo design

In order to fully exploit the large dataset at our disposal, before we undertake analysis we randomly divide the 6,164,114 observations into two equally sized groups: an 'estimation' set and a 'validation' set (each with 3,082,057 observations). Because researchers using observational data from social surveys typically have fewer observations in their datasets than are present in our 'estimation' set, we draw samples from within the 'estimation' set. On these samples we estimate the regressions that will later be evaluated using the 'validation' set data. In total we randomly draw 300 samples with replacement: 100 samples of each size N_s ($N_s \in$ 5,000; 10,000; 50,000), where samples with $N_s =$ 5,000 or 10,000 may be thought of as having a similar number of observations as small to moderately sized datasets (Basu and Manning, 2009). We estimate 14 methods using the outcome and regressor data from each sample, where each method can be used to construct a counterfactual distribution of costs $F(y|X)$ (more details on each method are found in the Empirical Models section).

Then using all 3,082,057 observations in the 'validation' set, we use the covariates from the data (but not the outcome variable) to construct $F(y|X)$ for each method. Depending upon which method is being considered, we can either directly obtain $P(y > k|X)$, which we then integrate out over values of X to produce an estimate of $P(y > k)$, or we can use $F(y|X)$, which we integrate out over values of X, to give $F(y)$, to then estimate $P(y > k)$. This process could be carried out for a set of specific X values – for example patients aged over 65 years old – or over all X values (as in this paper). Once the estimate of $P(y > k)$ is produced for the 'validation' set using either method, it can be compared to the observed empirical proportion of costs in the data that exceeds the threshold k.[15] In this paper we choose round values for k throughout the distribution of the outcome variable (numbers in brackets correspond to % of population mean): $k \in$ £500 (19%); £1,000 (38%); £2,500 (96%); £5,000 (192%); £7,500 (287%); £10,000 (383%).[16] Results displayed look at performance across each replication for given method with a given sample

[15] It is worth noting that the practice of comparing observed versus empirical probabilities forms the basis of the Andrews (1988) chi-square test, although this is designed for use with parametric methods only, and as such is not implemented in this paper, where we are interested in the performance of both parametric and semi-parametric approaches.

[16] Table 4.1 gives the proportion of observations in the population that exceed these thresholds.

size. We construct a ratio of predicted $P(y > k)$ to observed $P(y > k)$ and look at the average of these across all replications. In addition, we analyse the variability of these ratios, for each method and a given sample size, using the average absolute deviation from the average computed ratio, as well as their standard deviation and their range.

4.3 Empirical models

4.3.1 Overview

We compare, in total, the performance of 14 different estimators, which we divide into two groups: parametric methods and distributional methods. In addition, we compare results to a naïve estimate based purely on the sample, where the researcher is assumed to forecast the same tail probability for the 'validation' set as observed in the 'estimation' sample (without considering the observed covariates in either dataset). First we describe each of the parametric distributions and provide its conditional probability density function – $f(y|X)$ – the equation to calculate $P(y > k|X)$, as well as the procedure for integrating over X in order to produce an estimate of $P(y > k)$. For the remaining five methods, the procedure is more varied and complex, so we provide a detailed account of the steps required to produce estimates of $P(y > k)$ for all of these distributions. Table 4.2 provides a key for the abbreviations used for each method throughout the remainder of the paper.

GB2_LOG	generalised beta of the second kind (log link)
GB2_SQRT	generalised beta of the second kind ($\sqrt{}$-link)
GG	generalised gamma (log link)
GAMMA	two-parameter gamma (log link)
LOGNORM	log-normal (log link)
WEIB	Weibull (log link)
EXP	exponential (log link)
FMM_LOG	two-component finite mixture of gamma densities (log link)
FMM_SQRT	two-component finite mixture of gamma densities ($\sqrt{}$-link)
HH	Han and Hausman
FP	Foresi and Peracchi
CH	Chernozhukov, Fernández-Val and Melly (linear probability model)
MM	Machado and Mata – Melly (log-transformed outcome)
RIF	recentered-influence-function regression (linear probability model)

Table 4.2: Key for method labels

4.3.2 Parametric methods

All nine of the parametric approaches that we consider, including two variants of finite mixture models[17], are estimated by specifying the full conditional distribution of healthcare costs using between one and five parameters. While it is possible in principle to allow shape parameters to vary with covariates, preliminary work showed that this produced unreliable and uninterpretable results, so in all cases we only specify location parameters as functions of covariates. This means that all models have only one parameter depending upon covariates, except FMM_LOG and FMM_SQRT which have scale parameters in each component that are allowed to vary with covariates. All other parameters are estimated as scalars. In Table 4.3 we give the conditional probability density function and the conditional survival function for each model we compare.[18]

[17]These are elsewhere considered to be semi-parametric, since the number of components can vary, but we fix the number of components as two, meaning that they are essentially parametric.

[18]Note that certain distributions' notation could be simplified, the parameterisation is chosen to maximise the reader's ability to see how distributions are related to one another.

Model	$\mathbf{f}(\mathbf{y}\|\mathbf{X}) =$	$\mathbf{P}(\mathbf{y} > \mathbf{k}\|\mathbf{X}) =$
GB2_LOG	$\frac{ay^{ap-1}}{\exp(X\beta)^{ap}B(p,q)[1+(\frac{y}{\exp(X\beta)})^a](p+q)}$	$1 - I_Z(p,q)^*$ where $z = \left(\frac{k}{\exp(X\beta)}\right)^a$
GB2_SQRT	$\frac{ay^{ap-1}}{(X\beta)^{2ap}B(p,q)[1+(\frac{y}{(X\beta)^2})^a](p+q)}$	$1 - I_Z(p,q)^*$ where $z = \left(\frac{k}{(X\beta)^2}\right)^a$
GG	$\frac{\kappa}{\sigma y\Gamma(\kappa^{-2})}\left(\kappa^{-2}\left(\frac{y}{\exp(X\beta)}\right)^{\kappa/\sigma}\right)^{\kappa^{-2}} exp\left(-\kappa^{-2}\left(\frac{y}{\exp(X\beta)}\right)^{\kappa/\sigma}\right)$	if $\kappa > 0$: $1 - \Gamma(z;\kappa^{-2})^{**}$ if $\kappa < 0$: $\Gamma(z;\kappa^{-2})^{**}$ where $z = \kappa^{-2}\left(\frac{k}{\exp(X\beta)}\right)^{\kappa/\sigma}$
GAMMA	$\frac{1}{y\Gamma(\kappa^{-2})}\left(\kappa^{-2}\left(\frac{y}{\exp(X\beta)}\right)\right)^{\kappa^{-2}} exp\left(-\kappa^{-2}\left(\frac{y}{\exp(X\beta)}\right)\right)$	$\kappa > 0$: $1 - \Gamma(z;\kappa^{-2})^{**}$ if $\kappa < 0$:$\Gamma(z;\kappa^{-2})^{**}$ where $z = \kappa^{-2}\left(\frac{k}{\exp(X\beta)}\right)$
LOGNORM	$\frac{1}{\sigma y\sqrt{2\pi}}\exp\left(\frac{-(\ln y - X\beta)^2}{2\sigma^2}\right)$	$1 - \Phi\left(\frac{\ln k - X\beta}{\sigma}\right)$
WEIB	$\frac{1}{\sigma y}\left(\frac{y}{\exp(X\beta)}\right)^{\frac{1}{\sigma}} exp\left(-\left(\frac{y}{\exp(X\beta)}\right)^{\frac{1}{\sigma}}\right)$	$\exp\left(-\left(\frac{k}{\exp(X\beta)}\right)^{\frac{1}{\sigma}}\right)$
EXP	$\frac{1}{\exp(X\beta)}\exp\left(\frac{-y}{\exp(X\beta)}\right)$	$\exp\left(-\frac{k}{\exp(X\beta)}\right)$
FMM_LOG	$\sum_j^2 \pi_j \frac{y^{\alpha_j}}{y\Gamma(\alpha_j)\exp(X\beta_j)^{\alpha_j}}\exp\left(-\left(\frac{y}{\exp(X\beta_j)}\right)\right)$	$\sum_j^2 \pi_j (1 - \Gamma(z;\alpha_j))^{***}$ where $z = \frac{k}{\exp(X\beta_j)}$
FMM_SQRT	$\sum_j^2 \pi_j \frac{y^{\alpha_j}}{y\Gamma(\alpha_j)(X\beta_j)^{2\alpha_j}}\exp\left(-\left(\frac{y}{(X\beta_j)^2}\right)\right)$	$\sum_j^2 \pi_j (1 - \Gamma(z;\alpha_j))^{***}$ where $z = \frac{k}{(X\beta_j)^2}$

*where $I_Z(p,q) = \frac{1}{B(p,q)}\int_0^z \frac{t^{p-1}}{(1+t)^{p+q}}dt$ is the incomplete beta function ratio.

**where $\Gamma(z;\kappa^{-2}) = \frac{1}{\Gamma(\kappa^{-2})}\int_0^z t^{(\kappa^{-2}-1)}\exp(-t)dt$.

***where $\Gamma(z;\alpha_j) = \frac{1}{\Gamma(\alpha_j)}\int_0^z t^{(\alpha_j-1)}\exp(-t)dt$.

Table 4.3: Forms of density functions and survival functions for parametric distributions

The generalised beta of the second kind[19] is a four-parameter distribution that was applied to modelling healthcare costs by Jones (2011) specifying the location parameter as a linear function of covariates using software developed by Jenkins (2009). Jones et al. (2014) estimated the distribution with a log link (GB2_LOG) making it more comparable with commonly used approaches. With this specification, for example, GG (as proposed by Manning et al., 2005) becomes a limiting case of GB2_LOG. Jones et al. (2013) also compared GB2_SQRT as well as GB2_LOG against a broad range of models, finding that the GB2_SQRT performed particularly well in terms of accurately predicting mean individual healthcare costs. GG has been compared more extensively in terms of predicting mean healthcare costs, having been found to out-perform a GLM log link with gamma-distribution in the presence of heavy tails using simulated data (Manning et al., 2005), and a number of models within the GLM framework when a log link is appropriate using American survey data; the Medical Expenditures Panel Survey (Hill and Miller, 2010). GB2_LOG, GG and LOGNORM are compared in Jones et al. (2014), with some indication that GB2_LOG better fits the entire distribution with lower AIC and BIC, although LOGNORM better predicts tail probabilities associated with the majority of high costs considered. We also consider further special cases of GG (and GB2_LOG) with two parameters (GAMMA and WEIB) and with one parameter (EXP).[20]

Finite mixture models have been used in health economics in order to allow for heterogeneity both in response to observed covariates and in terms of unobserved latent classes (Deb and Trivedi, 1997). Heterogeneity is modelled through a number of components, denoted C, each of which can take a different specification of covariates (and shape parameters, where specified), written as $f_j(y|X)$, with an associated parameter for the probability of belonging to each component, π_j. The general form of the probability density function of finite mixture models is given as:

$$f(y|X) = \sum_{j}^{C} \pi_j f_j(y|X) \tag{4.1}$$

[19] Also known as generalised-F, see Cox (2008).

[20] The parametric distributions chosen are the set of distributions that are typically used in health economics. There are many other candidate distributions, for example Walls (2005) uses the skew-normal distribution to model film returns (which should exhibit empirically similar distributions).

We use two gamma-distributed components in our comparison.[21] In one of the models used, we allow for log links in both components (FMM_LOG), and in the other we allow for a square root link in both components (FMM_SQRT). In both, the probability of class membership is treated as constant for all individuals. Unlike the other parametric methods, this approach can allow for a multi-modal distribution of costs. In this way, finite mixture models represent a flexible extension of parametric models (Deb and Burgess, 2003). Using increasing numbers of components, it is theoretically possible to fit any distribution, although in practice researchers tend to use few components (two or three) and achieve good approximation to the distribution of interest (Heckman, 2001).

Once we have obtained estimates of location parameters (all βs for each regressor) and shape parameters for each distribution, these are stored in memory and then used to generate estimates of $P(y > k|X)$, where values for X are the observed covariates in the 'validation' set. These estimated conditional tail probabilities will vary across each possible combination of X, and hence for any given individual i, and so we take the average in order to 'integrate out' these to provide us with a single estimate of $P(y > k)$ for each method and replication, which can be compared to the proportion of costs empirically observed to exceed k. We then take the average across all replications of $P(y > k)$ for each method in order to assess bias and analyse the variability across replications as an indicator of precision.

4.3.3 Distributional methods

4.3.4 Methods using the cumulative distribution function

Of the remaining five methods that we compare, three involve estimation of the conditional distribution function and two operate through the quantile function. First we consider the methods which estimate the conditional distribution function $F(y|X)$. Han and Hausman (1990) adopts a proportional hazards specification, where the baseline hazard is allowed to vary non-parametrically across a number, denoted D_{HH}, of intervals of a discretised continuous outcome variable. The logarithm of the integrated baseline

[21] Preliminary work showed that models with a greater number of components lead to problems with convergence in estimation. Empirical studies such as Deb and Trivedi (1997) provide support for the two components specification for healthcare use.

hazard for each of the $D_{HH} - 1$ intervals (one is arbitrarily omitted for estimation) is estimated as a constant $\delta_{D_{HH}}$. The effects of covariates are estimated using a particular functional form, which is typically linear. This approach is similar to the semi-parametric Cox proportional hazard model (Cox, 1972), but differs in that the baseline hazard is not regarded as a nuisance parameter and is better suited to data with many ties of the outcome variable (or in the case of a discrete outcome). In order to implement this method, we construct a categorical variable for each observation, indicating the interval into which the value of the outcome variable falls. This is then used as the dependent variable in an ordered logit regression on the covariates. The cut-points are estimates of the baseline hazard within each interval $\delta_{D_{HH}}$. The authors argue that given a large sample size, finer intervals should improve the efficiency of the estimator, without providing guidance on a specific number of intervals to be used. As a result we carried out preliminary work to establish the largest number of intervals that could be used for each sample size whilst maintaining good convergence performance,[22] which resulted in a maximum of 33 intervals for sample sizes 5,000 and 10,000, and 36 intervals for a sample size of 50,000.

Foresi and Peracchi's (1995) method is similar to Han and Hausman's (1990) in that it divides the data into a set of discrete intervals. Rather than using an ordered logit specification, Foresi and Peracchi (1995) estimate a series of logit regressions. For each upper boundary of the $D_{FP} - 1$ intervals (the highest value interval is excluded), an indicator variable is created which is equal to one if the observation's observed cost is less than or equal to the upper boundary, and zero otherwise. These are then used as dependent variables in $D_{FP} - 1$ logit regressions each using the full set of regressors. In their application to excess returns in their paper they use zero, as well as the 10th, 15th, 20th, ... , 80th, 85th and 90th percentiles as boundaries. While we do not have information on patients with zero costs in our dataset, we base our intervals on their specification of the dependent variables by using the 5th, 10th, 15th, ... , 85th, 90th and 95th percentiles (vigiciles).

The third approach that we compare is an extension of Foresi and Peracchi (1995) and is described in Chernozhukov et al. (2013). The crucial difference between the methods is that Chernozhukov et al. (2013) argue that a logit regression should be used for each

[22] This was taken to mean that the model converges at least 95 times out of the 100 samples.

unique value of the outcome variable. A continuum of indicator variables needs to be generated and then regression models are used to construct the conditional distribution functions for each value. Given the computational demand of this approach, and lack of variation in the indicator variables at low and high costs, de Meijer et al. (2013) use linear probability models in place of logit regressions. We also adopt this approach in our comparison, since preliminary work showed that, where it was possible to estimate both logit and linear probability models, there was little difference between the methods.

All of these methods are similar in that they can produce estimates of $P(y > k^*|X)$, where k^* represents one of the boundaries of the intervals generated using either Han and Hausman (1990) or Foresi and Peracchi (1995), or any cost value observed in the sample when implementing Chernozhukov et al. (2013). Since models are estimated without knowing what thresholds (k) the policymaker might be interested in, it is not always the case that $k^* = k$. Therefore, for all three methods described above, we use a weighted average of $P(y > k^*|X)$ for the nearest two values of k^* to k when $k^* \neq k$. Our weight is based on a simple linear interpolation:

$$P(y > k|X) = P(y > k_a^*|X) + \left(\frac{k - k_a^*}{k_b^* - k_a^*}\right)\left(P(y > k_b^*|X) - P(y > k_a^*|X)\right) \tag{4.2}$$

where k_a^* and k_b^* represent the thresholds analysed in estimation closest below and closest above k, respectively.[23]

Since we end up with an estimate for each observation of $P(y > k|X)$, we carry out the same procedure as with the parametric distributions. This means that we take the average of $P(y > k|X)$, thus 'integrating out' over all possible combinations of X and giving us an estimate of $P(y > k)$ to be compared against the empirical proportion.

4.3.5 Methods using the quantile function

Machado and Mata (2005) propose a method for constructing a counterfactual distribution based on a series of quantile regressions using the logged outcome variable. They suggest that a quantile (τ) is chosen at random by drawing from a uniform probability

[23] This should work well when there are a large number of k^* spaced throughout the distribution. When interested in high values of k this linear interpolation may be inappropriate if there are few high values of k^*, given the often large distances between a high cost and the next highest observed cost.

distribution between zero and one. After running the quantile regression for the drawn value, the set of estimated coefficients is used to predict the quantile given the covariate values observed for a randomly selected observation. The authors repeat this process 4500 times with replacement, generating a full counterfactual distribution. The theoretical motivation for this procedure is that each predicted quantile based on $q_\tau(X)$ represents a draw from the conditional distribution of healthcare costs ($f(y|X)$). Therefore drawing a random observation and forecasting q_τ enough times with random τ effectively integrates out X. Running such a large number of quantile regressions is computationally expensive, and so Melly (2005) suggest running a regression for a fixed number of quantiles spread over the full range of the distribution, e.g. for each percentile, rather than drawing a quantile at random. We use the Melly (2005) approach for the MM method, running quantile regressions for each percentile on the 'estimation' set, after log-transforming the outcome variable, and randomly choosing one of these quantiles to forecast for each observation in the 'validation' set.[24] Once this has been done, the forecasted values represent the counterfactual distribution of healthcare costs belonging to the 'validation' set. Therefore to produce an estimate of $P(y > k)$ we observe the proportion of the observations in the counterfactual distribution that exceed k.

Another method to estimate quantiles of the distribution is developed by Firpo et al. (2009), which employs recentred-influence-function regressions. For a given observed quantile (q_τ), a recentred-influence-function (RIF) is generated, which can take one of two values depending upon whether or not the observation's value of the outcome variable is less than or equal to the observed quantile:

$$RIF\left(y; q_\tau\right) = q_\tau + \frac{\tau - 1\left[y \leq q_\tau\right]}{f_y\left(q_\tau\right)} \tag{4.3}$$

Here, q_τ is the observed sample (τ) quantile, $1\left[y \leq q_\tau\right]$ is an indicator variable which takes the value one if the observation's value of the outcome variable is less than or equal to the observed quantile and zero otherwise, and $f_y\left(q_\tau\right)$ is the estimated kernel density of the distribution of the outcome variable at the value of the observed quantile. The recentred-influence-function is then used as the dependent variable in an OLS regression

[24] The prediction is exponentiated to achieve the quantile of the distribution of the levels of healthcare costs.

on the chosen covariates, which effectively constitutes a rescaled linear probability model. These estimated coefficients can then be used to predict the quantile being analysed for a given observation's covariates. Following the same thought process as MM, predictions based on $q_\tau(X)$ represent a draw from $f(y|X)$. This means that we can use the estimated quantile functions to predict a counterfactual distribution in the same way for the RIF method as we do for the MM method.[25]

4.4 Results

When analysing the performance of the methods, we calculate a ratio of the estimated $P(y > k)$ to the actual proportion of costs in the 'validation' set observed to exceed the threshold value k (see Table 4.4). Using a ratio allows for greater comparability when looking at performance at different thresholds. We will look at the average ratio across replications (with methods estimated on different samples drawn from the 'estimation' set[26]) as well as the variability of the ratios. The former indicates the bias associated with each method at a given k, while the latter indicates precision of the method. First we will look at results across methods for a given sample size and threshold cost value: $N_s = 5,000$ and $k = £10,000$.[27] Second we consider performance for a given sample size, with a range of values for the threshold cost value, since different methods may be better at fitting different parts of the distribution of healthcare costs: $N_s = 5,000$ and ($k \in$ £500; £1,000; £2,500; £5,000; £7,500; £10,000). Lastly performance at different sample sizes is evaluated at a given threshold cost value: ($N_s \in$ 5,000; 10,000; 50,000) and $k = £10,000$.

In Figure 4.4 we present the performance of the 14 methods in predicting the probability of a cost exceeding £10,000 in the validation set, when samples with $N_s = 5,000$ observations are used. The bars indicate the ratio of estimated to actual probability, and the capped spikes indicate the range of ratios across all of the replications. A ratio of one represents a perfect fit, i.e. the method correctly predicted that 4.10% of observations

[25] We calculate the recentred-influence-function using the level of costs and so no re-transformation is required unlike when using MM.

[26] Three samples were discarded when $N_s = 5,000$, due to being unable to form the categorical variable for HH. Only one sample was discarded when $N_s = 10,000$ and $N_s = 50,000$.

[27] We choose these values of N_s and k since they are the smallest and most challenging sample size and the largest and most economically interesting threshold value, respectively.

k	% observations in 'validation' set $> k$
£500	82.93%
£1,000	55.89%
£2,500	27.04%
£5,000	13.84%
£7,500	6.94%
£10,000	4.10%

Table 4.4: Actual empirical proportion of observations greater than k in the 'validation' set

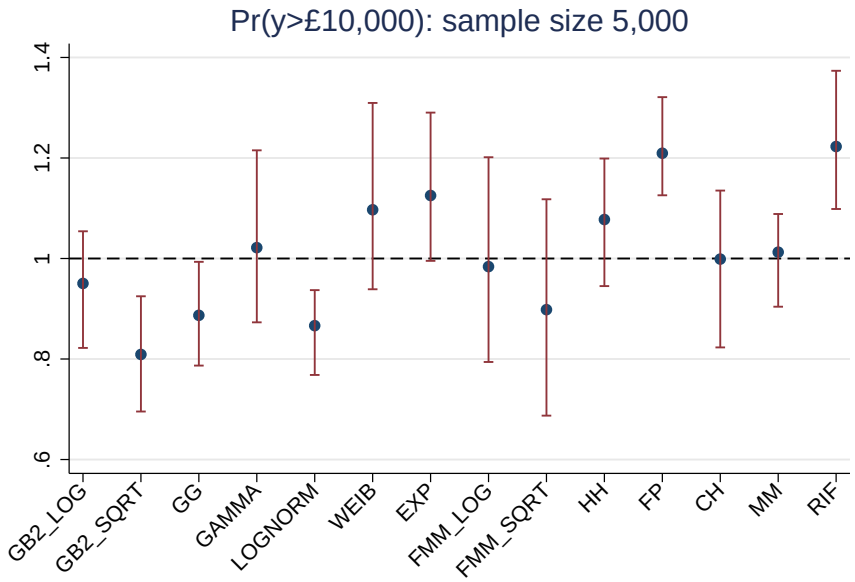


Figure 4.4: Performance of methods predicting the probability of a cost exceeding £10,000 at sample size 5,000

would exceed £10,000.

From Figure 4.4, it is clear that performance of the methods varies both in terms of bias (the point – the average ratio) and precision (the variability of ratios as depicted by the capped spikes showing the range). There is no clear pattern in terms of parametric versus distributional methods, since in both groups there are methods where the average ratio is seen to be near the desired value of one, as well as methods in both groups where the range of computed ratios does not contain one. In terms of bias, the best method is CH with an average ratio of almost exactly one. It appears that this is not the most precise method for $k = £10,000$, however, with a range of ratios: $0.82 - 1.14$, that is the fifth largest of all methods compared (the largest belongs to FMM_SQRT). To more

clearly represent the tradeoff between bias and precision, see Table 4.5, which gives the rankings of each method in terms of bias (absolute value of one minus the average ratio), the range of ratios and also the standard deviation of ratios.

Method	Bias	Range	Standard deviation
GB2_LOG	5th	6th	6th
GB2_SQRT	12th	5th	3rd
GG	9th	4th	2nd
GAMMA	4th	11th	11th
LOGNORM	11th	1st	1st
WEIB	7th	12th	12th
EXP	10th	9th	8th
FMM_LOG	3rd	13th	14th
FMM_SQRT	8th	14th	13th
HH	6th	7th	9th
FP	13th	3rd	4th
CH	1st	10th	10th
MM	2nd	2nd	5th
RIF	14th	8th	7th
NAÏVE	2nd	11th	13th

Table 4.5: Rankings of methods based on threshold of £10,000 at sample size 5,000

From Table 4.5 it can be seen that three of the parametric distributions – GB2_SQRT, GG and LOGNORM – demonstrate significant potential in terms of the variability of their predictions as the three methods with the lowest standard deviations of ratios. MM performs consistently well across all three measures of performance, especially when variability is measured by the range of ratios, although the standard deviation is still among the five lowest of methods compared. From these results it is unclear which method is the best for forecasting costs greater than £10,000, since there is no outright winner over the three metrics. Some methods actually perform worse than the naïve sample-based method across all three metrics, namely FMM_LOG and FMM_SQRT (with WEIB and GAMMA worse on two of three metrics).

Whilst the results outlined previously give some indication of the methods' respective abilities to forecast high costs, we are interested in the performance of the regression methods at all points in the distribution. For this reason we carry out a similar analysis across a range of cost threshold values. To present these results, once again we plot the average ratio and the range of ratios across the replications. The results presented in

Figure 4.5 are undertaken using samples with 5,000 observations.

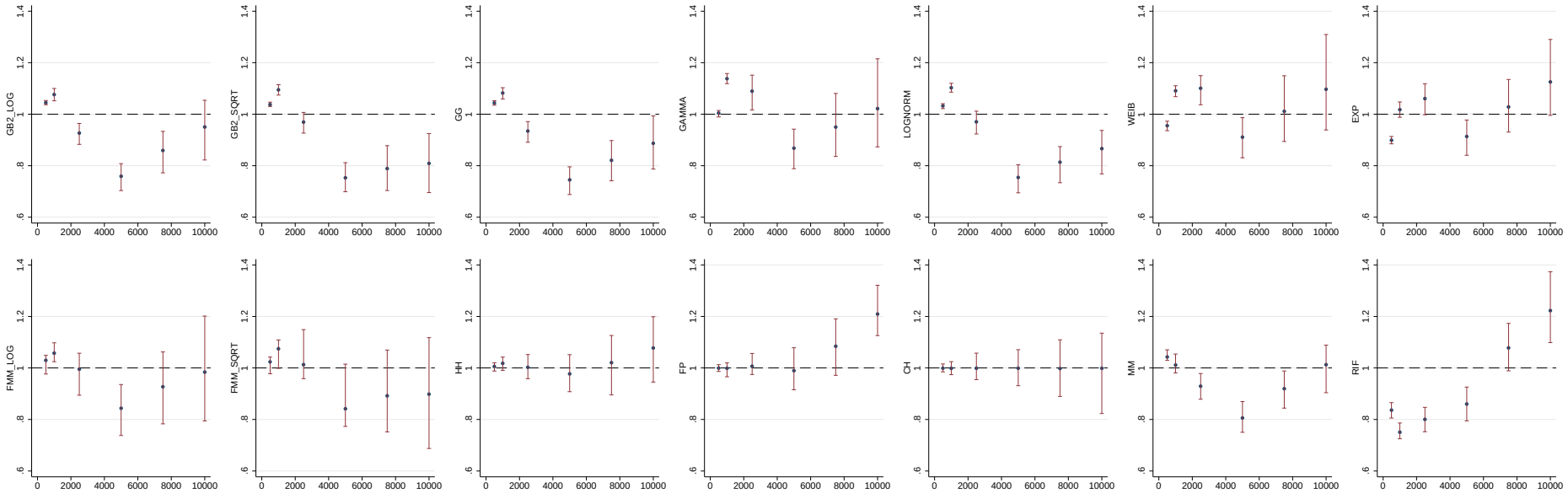


Figure 4.5: Performance of methods predicting the probability of costs exceeding various thresholds at sample size 5,000

There is a clear pattern in Figure 4.5: the higher the cost threshold being considered, the greater the variability in ratio of estimated to actual probability. Besides this, the way in which performance varies across different thresholds, including by how much variability increases with higher thresholds, is different for all methods.

Beginning with the parametric distributions, with log links, there seems to be little difference in the performance of GB2_LOG and GG, except for that GB2_LOG performs slightly better at the higher costs considered in terms of bias. Looking at the gamma-type models, LOGNORM demonstrates potential in terms of producing precise estimates of tail probabilities if not in terms of bias. Since FMM_LOG represents a two-component version of GAMMA, comparing the performance of these methods provides some insight into the returns from using more complex mixture specifications. The pattern of performance at different thresholds is quite similar for these, and the main difference seems to be that FMM_LOG produces more variable estimates, especially at low cost thresholds. WEIB and EXP seem to perform similarly, with high variability forecasts. It is interesting to note that the square-root link methods differ from their log link counterparts, particularly in terms of having worse high cost forecasts.

There is considerable variation in performance between the distributional methods. The methods that use the cumulative distribution function seem to vary predominantly according to the number of intervals that are used, rather than the specification for predicting interval membership. CH is practically unbiased for all cost thresholds, illustrating the strength of this method in forecasting $P(y > k)$ for a range of values of k. As pointed out earlier, however, the variability of the forecasts across replications is wider than the majority of other methods considered in this paper. It seems therefore that much of the bias in HH and FP stems from when k_a^* and k_b^* are not close to the value of k being investigated. This is more likely to be the case with FP than with HH, since FP has fewer intervals (and is highly unlikely using CH – in our application). This is particularly clear with $k = £10,000$, since with HH and FP in this case k_b^* will often be the highest observed cost in the sample. When this occurs, the linear interpolation that we employ is likely to lead to an overestimation of the forecasted probability (see equation 4.2 for details). For these three methods the variability of ratios is roughly similar, but when looking also at the methods using the quantile function, it is clear that MM offers an improvement upon

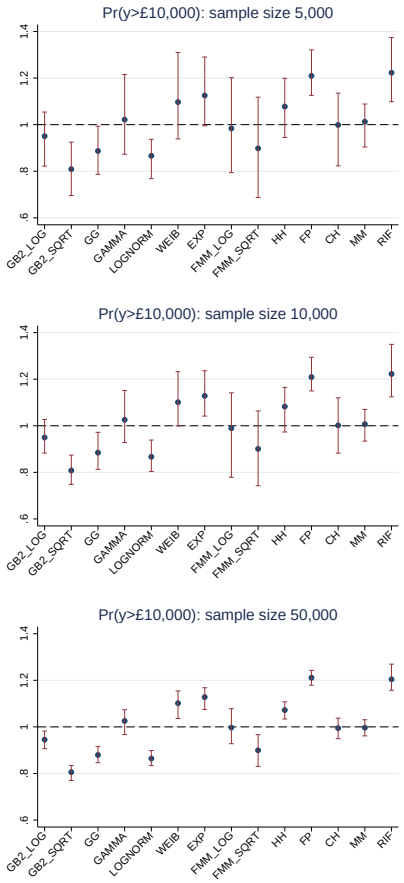


Figure 4.6: Performance of methods predicting the probability of a cost exceeding £10,000 at different sample sizes

the variability. Its performance, however, in terms of bias varies across values of k. RIF seems to perform badly both in terms of bias and precision.

Finally, we look into how our analysis is affected by the number of observations that are present in the drawn samples. To do this, we return to the style of graph used for Figure 4.4, but illustrate performances for the three sample sizes analysed ($N_s \in$ 5,000; 10,000; 50,000). The results are therefore only for one value of k, but results at other values followed a similar pattern.

From Figure 4.6 we can see that there is a clear effect of sample size on the performance of the regression methods fitting the whole distribution. Having more observations does not particularly affect the bias of each method, but, as expected, it reduces the variability of the estimates. This therefore means that methods such as CH perform relatively better at bigger sample sizes since they remain unbiased, but forecast costs with increased

precision.

4.5 Discussion

The results of this paper are the first to provide a comparative assessment of parametric and distributional methods designed to estimate a counterfactual distribution. This makes them different to most studies concerning econometric modelling of healthcare costs where performance has largely been judged on the basis of the ability to predict conditional means. Jones et al. (2014) compare parametric distributions (but not distributional methods) against one another for predicting tail probabilities as well as in-sample fit of the whole distribution based on log-likelihood statistics. The analysis presented here builds on this work with a range of thresholds for tail probabilities as well as a broader range of parametric distributions including mixture distributions and models with a square-root link as well as those with a log link.

As mentioned in the methodology section of the paper, some of these methods have been automated in order to make the quasi-Monte Carlo study design feasible. For instance, we only allow location parameters to vary with covariates and we restrict the number of mixtures used in FMM_LOG and FMM_SQRT. In practice, analysts are likely to train their model for a given sample – testing the appropriateness of covariates in the specification as well as the number of mixtures that are required etc. Since all methods have been restricted to some degree, e.g. the regressors are the same for all methods, the results of this paper give some indication of the relative performance of these methods and illustrate their pitfalls and strengths.

For our application, CH demonstrates potential even for forecasting probabilities of high costs – such as costs that exceed £10,000. A function of the methodology is that CH (as well as HH and FP) is unable to extrapolate beyond the observed sample, and so in applications where sample size is small, or if the decision-maker is interested in the probability of extremely high costs beyond the largest observed, this method would be unable to provide any information on this parameter. This represents a limitation for this type of method for fitting the distribution of healthcare costs, where the underlying data generating process is heavy-tailed, and any observed sample is unlikely to contain some

of the potential extreme outcomes. This could be overcome by applying some smoothing techniques and moving beyond the non-smooth methodology adopted in this paper.

There is considerable variation in the best performing parametric distributions according to the specific tail probability being considered. When considering costs that exceed £10,000, FMM_LOG is the least biased parametric method, but is the most imprecise of all methods considered. At other thresholds, the distribution with the best fit on average varies: for example WEIB performs best among parametric distributions for costs that exceed £7,500. This means that the preferred parametric distribution would depend upon the decision-maker's loss function. Some distributions are particularly imprecise at all tails investigated, notably the mixture models – FMM_LOG and FMM_SQRT – as well as some of the more restrictive distributions – GAMMA, WEIB and EXP. LOGNORM is the most precise and thus demonstrates its potential for modelling the whole distribution of costs. Whilst other papers have focused on the importance of the link function, which seems to have a large impact on performance when it comes to predicting mean healthcare costs (see for example Basu et al., 2006), this paper finds that when we are concerned with predicting tail probabilities the link function is less of an issue than are the distributional assumptions more generally.

The distributional methods show promise for modelling the full distribution of healthcare costs. In particular, CH is practically unbiased in terms of all forecasted tail probabilities considered. The related methods of FP and HH also perform well in terms of bias, but not when considering costs that exceed £10,000, because £10,000 is likely to fall in the highest quantile of costs in either method. CH is better placed to model this tail probability, since each unique value of costs that is encountered in the sample is used as the basis for an indicator variable for a separate regression, and using a linear probability model does not require variation across all covariates for each value of the dependent variable. At the smallest sample size of 5,000 observations, these three methods exhibit highly imprecise forecasted probabilities, but this becomes less of an issue at larger sample sizes where the variability is lower for all 14 methods. MM delivers better precision, but its performance on average varies across the different tail probabilities. RIF appears to be the worst among the distributional methods for this dataset and specification.[28]

[28]Our results are in line with results from a simulation comparing quantile and distribution regression

Acknowledgements

The authors gratefully acknowledge funding from the Economic and Social Research Council (ESRC) under grant reference RES-060-25-0045. We would also like to thank members of the Health, Econometrics and Data Group (HEDG), at the University of York, and John Mullahy for useful discussions. In addition, we are grateful to Blaise Melly for extensive comments on this work. Presentations of this paper at the European Workshop on Econometrics and Health Economics at the University of Munich and Max Planck Institute for Social Law and Social Policy and the Annual Health Econometrics Workshop at the University of Toronto have both been extremely useful for refining this paper. In particular, we would like to thank Anirban Basu, Fabrice Etilé, Owen O'Donnell, Bill Greene, Frank Windmeijer and Jeffrey Racine for their suggestions and support.

methods conducted in supplemental material of Chernozhukov et al. (2013), which show that quantile regression methods perform worse when there is a non-continuous outcome variable such as ours (given the observed number of mass-points).

References

Andrews DW. 1988. Chi-square diagnostic tests for econometric models: Introduction and applications. *Journal of Econometrics* **37**: 135 – 156.

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. 2009. Issues for the next generation of health care cost analyses. *Medical Care* **47**: S109–S114.

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

Bilger M, Manning WG. 2014. Measuring overfitting in nonlinear models: A new method and an application to health expenditures. *Health Economics* In Press.

Bitler MP, Gelbach JB, Hoynes HW. 2006. What mean impacts miss: Distributional effects of welfare reform experiments. *The American Economic Review* **96**: 988–1012.

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

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.

Cawley J, Meyerhoefer C. 2012. The medical care costs of obesity: An instrumental variables approach. *Journal of Health Economics* **31**: 219 – 230.

Chernozhukov V, Fernndez-Val I, Melly B. 2013. Inference on counterfactual distributions. *Econometrica* **81**: 2205–2268.

Cook BL, Manning WG. 2009. Measuring racial/ethnic disparities across the distribution of health care expenditures. *Health Services Research* **44**: 1603–1621.

Cox C. 2008. The generalized F distribution: An umbrella for parametric survival analysis. *Statistics in Medicine* **27**: 4301–4312.

Cox DR. 1972. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)* **34**: 187–220.

de Meijer C, ODonnell O, Koopmanschap M, van Doorslaer E. 2013. Health expenditure growth: Looking beyond the average through decomposition of the full distribution. *Journal of Health Economics* **32**: 88 – 105.

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

Deb P, Trivedi PK. 1997. Demand for medical care by the elderly: A finite mixture approach. *Journal of Applied Econometrics* **12**: 313–336.

Dixon J, Smith P, Gravelle H, Martin S, Bardsley M, Rice N, Georghiou T, Dusheiko M, Billings J, Lorenzo MD, Sanderson C. 2011. A person based formula for allocating commissioning funds to general practices in england: development of a statistical model. *BMJ* **343**: d6608.

Duan N, Manning WG, Morris CN, Newhouse JP. 1983. A comparison of alternative models for the demand for medical care. *Journal of Business & Economic Statistics* **1**: 115–126.

Firpo S, Fortin NM, Lemieux T. 2009. Unconditional quantile regressions. *Econometrica* **77**: 953–973.

Foresi S, Peracchi F. 1995. The conditional distribution of excess returns: An empirical analysis. *Journal of the American Statistical Association* **90**: 451–466.

Fortin N, Lemieux T, Firpo S. 2011. Decomposition methods in economics. volume 4, Part A of *Handbook of Labor Economics*. Elsevier, 1 – 102.

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

Han A, Hausman JA. 1990. Flexible parametric estimation of duration and competing risk models. *Journal of Applied Econometrics* **5**: 1–28.

Heckman JJ. 2001. Micro data, heterogeneity, and the evaluation of public policy: Nobel lecture. *Journal of Political Economy* **109**: 673–748.

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.

Hoch JS, Briggs AH, Willan AR. 2002. Something old, something new, something borrowed, something blue: a framework for the marriage of health econometrics and cost-effectiveness analysis. *Health Economics* **11**: 415–430.

Holly A. 2009. Modeling risk using fourth order pseudo maximum likelihood methods. Institute of Health Economics and Management (IEMS), University of Lausanne, Switzerland.

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* **S456823**. Boston College Department of Economics.

Johnson E, Dominici F, Griswold M, L Zeger S. 2003. Disease cases and their medical costs attributable to smoking: an analysis of the national medical expenditure survey. *Journal of Econometrics* **112**: 135–151.

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

Jones AM, Lomas J, Moore P, Rice N. 2013. A quasi-Monte Carlo comparison of developments in parametric and semi-parametric regression methods for heavy tailed and non-normal data: with an application to healthcare costs. *Health Econometrics and Data Group Working Paper* **13/30**.

Jones AM, Lomas J, Rice N. 2014. Applying beta-type size distributions to healthcare cost regressions. *Journal of Applied Econometrics* **29**: 649–670.

Machado JAF, Mata J. 2005. Counterfactual decomposition of changes in wage distributions using quantile regression. *Journal of Applied Econometrics* **20**: 445–465.

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

Manning WG, Duan N, Rogers W. 1987. Monte Carlo evidence on the choice between sample selection and two-part models. *Journal of Econometrics* **35**: 59 – 82.

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

McDonald JB, Sorensen J, Turley PA. 2013. Skewness and kurtosis properties of income distribution models. *Review of Income and Wealth* **59**: 360–374.

Melly B. 2005. Decomposition of differences in distribution using quantile regression. *Labour Economics* **12**: 577 – 590.

Mora T, Gil J, Sicras-Mainar A. 2014. The influence of obesity and overweight on medical costs: a panel data perspective. *European Journal of Health Economics* In press.

Mullahy J. 2009. Econometric modeling of health care costs and expenditures: a survey of analytical issues and related policy considerations. *Medical Care* **47**: S104–S108.

Pentsak Y. 2007. Addressing skewness and kurtosis in health care econometrics. PhD Thesis, University of Lausanne.

Van de Ven WP, Ellis RP. 2000. Risk adjustment in competitive health plan markets. In Culyer AJ, Newhouse JP (eds.) *Handbook of Health Economics*, volume 1 of *Handbook of Health Economics*, chapter 14. Elsevier, 755–845.

Vanness DJ, Mullahy J. 2007. Perspectives on mean-based evaluation of health care. In Jones AM (ed.) *The Elgar Companion to Health Economics*. Elgar Original Reference.

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

Walls WD. 2005. Modelling heavy tails and skewness in film returns. *Applied Financial Economics* **15**: 1181–1188.

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

4.6 Appendix A

We use the variables shown in Table 4A1 to construct our regression models. They are based on the ICD10 chapters, which are given in Table 4A2.

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 4A1: Classification of morbidity characteristics

ICD10 codes beginning with U were dropped because there were no observations in the 6,164,114 used. Only a small number (3,170) 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 4A2: ICD10 chapter codes

4.7 Appendix B

	Cost threshold ($£k$)					
Method	**500**	**1000**	**2500**	**5000**	**7500**	**10000**
GB2_LOG	1.045	1.076	0.927	0.759	0.859	0.950
GB2_SQRT	1.038	1.095	0.969	0.753	0.789	0.809
GG	1.044	1.082	0.935	0.745	0.821	0.887
GAMMA	1.004	1.138	1.089	0.868	0.950	1.022
LOGNORM	1.032	1.103	0.970	0.754	0.814	0.866
WEIB	0.955	1.091	1.101	0.911	1.011	1.097
EXP	0.899	1.018	1.061	0.913	1.028	1.125
FMM_LOG	1.030	1.058	0.995	0.844	0.927	0.984
FMM_SQRT	1.024	1.074	1.013	0.842	0.892	0.898
HH	1.006	1.018	1.003	0.977	1.021	1.078
FP	1.000	0.999	1.007	0.990	1.084	1.209
CH	0.999	0.999	0.999	0.999	0.998	0.999
MM	1.043	1.011	0.929	0.806	0.920	1.013
RIF	0.836	0.751	0.800	0.860	1.078	1.223

Table 4B1: Mean ratios of predicted to actual survival probabilities, sample size 5,000

	Cost threshold ($£k$)					
Method	**500**	**1000**	**2500**	**5000**	**7500**	**10000**
GB2_LOG	0.016	0.048	0.081	0.104	0.161	0.232
GB2_SQRT	0.017	0.040	0.080	0.113	0.174	0.229
GG	0.017	0.044	0.080	0.108	0.156	0.206
GAMMA	0.025	0.039	0.135	0.153	0.245	0.342
LOGNORM	0.019	0.034	0.089	0.109	0.140	0.169
WEIB	0.037	0.042	0.113	0.156	0.255	0.371
EXP	0.028	0.060	0.121	0.137	0.204	0.295
FMM_LOG	0.072	0.073	0.163	0.198	0.279	0.407
FMM_SQRT	0.065	0.110	0.191	0.242	0.318	0.431
HH	0.032	0.051	0.094	0.144	0.231	0.254
FP	0.026	0.054	0.082	0.163	0.219	0.195
CH	0.030	0.050	0.103	0.140	0.220	0.312
MM	0.041	0.073	0.100	0.119	0.144	0.184
RIF	0.060	0.061	0.095	0.131	0.184	0.275

Table 4B2: Range of ratios of predicted to actual survival probabilities, sample size 5,000

	Cost threshold ($£k$)					
Method	**500**	**1000**	**2500**	**5000**	**7500**	**10000**
GB2_LOG	0.003	0.008	0.015	0.021	0.034	0.047
GB2_SQRT	0.003	0.007	0.015	0.021	0.031	0.041
GG	0.003	0.008	0.015	0.021	0.031	0.040
GAMMA	0.005	0.008	0.027	0.034	0.047	0.061
LOGNORM	0.003	0.007	0.016	0.021	0.030	0.039
WEIB	0.009	0.009	0.022	0.034	0.049	0.065
EXP	0.006	0.012	0.024	0.029	0.040	0.053
FMM_LOG	0.016	0.016	0.029	0.042	0.071	0.095
FMM_SQRT	0.018	0.020	0.035	0.036	0.056	0.089
HH	0.007	0.010	0.021	0.029	0.049	0.057
FP	0.005	0.011	0.017	0.035	0.045	0.045
CH	0.006	0.011	0.019	0.030	0.042	0.060
MM	0.006	0.012	0.019	0.024	0.034	0.045
RIF	0.012	0.014	0.022	0.028	0.040	0.053

Table 4B3: Standard deviation of ratios of predicted to actual survival probabilities, sample size 5,000

	Cost threshold ($£k$)					
Method	**500**	**1000**	**2500**	**5000**	**7500**	**10000**
GB2_LOG	1.045	1.077	0.928	0.759	0.859	0.950
GB2_SQRT	1.037	1.095	0.970	0.753	0.789	0.808
GG	1.043	1.083	0.936	0.745	0.820	0.885
GAMMA	1.004	1.138	1.092	0.871	0.954	1.026
LOGNORM	1.032	1.103	0.971	0.755	0.814	0.867
WEIB	0.953	1.088	1.102	0.914	1.015	1.101
EXP	0.900	1.019	1.063	0.916	1.031	1.128
FMM_LOG	1.034	1.055	0.988	0.845	0.931	0.989
FMM_SQRT	1.028	1.076	1.002	0.835	0.890	0.901
HH	1.006	1.018	1.001	0.978	1.020	1.083
FP	0.999	0.999	1.004	0.988	1.083	1.209
CH	0.999	0.999	0.999	1.000	0.997	1.002
MM	1.043	1.010	0.929	0.804	0.915	1.007
RIF	0.836	0.747	0.800	0.862	1.080	1.222

Table 4B4: Mean ratios of predicted to actual survival probabilities, sample size 10,000

	Cost threshold ($£k$)					
Method	**500**	**1000**	**2500**	**5000**	**7500**	**10000**
GB2_LOG	0.012	0.031	0.059	0.075	0.111	0.144
GB2_SQRT	0.010	0.028	0.058	0.070	0.093	0.126
GG	0.011	0.026	0.055	0.081	0.122	0.159
GAMMA	0.019	0.027	0.077	0.105	0.166	0.224
LOGNORM	0.011	0.029	0.062	0.075	0.107	0.135
WEIB	0.036	0.039	0.068	0.101	0.166	0.233
EXP	0.016	0.034	0.070	0.088	0.140	0.195
FMM_LOG	0.054	0.050	0.126	0.149	0.265	0.363
FMM_SQRT	0.073	0.096	0.112	0.135	0.213	0.321
HH	0.022	0.038	0.073	0.102	0.161	0.191
FP	0.020	0.036	0.060	0.125	0.145	0.144
CH	0.020	0.035	0.064	0.094	0.138	0.238
MM	0.019	0.052	0.076	0.074	0.100	0.136
RIF	0.043	0.062	0.104	0.103	0.158	0.225

Table 4B5: Range of ratios of predicted to actual survival probabilities, sample size 10,000

	Cost threshold ($£k$)					
Method	**500**	**1000**	**2500**	**5000**	**7500**	**10000**
GB2_LOG	0.002	0.006	0.011	0.014	0.022	0.030
GB2_SQRT	0.002	0.005	0.011	0.015	0.021	0.027
GG	0.002	0.005	0.010	0.014	0.022	0.029
GAMMA	0.004	0.006	0.017	0.023	0.033	0.044
LOGNORM	0.002	0.005	0.011	0.014	0.021	0.027
WEIB	0.006	0.007	0.014	0.023	0.035	0.047
EXP	0.004	0.008	0.016	0.020	0.028	0.038
FMM_LOG	0.013	0.010	0.021	0.027	0.050	0.069
FMM_SQRT	0.013	0.015	0.020	0.024	0.042	0.064
HH	0.005	0.007	0.015	0.022	0.035	0.042
FP	0.004	0.008	0.011	0.026	0.032	0.028
CH	0.004	0.008	0.012	0.021	0.032	0.046
MM	0.004	0.009	0.015	0.015	0.021	0.030
RIF	0.009	0.012	0.017	0.020	0.032	0.043

Table 4B6: Standard deviation of ratios of predicted to actual survival probabilities, sample size 10,000

	Cost threshold ($£k$)					
Method	**500**	**1000**	**2500**	**5000**	**7500**	**10000**
GB2_LOG	1.045	1.078	0.928	0.758	0.857	0.945
GB2_SQRT	1.037	1.096	0.970	0.753	0.787	0.806
GG	1.043	1.084	0.937	0.744	0.817	0.879
GAMMA	1.004	1.139	1.092	0.871	0.954	1.025
LOGNORM	1.032	1.103	0.971	0.754	0.813	0.864
WEIB	0.951	1.086	1.101	0.914	1.015	1.102
EXP	0.900	1.020	1.063	0.915	1.031	1.128
FMM_LOG	1.038	1.053	0.981	0.845	0.935	0.997
FMM_SQRT	1.033	1.079	0.996	0.828	0.885	0.899
HH	1.004	1.017	0.998	0.984	1.011	1.072
FP	0.999	1.001	1.004	0.985	1.076	1.211
CH	1.000	1.000	0.999	0.999	0.994	0.995
MM	1.043	1.010	0.929	0.803	0.908	0.997
RIF	0.834	0.745	0.803	0.861	1.072	1.204

Table 4B7: Mean ratios of predicted to actual survival probabilities, sample size 50,000

	Cost threshold ($£k$)					
Method	**500**	**1000**	**2500**	**5000**	**7500**	**10000**
GB2_LOG	0.006	0.013	0.029	0.034	0.055	0.075
GB2_SQRT	0.006	0.011	0.030	0.037	0.052	0.064
GG	0.006	0.012	0.028	0.037	0.053	0.070
GAMMA	0.010	0.012	0.045	0.064	0.087	0.107
LOGNORM	0.005	0.011	0.028	0.038	0.053	0.065
WEIB	0.015	0.017	0.034	0.064	0.093	0.119
EXP	0.009	0.018	0.041	0.055	0.075	0.093
FMM_LOG	0.024	0.019	0.082	0.099	0.114	0.150
FMM_SQRT	0.008	0.016	0.034	0.059	0.101	0.136
HH	0.011	0.022	0.028	0.040	0.060	0.074
FP	0.011	0.021	0.026	0.053	0.075	0.063
CH	0.010	0.016	0.026	0.041	0.079	0.088
MM	0.011	0.024	0.038	0.036	0.044	0.069
RIF	0.019	0.025	0.038	0.049	0.080	0.112

Table 4B8: Range of ratios of predicted to actual survival probabilities, sample size 50,000

	Cost threshold ($£k$)					
Method	**500**	**1000**	**2500**	**5000**	**7500**	**10000**
GB2_LOG	0.001	0.003	0.005	0.007	0.010	0.014
GB2_SQRT	0.001	0.003	0.006	0.007	0.010	0.013
GG	0.001	0.003	0.005	0.007	0.010	0.013
GAMMA	0.002	0.002	0.008	0.010	0.015	0.020
LOGNORM	0.001	0.002	0.005	0.007	0.010	0.013
WEIB	0.003	0.003	0.006	0.011	0.016	0.022
EXP	0.002	0.003	0.007	0.009	0.013	0.017
FMM_LOG	0.002	0.004	0.009	0.013	0.019	0.025
FMM_SQRT	0.001	0.003	0.007	0.010	0.016	0.022
HH	0.002	0.004	0.006	0.008	0.012	0.017
FP	0.002	0.005	0.005	0.012	0.014	0.011
CH	0.002	0.004	0.006	0.009	0.015	0.019
MM	0.002	0.004	0.007	0.007	0.010	0.013
RIF	0.004	0.005	0.009	0.010	0.017	0.022

Table 4B9: Standard deviation of ratios of predicted to actual survival probabilities, sample size 50,000

Chapter 5

Conclusions

The common theme of the chapters in this thesis is the comparative empirical assessment of econometric methods for modelling the distribution of healthcare costs.

In chapter 2, the generalised beta of the second kind distribution (GB2) is shown to be a flexible parametric distribution which features prominent distributions in the modelling of healthcare costs as special cases. Importantly, it allows for greater flexibility in estimating skewness and kurtosis than the generalised gamma distribution and so builds upon the work of Manning et al. (2005). Whilst these properties are already known in the statistics literature, chapter 2 provides a rigorous assessment of the extent to which these lead to better ability to forecast healthcare costs using real data. When considering flexible models, a major concern is overfitting the data. Chapter 2 adopts two approaches to taking account of this possibility. First, tests based on the 'estimation' sample build in some penalty for additional parameters. And second, most of the results are obtained using data from the 'validation' set, which was not used in the estimation of models. With tests based on the 'estimation' sample, the generalised beta of the second kind is found to provide the best fit of the distribution according to the Akaike Information Criterion (AIC). With a larger penalty for additional parameters, as imposed using the Bayesian Information Criterion (BIC), the nested beta of the second kind distribution is the best. The results based on out-of-sample forecasts of conditional means and tail probabilities (for very high costs) suggest that the additional flexibility provided by GB2 is not helpful in modelling these parameters. The special cases of beta of the second kind, generalised gamma and log-normal distributions perform the best of models compared when looking

at bias, goodness-of-fit and accuracy of forecasted conditional means, respectively, and the generalised gamma and log-normal distributions perform best in terms of the forecasted tail probabilities considered.

The results from chapter 2 also open up new research questions that are addressed in later chapters of the thesis. Firstly, while chapter 2 investigates the performance of GB2 relative to other parametric distributions, it does not compare GB2 to commonly used methods in this area. Secondly, it is shown that GB2 performs best in fitting the whole distribution of healthcare costs (according to AIC measure), but that this is not borne out in performance in terms of fitting either the conditional mean, or the probability of very high costs. The first of these is addressed by chapter 3, and the second by chapter 4.

In addition, there is scope for further research based on chapter 2. When considering multiple-parameter functional forms of the distribution, it is not clear which parameters should be allowed to vary with covariates (and the covariates that should be included). In chapter 2, the GB2 was parameterised with its scale parameter, b, as an exponential function of covariates, but future research may allow one or more of its shape parameters: a, p and q to also be specified as functions of covariates. In Manning et al. (2005), the authors allow for both its μ and σ parameters to be functions of covariates, for example. Furthermore, there are other multiple-parameter distributions that could be used as alternatives to GB2 for modelling healthcare costs, for example Holly and Pentsak (2006) propose the use of Pearson Type IV distribution. These competing approaches could be compared against each other, once coding issues are resolved.

As well as comparing GB2 to standard practice, chapter 3 implements a systematic comparison of developments of both a parametric and semi-parametric nature against both each other and commonly used methods. Chapter 3 begins with a review of comparative studies, with a particular focus on work using real, as opposed to synthesised, data. From this it is clear that there is a need for a thorough comprehensive study, which chapter 3 provides. It also provides details for implementing recently developed econometric approaches, including the conditional density approximation estimator. In this chapter, all evaluation is based on forecasting the conditional mean of the distribution. The results indicate that the link function plays a particularly important role. The methods with square root link functions perform best with the chosen specification of covariates and data. Ad-

ditional flexibility in the functional form of the whole distribution improves the accuracy of forecasts, with a two component gamma model and GB2 performing among the top four. The results clearly show that the conditional density approximation estimator is a promising method for the econometric analysis of healthcare costs.

It is interesting to note that the conditional density approximation estimator performs so well in terms of forecasting conditional means. This method essentially approximates the whole distribution by dividing the outcome into discrete intervals and modelling the probability of an outcome taking a value within each interval. As such, an important research question revolves around the accuracy of the approximation of the entire distribution. Since this approach is estimated using an approach inspired by Han and Hausman (1990), this question is addressed in chapter 4, where numerous distributional regression methods, including the Han and Hausman (1990) approach, are compared against parametric methods for fitting the distribution.

The main result of chapter 3 is the illustration of the sensitivity of results to the choice of econometric method. However it is not clear which specific aspects of the data generating process of healthcare costs drives this result. As such, there is the potential for illuminating research using traditional Monte Carlo studies, where the performance of each method can be evaluated under controlled circumstances by individually changing certain assumptions about the data generating process. Another result of chapter 3 is that the Pearson correlation coefficient test performs well in discriminating between methods (particularly in terms of the bias of forecasts). A number of model misspecification tests and model selection algorithms have been suggested for econometric modelling of healthcare costs – see *inter alia* Manning and Mullahy (2001); Gilleskie and Mroz (2004); Basu et al. (2006) – and research into the effectiveness of these methods in a real data context could be highly informative to applied researchers.

Chapter 4 seeks to directly address the research question regarding the appropriateness of econometric methods for fitting the whole distribution of healthcare costs. Many of the methods proposed in modelling healthcare costs are motivated by this underlying objective, however comparative empirical work has evaluated performance based solely on the conditional mean of the distribution. The chapter begins by reviewing econometric methods which can be used to produce an estimate of the conditional distribution. These

include approaches used before with healthcare costs, as well as approaches used in labour and financial economics. Performance in chapter 4 is based entirely upon forecasting tail probabilities, which, by considering various tails, provides an assessment of the fit of the whole distribution (and has precedent in providing the basis for the Andrews (1988) test). The chapter's results are highly informative, since little empirical work has been carried out on methods for estimating tail probabilities for healthcare costs. There is a trade-off between how well methods perform on average (bias), and the variability of performance across replications (precision). More flexible parametric methods are found to produce precise results, but they are biased (the extent to which depends upon the tail being considered). Additional flexibility does not appear to necessarily increase precision, with the most precise method being the two-parameter log-normal distribution. On the other hand, distributional regressions, in particular Chernozhukov et al. (2013), produce unbiased results with a lack of precision. The chapter concludes that at large sample sizes, distributional regressions may be the best approach for this type of application, since the precision of all methods improves with increasing sample size. One important caveat is provided: that distributional regressions have less ability to extrapolate beyond observed values of the outcome in the sample compared to parametric approaches. Such extrapolation outside-of-support could be conducted with the adoption of some sort of smoothing technique, while the implementation here is non-smooth. A result of this might be worse performance at the threshold values considered here and as such a trade-off may be encountered. Further research into smoothing and non-parametric methods - more generally - would be very interesting in this field.

The results from chapter 4 concerning parametric distributions can be combined with some of the results from chapter 2. It is interesting to note that distributions that performed well in forecasting tail probabilities for very high costs in chapter 2 – log-normal and generalised gamma – also perform well for tail probabilities throughout the distribution in chapter 4. As such, the additional flexibility provided by GB2 and the impressive performance based on AIC in chapter 2 is not borne out in either of the chapter's tail probability evaluations. In contrast to the finding surrounding the importance of the link function in chapter 3, the link function is found to have little impact when considering tail probabilities as opposed to the conditional mean.

While it is clear that the mean is generally not the only informative feature of the distribution of healthcare costs, further research is required on the loss-functions of policymakers in order to direct future methodological studies. Considering cost-effectiveness analysis as a potential application of these methods, characterising uncertainty for probabilistic sensitivity analysis requires modelling of the entire distribution. In addition, while decisions in health technology assessment are currently based on expected net health benefit, there are arguments for altering the decision rule in order to account for societal risk aversion if considering spending on health technologies as a portfolio of risky investments (Zivin, 2001; Bridges, 2004). Given this perspective, other features of the distribution of healthcare costs are relevant, including higher order moments (Elbasha, 2005).

Throughout the work contained in the thesis, a number of simplifying assumptions are made in order to facilitate the automated quasi-Monte Carlo study design. This is necessary given that models are estimated on hundreds of different samples. To some extent, simplification is common to all types of methods, since the specification of regressors was fixed. However, the more complex methods were simplified in other ways too. For example, when discretising the outcome variable before estimating the conditional density approximation estimator, Gilleskie and Mroz (2004) propose an algorithm for deciding upon the optimal number of bins and bin widths, which was infeasible to implement as part of the study design adopted throughout this thesis. By fixing these elements, some of the 'semi-parametric' methods could be better considered as extensions to parametric methods, rather than being truly semi-parametric. As mentioned previously, similar restrictions were placed on flexible parametric distributions by allowing only one parameter to vary with covariates. As a result there is scope for a great deal of research into model selection algorithms. Chapter 2 makes some progress towards this by considering tests of restrictions on GB2 for its special cases, where GB2 can be thought of as an umbrella distribution (Cox, 2008). Similar work could consider the number of components in mixtures and the components considered. There is considerable debate about mixture models, as to whether few-but-complex (Villani et al., 2009) components are preferable to many simple components, although it is not clear which approach is better suited to modelling healthcare costs. In the related field of modelling health-related quality of life, it is common to use more than two components when applying mixture models (see *inter alia* Austin and

Escobar, 2003 and Hernández Alava et al., 2012).

Finally, this thesis considers the specification of the functional form of the distribution in regression methods, applied to a dataset with information from a small period of time (2007/2008 financial year). There are numerous methodological strides possible if patients are followed over time, and many interesting research questions that would follow with such data. In addition, methods such as matching could be used in place of, or in conjunction with, regression methods (Kreif et al., 2012). In this thesis, potential problems with endogeneity, which are important in many applications, are not considered and research is ongoing in this area too (Garrido et al., 2012). Often there are multiple interdependent outcomes that need to be considered by policymakers, and future work is required in this area. There are a whole host of econometric challenges that require future methodological research.

Bibliography

Andrews DW. 1988. Chi-square diagnostic tests for econometric models: Introduction and applications. *Journal of Econometrics* **37**: 135 – 156.

Appleby J. 2013. Spending on health and social care over the next 50 years.

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

Austin P, Escobar M. 2003. The use of finite mixture models to estimate the distribution of the health utilities index in the presence of a ceiling effect. *Journal of Applied Statistics* **30**: 909–923.

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. 2009. Issues for the next generation of health care cost analyses. *Medical Care* **47**: S109–S114.

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

Basu A, Rathouz PJ. 2005. Estimating marginal and incremental effects on health outcomes using flexible link and variance function models. *Biostatistics* **6**: 93–109.

Bilger M, Manning WG. 2014. Measuring overfitting in nonlinear models: A new method and an application to health expenditures. *Health Economics* In Press.

Bitler MP, Gelbach JB, Hoynes HW. 2006. What mean impacts miss: Distributional effects of welfare reform experiments. *The American Economic Review* **96**: 988–1012.

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

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.

Bridges J. 2004. Understanding the risks associated with resource allocation decisions in health: An illustration of the importance of portfolio theory. *Health, Risk & Society* **6**: 257–275.

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.

Cawley J, Meyerhoefer C. 2012. The medical care costs of obesity: An instrumental variables approach. *Journal of Health Economics* **31**: 219 – 230.

Chernozhukov V, Fernndez-Val I, Melly B. 2013. Inference on counterfactual distributions. *Econometrica* **81**: 2205–2268.

Cook BL, Manning WG. 2009. Measuring racial/ethnic disparities across the distribution of health care expenditures. *Health Services Research* **44**: 1603–1621.

Copas JB. 1983. Regression, prediction and shrinkage. *Journal of the Royal Statistical Society. Series B (Methodological)* **45**: pp. 311–354.

Cox C. 2008. The generalized F distribution: An umbrella for parametric survival analysis. *Statistics in Medicine* **27**: 4301–4312.

Cox DR. 1972. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)* **34**: 187–220.

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.

de Meijer C, ODonnell O, Koopmanschap M, van Doorslaer E. 2013. Health expenditure growth: Looking beyond the average through decomposition of the full distribution. *Journal of Health Economics* **32**: 88 – 105.

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

Deb P, Trivedi PK. 1997. Demand for medical care by the elderly: A finite mixture approach. *Journal of Applied Econometrics* **12**: 313–336.

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, Smith P, Gravelle H, Martin S, Bardsley M, Rice N, Georghiou T, Dusheiko M, Billings J, Lorenzo MD, Sanderson C. 2011. A person based formula for allocating commissioning funds to general practices in england: development of a statistical model. *BMJ* **343**: d6608.

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

Duan N, Manning WG, Morris CN, Newhouse JP. 1983. A comparison of alternative models for the demand for medical care. *Journal of Business & Economic Statistics* **1**: 115–126.

Elbasha EH. 2005. Risk aversion and uncertainty in cost-effectiveness analysis: the expected-utility, moment-generating function approach. *Health Economics* **14**: 457–470.

Firpo S, Fortin NM, Lemieux T. 2009. Unconditional quantile regressions. *Econometrica* **77**: 953–973.

Foresi S, Peracchi F. 1995. The conditional distribution of excess returns: An empirical analysis. *Journal of the American Statistical Association* **90**: 451–466.

Fortin N, Lemieux T, Firpo S. 2011. Decomposition methods in economics. volume 4, Part A of *Handbook of Labor Economics*. Elsevier, 1 – 102.

Garrido MM, Deb P, Burgess JF, Penrod JD. 2012. Choosing models for health care cost analyses: Issues of nonlinearity and endogeneity. *Health Services Research* **47**: 2377–2397.

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

Han A, Hausman JA. 1990. Flexible parametric estimation of duration and competing risk models. *Journal of Applied Econometrics* **5**: 1–28.

Heckman JJ. 2001. Micro data, heterogeneity, and the evaluation of public policy: Nobel lecture. *Journal of Political Economy* **109**: 673–748.

Hernández Alava M, Wailoo AJ, Ara R. 2012. Tails from the peak district: Adjusted limited dependent variable mixture models of EQ-5D questionnaire health state utility values. *Value in Health* **15**: 550 – 561.

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.

Hoch JS, Briggs AH, Willan AR. 2002. Something old, something new, something borrowed, something blue: a framework for the marriage of health econometrics and cost-effectiveness analysis. *Health Economics* **11**: 415–430.

Holly A. 2009. Modeling risk using fourth order pseudo maximum likelihood methods. Institute of Health Economics and Management (IEMS), University of Lausanne, Switzerland.

Holly A, Monfort A, Rockinger M. 2011. Fourth order pseudo maximum likelihood methods. *Journal of Econometrics* **162**: 278–293.

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.

Hosmer DW, Lemeshow S. 1980. Goodness of fit tests for the multiple logistic regression model. *Communications in Statistics - Theory and Methods* **9**: 1043.

Huber M, Lechner M, Wunsch C. 2013. The performance of estimators based on the propensity score. *Journal of Econometrics* **175**: 1–21.

Jenkins S. 2009. GB2FIT: stata module to fit generalized beta of the second kind distribution by maximum likelihood. *Statistical software components* **S456823**. Boston College Department of Economics.

Johnson E, Dominici F, Griswold M, L Zeger S. 2003. Disease cases and their medical costs attributable to smoking: an analysis of the national medical expenditure survey. *Journal of Econometrics* **112**: 135–151.

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.

Jones AM, Lomas J, Moore P, Rice N. 2013. A quasi-Monte Carlo comparison of developments in parametric and semi-parametric regression methods for heavy tailed and non-normal data: with an application to healthcare costs. *Health Econometrics and Data Group Working Paper* **13/30**.

Jones AM, Lomas J, Rice N. 2014. Applying beta-type size distributions to healthcare cost regressions. *Journal of Applied Econometrics* **29**: 649–670.

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

Kreif N, Grieve R, Radice R, Sadique Z, Ramsahai R, Sekhon JS. 2012. Methods for estimating subgroup effects in cost-effectiveness analyses that use observational data. *Medical Decision Making* **32**: 750–763.

Machado JAF, Mata J. 2005. Counterfactual decomposition of changes in wage distributions using quantile regression. *Journal of Applied Econometrics* **20**: 445–465.

Mandelbrot B. 1963. New methods in statistical economics. *Journal of Political Economy* **71**: 421–440.

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

Manning WG, Duan N, Rogers W. 1987. Monte Carlo evidence on the choice between sample selection and two-part models. *Journal of Econometrics* **35**: 59 – 82.

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

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, Sorensen J, Turley PA. 2013. Skewness and kurtosis properties of income distribution models. *Review of Income and Wealth* **59**: 360–374.

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

Melly B. 2005. Decomposition of differences in distribution using quantile regression. *Labour Economics* **12**: 577 – 590.

Mihaylova B, Briggs A, O'Hagan A, Thompson SG. 2011. Review of statistical methods for analysing healthcare resources and costs. *Health Economics* **20**: 897–916.

Mora T, Gil J, Sicras-Mainar A. 2014. The influence of obesity and overweight on medical costs: a panel data perspective. *European Journal of Health Economics* In press.

Mullahy J. 1997. Heterogeneity, excess zeros, and the structure of count data models. *Journal of Applied Econometrics* **12**: 337–350.

Mullahy J. 2009. Econometric modeling of health care costs and expenditures: a survey of analytical issues and related policy considerations. *Medical Care* **47**: S104–S108.

Organisation for Economic Co-operation and Development. 2013. OECD health statistics 2013 - frequently requested data.

Pentsak Y. 2007. Addressing skewness and kurtosis in health care econometrics. PhD Thesis, University of Lausanne.

Pregibon D. 1980. Goodness of link tests for generalized linear models. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* **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.

Van de Ven WP, Ellis RP. 2000. Risk adjustment in competitive health plan markets. In Culyer AJ, Newhouse JP (eds.) *Handbook of Health Economics*, volume 1 of *Handbook of Health Economics*, chapter 14. Elsevier, 755–845.

Vanness DJ, Mullahy J. 2007. Perspectives on mean-based evaluation of health care. In Jones AM (ed.) *The Elgar Companion to Health Economics*. Elgar Original Reference.

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

Villani M, Kohn R, Giordani P. 2009. Regression density estimation using smooth adaptive Gaussian mixtures. *Journal of Econometrics* **153**: 155 – 173.

Vuong QH. 1989. Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica* **57**: 307–333.

Walls WD. 2005. Modelling heavy tails and skewness in film returns. *Applied Financial Economics* **15**: 1181–1188.

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

Zivin JG. 2001. Cost-effectiveness analysis with risk aversion. *Health Economics* **10**: 499–508.