HEDG Working Paper 10/01

Models For Health Care

Andrew M Jones

January 2010

york.ac.uk/res/herc/hedgwp

MODELS FOR HEALTH CARE

ANDREW M. JONES
University of York

Abstract

This chapter reviews the econometric methods that are used by health economists to model health care costs. These methods are used for prediction, projection and forecasting, in the context of risk adjustment, resource allocation, technology assessment and policy evaluation. The chapter reviews the literature on the comparative performance of the methods, especially in the context of forecasting individual health care costs, and concludes with an empirical case study.

Acknowledgements: I gratefully acknowledge funding from the Economic and Social Research Council (ESRC) under the Large Grant Scheme, reference RES-060-25-0045. I am especially grateful to Will Manning for his detailed reading and extensive comments and for advice and access to Stata code from Anirban Basu, Partha Deb, Donna Gilleskie, Edward Norton and Nigel Rice.

Contents

1. Introduction

Health care costs pose particular challenges for econometric modelling. Individual-level data on medical expenditures or costs of treatment typically feature a spike at zero and a strongly skewed distribution with a heavy right-hand tail. This non-normality stems from the fact that, due to clinical complications and comorbidities, the more severe patients may attract substantial and costly services. Relatively rare events and medical procedures might be very expensive, creating outliers in the right-hand tail of the distribution. Often, a small minority of patients are responsible for a high proportion of health care costs and mean costs are well above median costs. In econometric models of costs the error term will typically exhibit a high degree of heteroskedasticity, reflecting both the process driving costs and heterogeneity across patients[1]. The relationship between costs and covariates may not be linear and the appropriate regression specification for such data may be nonlinear.

When the cost data represent the population as a whole, rather than just the users of health care, the distribution will typically have a large mass point at zero (with costs truncated at zero). The presence of a substantial proportion of zeros in the data has typically been handled by using a two-part model (2PM), which distinguishes between a binary indicator, used to model the probability of any costs, and a conditional regression model for the positive costs. An alternative approach is to use sample selection or generalised Tobit models to deal with the zeros. The relative merits of the two approaches are discussed in Jones (2000). Binary, multinomial and count data models for health care utilisation have been reviewed elsewhere (see e.g., Jones, 2000; Jones, 2007; Jones et al., 2007; Jones, 2009). The modelling of count data for doctor visits has strong affinities with the modelling of cost data, as both have non-normal heavily skewed distributions, but this chapter focuses specifically on econometric models for non-zero health care costs

Linear regression applied to the level of costs may perform poorly, due to the high degree of skewness and excess kurtosis; OLS minimises the sum of squared residuals on the cost scale and may be sensitive to extreme observations. As a result, in applied work costs are often transformed prior to estimation. The most common transformation is the logarithm of y, although the square root is sometimes used as well. More recently the literature has moved away from linear regression towards inherently nonlinear specifications, these include generalized linear models and extensions, such as the extended estimating equations approach, as well as more semiparametric approaches, such as finite mixture and discrete conditional density estimators.

Econometric models for health care costs are used in many areas of health economics and policy evaluation. Two areas where they are used frequently are risk adjustment and cost-effectiveness analysis. Cost-effectiveness analyses tend to work with smaller

[1] For example, if total costs are generated by the sum of discrete episodes of care times the costs of those episodes and the episodes follow a count distribution such as the Poisson which is inherently heteroskedastic.

datasets and the scope for parametric modelling may be more limited (Briggs et al., 2005). In cost-effectiveness analysis, and health technology assessment in general, the emphasis is often on costs incurred over a specific episode of treatment or over a whole lifetime. This introduces the issue of right censoring of cost data and the use of survival analysis.

Risk adjustment has been adopted by health care payers who use prospective or mixed reimbursement systems, such as Medicare in the United States. It is intended to address the incentives for providers to engage in cream skimming or dumping of potential patients (Van de Ven and Ellis, 2000). Risk adjustment also plays a role in the design of formulas for equitable geographic resource allocation (see e.g., Smith et al., 2001). In both cases regression models are used to predict health care costs for individuals or groups of patients. The specification of these models depends on their intended use but they typically condition on sociodemographic information, including age and gender, diagnostic indicators and controls for comorbidities such as the Diagnostic Cost Group (DCG) system (e.g., Ash et al., 2001). In risk adjustment the emphasis is on predicting the treatment costs for particular types of patient, often with very large datasets, and these costs are typically measured over a fixed period, such as a year. Risk adjustment entails making forecasts of health care costs for individual patients or groups of patients and is the motivation for exploring these methods here. It means that the focus is on individual level, rather than aggregate, data. Individual data comes from two broad sources: social surveys, in particular health interview surveys, and routine administrative datasets.

Administrative datasets include health care provider reimbursement and claims databases, and population registers of births, deaths, cancer cases, etc. (see, e.g., Atella et al., 2006; Chalkley and Tilley, 2006; Dranove et al., 2003; Dusheiko et al., 2004; Dusheiko et al., 2006; Dusheiko et al., 2007; Farsi and Ridder, 2006; Gravelle et al., 2003; Ho, 2002; Lee and Jones, 2004; Lee and Jones, 2006; Martin et al., 2007; Propper et al., 2002; Propper et al., 2004; Propper et al., 2005; Rice et al., 2000; Seshamani and Gray, 2004). These datasets are collected for administrative purposes and may be made available to researchers. Administrative datasets will often contain millions of observations and may cover a complete population, rather than just a random sample. As such they suffer from less unit and item non-response than survey data. They tend to be less affected by reporting bias, but as they are collected routinely and on a wide scale they may be vulnerable to data input and coding errors. Administrative datasets are not designed by and for researchers, which means they may not contain all of the variables that would be of interest to researchers, and different data sources may have to be combined.

This chapter provides an outline of the methods that are typically used to model individual health care costs. It reviews the literature on the comparative performance of the methods, especially in the context of forecasting individual health care costs, and concludes with an empirical case study. Section 2 begins with linear regression on the level of costs and on transformations of costs. Section 3 moves on to nonlinear regressions that are specified in terms of an exponential conditional mean. These can be estimated as nonlinear regressions or by exploiting their affinity with count data

regression and hazard models, which can provide specifications that give additional flexibility to the distribution of costs. Many recent studies of nonlinear specifications are embedded within the generalized linear models (GLM) framework. The language of the GLM approach is commonplace in the statistics literature but is less used in econometrics and is outlined in Section 4. Recent research has seen the development of more flexible parametric and semiparametric approaches and some of the key methods are described in Section 5. Section 6 reviews evidence on the comparative performance of methods that are most commonly used to model costs and for some of the recent methodological innovations. This is reinforced in Section 7 which presents an illustrative application of the methods with data from the US Medical Expenditure Panel Study (MEPS). Section 8 suggests some further reading.

2. Linear regression models

2.1 Cost regressions

Linear regression on the level of costs (*y)* is a natural starting point to model health care costs. It is familiar and straightforward to implement. Estimation by least squares is easy and fast to compute in standard software even when there are hundreds of regressors and millions of observations, which is often the case of risk adjustment models based on administrative data. The model is specified on the "natural" cost scale, measured directly as costs in dollars, pounds, etc., and no prior transformation is required. As the natural cost scale is used the effects of covariates (*x*) are on the same scale and are easy to compute and interpret:

$$y_i = x_i'\beta + \varepsilon_i$$

The model can be estimated by ordinary least squares (OLS) and predictions of the conditional mean of costs are given by:

$$\hat{\mu}(x_i) = x_i'\hat{\beta}$$

The specification of the regression model can be checked using a variety of diagnostic tests. These are presented here in the context of the linear cost regression model but can be extended to models for transformed costs and to the nonlinear regression models presented below.

With individual level data on medical costs there will typically be a high degree of heteroskedasticity in the distribution of the error term, as indicated by relevant diagnostic tests (Breusch-Pagan, 1979; Godfrey, 1978; Koenker, 1981; White, 1980). The norm is to estimate the model using robust standard errors and use these for inference (White, 1980).

A Ramsey (1969) RESET test, based on re-running the regression with squares and other powers of the fitted values included as auxiliary variables, is often used as a test for the reliability of the model specification. In the health economics literature Pregibon's (1980) link test is widely used as an alternative to the RESET, this adds the level of the fitted values rather than including the individual regressors. For the

nonlinear models discussed below the RESET and link tests may be augmented by a modified Hosmer-Lemeshow (1980, 1995) test and its variants. The idea here is to compute the fitted values and prediction errors for the model, on the raw cost scale. These prediction errors can then be regressed on the fitted values, testing whether the slope equals zero. In the modified Hosmer-Lemeshow test an F statistic is used to test for equality of the mean of the prediction errors over, say, deciles of the fitted values, often accompanied by a graphical residual-fitted value plot of the relationship on the cost scale. This can be implemented by regressing the prediction errors on binary indicators for the deciles of the fitted values and testing the joint significance of the coefficients.

A potential downside of heavily parameterized models is that they may over-fit a particular sample of data and perform poorly in terms of out-of-sample forecasts. When models are to be used for prediction, the Copas test provides a useful guide to out-of-sample performance and guards against over-fitting (Copas, 1983; Blough et al., 1999). The Copas test works by randomly splitting the data into an estimation, or training, sample and a forecast, or holdout, sample (see e.g., Buntin and Zaslavsky, 2004). The model is estimated on the former and used to form predictions on the latter. The predictions from the forecast data are then regressed on actual costs to test whether the coefficient on the predictions is significantly different from 1 over multiple replications of the random sampling. Evidence of a significant difference suggests a problem of over-fitting[2]. It should be noted that the tests for model specification - such as the RESET, link and Copas tests – are sensitive to the presence of outliers in the data and diagnostics for influential observations should be checked, particularly when split sample tests are used (Basu and Manning, 2009).

2.2 Regression on transformed costs

As health care cost data involves working with non-normal distributions on the raw scale, for both costs and for the model residuals, much of the early literature focused on transforming the cost data to produce a more symmetric distribution (see e.g., Carroll and Rupert, 1988; Manning, 1998; Manning et al., 2005; Mullahy, 1998). The most popular transformation is the log transformation but square-root transformations and other power functions are applied as well. The distinctive feature of the transformation approach is that the regression model is specified on the transformed scale and that the model no longer works with the raw cost scale.

[2] Split sample methods, such as balanced half samples, are inefficient, as only a portion of the data is used for estimation. A related approach is v-fold, or leave v out, cross validation; for each subset of v observations in the data the model is estimated with n-v observations and used to predict the v observations. Setting v=1, the leave one out approach, means estimating the model n times which may be computationally expensive. Ellis and Mookin (2008) propose an efficient Jacknife style variant of the Copas test which makes better use of the data than the conventional 50:50 split and, in the context of the classical linear model, avoids the need to estimate the model multiple times.

Log transformations

Using a logarithmic transformation of cost data typically reduces skewness, making the distribution more symmetric and closer to normality. This has lead to widespread use of regression models for the log of costs. One of the problems with this approach is that it requires arbitrary additional transformations if there are zero observations or the use of two-part specifications to deal with the zeros. More importantly standard regression estimates provide predicted costs on the log scale, while analysts typically want results that are expressed in terms of actual costs. Simple exponentiation of the predictions does not result in predictions on the original cost scale. To deal with this problem it is necessary to apply a smearing factor which is not always straightforward to implement. This weakens the case for working with transformed data and, in particular, problems arise with the retransformation if there is heteroskedasticity in the data on the transformed scale (Manning, 1998; Manning and Mullahy, 2001; Mullahy, 1998).

The log regression model takes the form:

$$\ln(y_i) = x_i'\beta + \varepsilon_i$$

The error term is assumed to have the standard properties:

$$E(\varepsilon) = 0 \qquad E(x'\varepsilon) = 0$$

Interest lies in predicting costs on the original scale and, given $E(\ln(y)) \neq \ln(E(y))$, this relies on retransforming to give[3]:

$$y_i = \exp(x_i'\beta + \varepsilon_i) = \exp(x_i'\beta)\exp(\varepsilon_i)$$

Then:

$$E(y_i \mid x_i) = \exp(x_i'\beta)E(\exp(\varepsilon_i) \mid x_i)$$

If the error term is normally distributed, with variance σ_ε^2 , then it is possible to estimate the conditional mean for the log-normal distribution using the OLS estimates of β and σ:

$$\hat{\mu}(x_i) = \exp\left(x_i'\hat{\beta} + 0.5\hat{\sigma}_\varepsilon^2\right)$$

If the error term is not normally distributed, but is homoskedastic, then the estimate based on log-normality will be biased. Instead the Duan (1983) smearing estimator can be applied. In this case the conditional mean is estimated using:

$$\hat{\mu}(x_i) = \hat{\varphi} \times \exp\left(x_i'\hat{\beta}\right)$$

where $\hat{\varphi}$ is the estimated smearing factor:

$$\hat{\varphi} = (n-k-1)^{-1}\sum_i \exp(\hat{\varepsilon}_i), \qquad \hat{\varepsilon}_i = \ln y_i - x_i'\hat{\beta}$$

where n is the sample size and k is the number of parameters in the regression. Typically this smearing factor lies between 1.5 and 4.0 in empirical applications with health care costs, illustrating the fact that ignoring the retransformation can lead to substantial underestimation of average costs.

[3] Basu et al., (2006) refer to the 'scale of interest' and the 'scale of estimation'.

If the error term on the log scale is heteroskedastic, Duan's homoskedastic smearing estimator will lead to bias, with the bias being a function of ***x*.** In the lognormal case:

$$\hat{\mu}(x_i) = \exp\left(x_i'\hat{\beta} + 0.5\hat{\sigma}_\varepsilon^2(x_i)\right)$$

In the general case:

$$\hat{\mu}(x_i) = \rho(x_i) \times \exp\left(x_i'\hat{\beta}\right)$$

This shows that eliminating bias in the predictions requires knowledge of the form of heteroskedasticity. This may be manageable if there are a limited number of binary regressors. For example, the approach adopted in the RAND Health Insurance Experiment was to split the sample by discrete *x* variables and apply separate smearing estimates (see e.g. Manning et al., 1987b). In general, this is difficult if the number of regressors is large and contains continuous variables. However, it is possible to exploit the fact that:

$$\rho(x_i) = E[\exp(\varepsilon_i) \mid x_i]$$

This suggests running a regression of the exponentiated residuals on *x* and using the fitted values as the smearing factor[4]. An alternative is to use separate smearing factors for different ranges of predicted costs, for example Buntin et al., (2004) use a separate smearing factor for the top decile. Ai and Norton (2000) provide standard errors for the retransformed estimates when there is heteroskedasticity.

Square root transformations

Square-root transformations have been favoured over log transformations in some applications. In this case the implied model is:

$$\sqrt{y_i} = x_i'\beta + \varepsilon_i$$

The smearing estimator can be adapted to the square root transformation to give estimates of the conditional mean:

$$\hat{\mu}(x_i) = \hat{\varphi} + \left(x_i'\hat{\beta}\right)^2$$

The smearing factor, assuming homoskedastic errors, is:

$$\hat{\varphi} = N^{-1}\sum_i \hat{\varepsilon}_i^2$$

In the heteroskedastic case predictions take the form:

$$\hat{\mu}(x_i) = \rho(x_i) + \left(x_i'\hat{\beta}\right)^2$$

Here the smearing factor can be estimated by running a regression of the squared residuals on functions of *x*, such as the fitted values of the linear index.

Box-Cox models

Rather than imposing a particular transformation, a Box-Cox transformation can be used to specify the cost regression (see Box and Cox 1964; Chaze 2005):

$$y_i^{(\lambda)} = \frac{y_i^\lambda - 1}{\lambda} = x_i'\beta + \varepsilon_i$$

[4] Veazie et al., (2003) adopt a variant of this approach using the fitted values of the linear index in place of *x* in the context of a square root transformation.

This includes levels (λ=1) and logs (λ=0) as special cases. Assuming ε has a normal distribution, λ can be estimated, along with the other parameters, by maximum likelihood estimation in packages such as Stata (more general models are also available that apply the Box-Cox transformation to the covariates as well). Retransformation of predictions to the cost scale is not straightforward, especially in the presence of heteroskedasticity. A more satisfactory use of the Box-Cox transformation is provided by the Extended Estimating Equations (EEE) approach that is discussed in Section 4 below.

Semiparametric transformation models

The flexibility of the Box-Cox transformation is taken a step further, while maintaining the idea of writing transformed costs as a linear function of the regressors, in a recent papers by Welsh and Zhou (2006) and Zhou et al. (2009). For example, Zhou et al. (2009) propose a semiparametric transformation model:

$$H(y_i) = x_i'\beta + \sigma(x_i'\gamma)\varepsilon_i$$

The specification is semiparametric in two senses: the transformation H(.) is treated as an unknown increasing function and the error ε~(0,1) has an unknown distribution. The function σ(.) captures heteroskedasticity and is assumed to be a known function. Estimation is based on an iterative algorithm that cycles between estimating β and γ, given H(.), and estimation of H(.) by nonparametric regression, given β and γ. Predictions are derived from an extended version of Duan's (1983) smearing estimator:

$$\hat{\mu}(x_i) = \frac{1}{n}\sum_{i=1}^{n} \hat{H}^{-1}\left[x_i'\hat{\beta} + \sigma(x_i'\hat{\gamma})\left(\frac{\hat{H}(y_i) - x_i'\hat{\beta}}{\sigma(x_i'\hat{\gamma})} \right) \right]$$

3. Nonlinear regression models

3.1 Exponential Conditional Mean models

The transformation approach discussed above deals with the non-normality of costs by finding a transformation that makes the outcome more symmetric and then estimating a linear regression on that scale. But these models can perform poorly and create the problem of retransforming predictions back to an economically meaningful scale. To avoid this problem, the exponential conditional mean (ECM) model assumes a nonlinear relationship for the cost regression, such that:

$$E\left[y_i \mid x_i \right] = \mu_i = \phi \exp\left(x_i'\beta \right)$$

The ECM model is written in a general form here, to encompass specifications where the conditional mean is proportional to the exponential function. The use of the exponential function recognises that the object of interest, health care costs, is a non-negative quantity and accommodates the typical skewed shape of the distribution. Notice also that this implies that the effect of covariates is proportional rather than additive, with a constant proportional effect (see e.g., Gilleskie and Mroz, 2004).

The ECM, and related extensions, can be estimated in a variety of ways. In practice this is done using nonlinear least squares (NLS); the Poisson quasi-maximum likelihood (QML) estimator; and using hazard models (for example, based on exponential, Weibull and generalized gamma distributions). Also, the ECM is closely related to generalized linear models (GLMs), which are covered in Section 4.

The ECM can be viewed as a nonlinear regression model:

$$E\left[y_i \mid x_i \right] = \exp\left(x_i'\beta \right)$$

This can be estimated by nonlinear least squares (NLLS) or, more generally, by the generalized method of moments (GMM**)**. The relevant first-order/moment conditions are solved iteratively to give estimates of the regression parameters:

$$\sum_i \left\{ y_i - \exp\left(x_i'\beta \right) \right\} x_i = 0$$

As this approach only uses the first moment rather than the full probability distribution, it may be more robust than maximum likelihood, but it may also be less efficient, depending on the form of the variance function.

3.2 Poisson regression

The basic model used for integer-valued count data is the Poisson model. This model, and extensions such as the negative binomial model, are often used in health economics to model the number of visits to a doctor but the models can also be applied to continuous measures of health care costs (see e.g., Jones, 2000).

In the Poisson model the dependent variable y_i is assumed to follow a Poisson distribution, with mean μ_i, defined as a function of the covariates x_i. Thus, the model is defined by the distribution:

$$P(y_i) = \frac{e^{-\mu_i} \mu_i^{y_i}}{y_i!}$$

where the conditional mean μ_i is specified by:

$$\mu_i = E\left[y_i \mid x_i \right] = \exp\left(x_i'\beta \right)$$

So the Poisson model has the ECM form and standard software designed for Poisson regression can be used to estimate the β parameters by maximum likelihood, even if the dependent variable is not an integer count, as in the case of the skewed distribution of health care costs. The quasi-maximum likelihood (QML) property of the Poisson estimator means that, so long as the mean is correctly specified, it is consistent even if higher moments, such as the conditional variance, are misspecified. In this case robust standard errors, computed using the sandwich estimator, are used in place of the standard ML estimates.

3.3 Hazard models

The ECM and its extensions can be estimated using standard estimation routines for parametric hazard models. These models are normally applied to duration data but, as

with count data regressions, can also be used for health care costs (see e.g., Jones, 2000).

For example, the Weibull model has a hazard function:

$$h\left[y_i \mid x_i\right] = \eta.\rho y_i^{\rho-1} \exp\left(x_i'\beta\right)$$

where ρ is known as the shape parameter. The hazard is monotonically increasing for ρ>1, showing increasing duration dependence, and monotonically decreasing for ρ<1, showing decreasing duration dependence. ρ=1 gives the exponential distribution. Standard maximum likelihood estimation can be used to obtain estimates of the parameters η, ρ and β. In the context of cost data the parameter ρ provides flexibility to capture the shape of the distribution and, in particular, to allow for its skewness. The Weibull model can be expressed in proportional hazard form but can also be written in what is called the accelerated time to failure format, which expresses the log of *y* as a function of the dependent variables and the shape parameter:

$$\log(y_i) = \frac{1}{\rho}\left\{-\log(h) - x_i'\beta + \log(-\log(S(y_i)))\right\}$$

where log(-log(S(y))) has an extreme value distribution. This provides an intuitive link to the ECM model and to log transformed models of costs.

The scope for parametric modelling of survival data is taken a step further by the generalized gamma model (GGM), which is often used as a flexible parametric distribution for survival models. Manning et al. (2005) propose the use of this distribution as a flexible way of modelling non-normal health care cost data. The generalized gamma has density function:

$$f(y_i;\kappa,\beta,\sigma) = \frac{\gamma^{\gamma}}{\sigma y_i \sqrt{\gamma}\,\Gamma(\gamma)} \exp(z_i\sqrt{\gamma} - u_i)$$

where

$$\gamma = \left|\kappa\right|^{-2}, z_i = sign(\kappa)\left\{\ln(y_i) - \mu_i\right\}, u_i = \gamma\exp(\left|\kappa\right| z_i)$$

$$\mu_i = x_i'\beta$$

Special cases of the distribution are the gamma ($\kappa = \sigma$) , Weibull ($\kappa = 1$), exponential ($\kappa = 1, \sigma = 1$), and lognormal ($\kappa = 0$). The model can be estimated by maximum likelihood, for example using the `streg` command in Stata, and the restrictions implied by the nested specifications can be tested explicitly.

In general the *r*th uncentred moment of the generalized gamma distribution is:

$$E(y^r) = \left(\exp(\mu)\kappa^{2\sigma/\kappa}\right)^r \left[\frac{\Gamma\left(\frac{1}{\kappa^2} + \frac{r\sigma}{\kappa}\right)}{\Gamma\left(\frac{1}{\kappa^2}\right)}\right]$$

So, the conditional mean of costs is:

$$E(y_i \mid x_i) = \mu_i = \exp(x_i'\beta)\left[\kappa^{2\sigma/\kappa}\frac{\Gamma\left(\frac{1}{\kappa^2}+\frac{\sigma}{\kappa}\right)}{\Gamma\left(\frac{1}{\kappa^2}\right)}\right] = \exp(x_i'\beta)\phi$$

This shows that the model fits within the ECM class, with the mean proportional to an exponential function[5]. It also highlights the form of the various special cases as well. For example, for the Weibull ($\kappa = 1$) :

$$E(y_i \mid x_i) = \exp(x_i'\beta)\Gamma(1+\sigma) = \exp(x_i'\beta)\Gamma\left(1+\frac{1}{\rho}\right)$$

For the Gamma distribution ($\kappa = \sigma$):

$$E(y_i \mid x_i) = \mu_i = \exp(x_i'\beta)\left[\kappa^{2}\frac{\Gamma\left(\frac{1}{\kappa^2}+1\right)}{\Gamma\left(\frac{1}{\kappa^2}\right)}\right] = \exp(x_i'\beta)$$

The conditional variance of the generalized gamma model (and of the standard gamma) is proportional to the square of the mean.

Manning et al. (2005) propose that, when there is evidence that κ is small (<0.1), it is better to use a specification with additional heteroskedasticity, generated by assuming $\sigma = \exp(z_i'\alpha)$ for a set of regressors z. This ensures that the special cases of the GGM, such as the lognormal model, allow for heteroskedasticity through σ.

The use of hazard models is taken further by Basu et al. (2004) who compare log-transformed models for health care costs to the semiparametric Cox (1972)

[5] There is a link here with the generalized beta of the second kind (GB2) distribution. This has been used to model the size distribution of earnings and in analyses of income inequality and it nests other distributions such as the Burr-Singh-Maddala (BSM) and Dagum, among others (see for example, Parker, 1999; Jenkins, 2009). Mullahy (2009) discusses the issue of heavy tailed distributions and the use of the BSM distribution but the GB2 distribution does not seem to have been applied to health care costs. The mean of the GB2 distribution is:

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

Using $b = \exp(x_i'\beta)$ and treating the other parameters as scalars puts this in the ECM class of models. The Burr-Singh-Maddala distribution is a special case when p=1, the Dagum is a special case when q=1 and p=q=1 gives the log-logistic. Also, the generalized gamma, and hence the gamma and Weibull, are limiting cases of the GB2.

proportional hazard model. In the Cox model the hazard function at *y* for individual *i* is:

$$h\left[y_i \mid x_i\right] = h_0(y_i)\exp\left(x_i'\beta\right)$$

Cox's method is described as being semiparametric because it does not specify the baseline hazard function $h_O(y)$. Estimation uses the partial log-likelihood function,

$$LogL = \sum_i \left\{ x_i'\beta - \log\left(\sum_{l \in R_i} \exp(x_l'\beta)\right)\right\}$$

where $l \in R_i$ are those observations in the risk set, R_i, at the point of exit of individual *i*. By conditioning on the risk set the baseline hazard $h_O(y)$ is factored out of the partial likelihood function. A drawback of the Cox model for modelling costs is that estimates of the baseline hazard are required to estimate the conditional mean. But the model does provide a benchmark for testing the 'proportional hazards' assumption that is implicit in the choice of an ECM specification.

6. Generalized linear models

6.1 Basic approach

The dominant approach to modelling health care costs in the recent literature has been the use of generalized linear models (see e.g., Blough et al 1999; Buntin and Zaslavsky, 2004; Manning and Mullahy, 2001; Manning et al., 2005; Manning, 2006). Generalized linear models (GLMs) specify the conditional mean function directly:

$$E\left[y_i \mid x_i\right] = \mu_i = f\left(x_i'\beta\right)$$

For example, with an exponential conditional mean (ECM) or 'log link':

$$E\left[y_i \mid x_i\right] = f\left(x_i'\beta\right) = \exp\left(x_i'\beta\right)$$

The first component of a GLM model is a link function *g(.)* that relates the conditional mean to the covariates:

$$g(\mu_i) = x_i'\beta$$
$$\Rightarrow \mu_i = g^{-1}(x_i'\beta) = f(x_i'\beta)$$

The second component is a distribution (D) that belongs to the linear exponential family. This is used to specify the relationship between the variance and the mean:

$$Var(y_i \mid x_i) = v(\mu_i)$$

Advantages of the GLM approach are that predictions are made on the raw cost scale, so that no retransformation is required, and that they allow for heteroskedasticity through the choice of distributional family, albeit limited to specifications of the conditional variance that are pre-specified functions of the mean.

The link function specifies the shape of the conditional mean function. The most commonly used link functions are the identity – where covariates act additively on mean, so that the interpretation of coefficients is the same as linear regression – and

the log link – where covariates act multiplicatively on mean. The link function characterises how the mean on the raw cost scale is related to the set of covariates. For example, with a log link:

and:

$$E\left[y_i \mid x_i\right] = \exp\left(x_i'\beta\right)$$
$$\ln\left(E\left[y_i \mid x_i\right]\right) = x_i'\beta$$

The chosen distribution is used to describe the relationship between the variance and conditional mean. Often this is specified as a power function:

$$\text{var}\left[y_i \mid x_i\right] \propto \left(E\left[y_i \mid x_i\right]\right)^{\upsilon} = \mu^{\upsilon}$$

Common distributional families based on the power function include:

- Gaussian: constant variance; υ=0
- Poisson: variance proportional to the mean; υ=1
- Gamma: variance proportional to the square of the mean; υ=2
- Inverse Gaussian: variance proportional to cube of the mean; υ=3

Other common distributions within the GLM framework use a quadratic function of the mean, in particular the Bernoulli, $\mu(1-\mu)$, and binomial, $n\mu(1-\mu)$.

These distributions allow considerable flexibility in modelling cost data, although the modelling of the variance is restricted to being a specified function of the mean. Note that the Gaussian distribution with an identity link function is comparable to linear regression. The distribution and link functions can be combined freely, although there are canonical links for each distribution. The most popular specification of the GLM for health care costs has been the log-link with a gamma error (Blough et al., 1999; Manning and Mullahy, 2001; Manning et al., 2005).

Estimation of GLMs is based on the classical "estimating equations" or quasi-score functions:

$$\sum_i \left\{ \frac{y_i - \mu_i(\beta)}{v_i(\mu)} \right\} \frac{\partial \mu_i}{\partial \beta} = \sum_i \left\{ \frac{r_i}{\sqrt{v_i(\mu)}} \right\} \frac{\partial \mu_i}{\partial \beta} = 0$$

where r is the Pearson or standardized residual and $\frac{1}{\sqrt{v_i(\mu)}} \frac{\partial \mu_i}{\partial \beta}$ are the standardized regressors (see, Wedderburn, 1974). GLMs are based on the linear exponential family of distributions:

$$f_{LEF} = \exp(a(\mu) + b(y) + c(\mu)y)$$

This means they have the pseudo- or quasi-ML property and estimates are consistent so long as the mean is correctly specified (Gourieroux et al., 1984)[6]. The estimator only specifies the conditional mean and variance functions, so more efficient

[6] Cantoni and Ronchetti (2006) propose a robust variant of GLM that modifies the quasi-score equations to make the estimator less sensitive to outliers.

estimators may be obtained that make use of correctly specified functions for higher moments, such as the skewness of the distribution.

The LEF density presented above is what is known as the mean parameterisation of the density, where:

$$E\left[y_i\right] = \mu = -\frac{a'(\mu)}{c'(\mu)}$$

GLMs are more typically presented in terms of the canonical parameterisation:

$$f_{GLM} = \exp\left(\frac{\theta y - b(\theta)}{a(\phi)} + c(y,\phi)\right)$$

Where:

$$E\left[y_i\right] = \mu = b'(\theta)$$

The canonical link is such that :

$$\eta = \theta \quad where \quad \eta = x'\beta$$

For example, with the Poisson distribution:

$$b(\theta) = \exp(\theta), \quad \mu = b'(\theta) = \exp(\theta), \quad \eta = \ln(\mu) = \ln(\exp(\theta)) = \theta$$

In applications the choice of link function and distribution is often guided by the use of the Pregibon link test, modified versions of Park's (1966) test for the distribution, and by the use of residual plots. The link test has been described above. In the context of GLMs it should be applied using the same link function and distribution as the model being tested, taking care to check for influential observations in the data. The idea of the modified Park test is that the GLM distribution should reflect the relationship between the variance and the mean, when this is based on a power function it implies:

$$\ln\left[\mathrm{var}\left(y_i \mid x_i\right)\right] = \ln\left(\alpha\right) + \upsilon \ln\left(\mu_i\right)$$

The test exploits this by regressing $\ln\left((y_i - \hat{y}_i)^2\right)$ on $\ln\left(\hat{y}_i\right)$ and a constant, typically using a GLM to estimate the model, having tested for the appropriate form of the link function to use (e.g., Manning and Mullahy, 2001). The estimated slope coefficient from the modified Park test provides guidance on the appropriate distributional family.

4.2 Extended estimating equations

In response to the problem of selecting the appropriate link and variance functions, Basu and Rathouz (2005) suggest a flexible semiparametric extension of the GLM model. Their model, which is labelled the extended estimating equations (EEE) approach, uses a Box-Cox transformation for the link function:

$$x_i'\beta = \frac{\mu_i^{\lambda} - 1}{\lambda} \quad where \quad \mu_i = E(y_i \mid x_i)$$

This includes the log-link as a special case along with other power functions of *y*.

This is combined with a general power function for the variance:

$$\mathrm{var}(y_i \mid x_i) = \upsilon_1 \mu_i^{\upsilon_2}$$

which gives a flexible specification that nests the common GLM distributions and allows the restrictions to be tested[7]. The additional parameters are estimated, along with the regression coefficients, by QML using the extended estimating equations. The EEE specification is heavily parameterized and care may be needed in calibrating numerical optimisation routines to estimate the model[8]. Basu et al. (2006) apply the EEE method to claims data on the incremental costs associated with heart failure.

5. Other nonlinear models

5.1 Finite mixture models

The proportional effect of covariates implied by the ECM may be too restrictive in some applications and evidence of heterogeneity, in the form of a multimodal distribution may indicate that costs can be modelled as a mixture. This can be done semiparametrically and finite mixture models have been applied to health care costs. For example Deb and Burgess (2007) use mixtures of gamma distributions.

To specify a finite mixture model, assume that each individual belongs to one of a set of latent classes j=1,…,C, and that individuals are heterogeneous across classes[9]. Conditional on the observed covariates, there is homogeneity within a given class j. Given the class that individual i belongs to, the outcomes have a density $f_j(y_i| x_i; \beta_j)$, such as a gamma distribution, where the β_j are vectors of parameters that are specific to each class. The probability of belonging to class j is π_{ij}, where $0< \pi_{ij} <1$ and $\sum_{j=1}^{C}\pi_{ij}=1$. Unconditional on the latent class the individual belongs to, the density of y_i is given by:

$$f\left(y_i \mid x_i;\pi_{i1},...,\pi_{iC};\beta_1,...,\beta_C\right)=\sum_{j=1}^{C}\pi_{ij}\, f_j\left(y_i \mid x_i;\beta_j\right)$$

The discrete distribution of the heterogeneity has C mass points and the πs need to be estimated along with the β_js. In most empirical applications of finite mixture models the class membership probabilities are treated as fixed parameters $\pi_{ij} = \pi_j$ but this can be relaxed (see for example, Deb and Trivedi, 1997; Deb and Holmes, 2000; Deb, 2001; Deb and Trivedi, 2002; Jiménez-Martin et al., 2002; Atella et al., 2004;

[7] Even greater flexibility is assumed by Chiou and Muller (1998) who leave the link and variance functions unspecified and estimated nonparametrically, by locally weighted least squares, as part of a three stage extension of the QML estimator. This method does not seem to have been applied to health care costs so far.

[8] Although, in Hill and Miller's (2009) comparative analysis of cost regression models the EEE estimator fails to converge in only 1.8 per cent of the 4,096 models they estimate.

[9] This section focuses on models for latent mixtures, where class membership is unobserved. Mixture models can of course be used when there is an observed split, such as two-part models, applied to zero and positive costs, or multi-part models, applied to different categories of inpatient and outpatient expenditures.

Conway and Deb, 2005; Bago d'Uva, 2006). After estimating the model, it is possible to calculate the posterior probability that each individual belongs to a given class. The posterior probability of membership of class *j* depends on the relative contribution of that class to the individual's likelihood function. This is given by:

$$P[i \in j] = \frac{\pi_{ij} f_j(y_i \mid x_i; \beta_j)}{\sum_{k=1}^{C} \pi_{ik} f_k(y_i \mid x_i; \beta_k)}$$

Each individual can then be assigned to the class that has the highest posterior probability for them and the predicted costs can be calculated separately for each class.

5.2 The discrete conditional density estimator

Gilleskie and Mroz (2004) propose a semiparametric approach that divides the data into a fixed number of discrete intervals then applies discrete hazard models, implemented as a sequence of logits, to estimate the conditional density function. From that the conditional mean and other conditional expectations can be formed. This approach can be seen as a generalisation of the two-part model into a multi-part model: in which a separate estimate of the conditional mean is used for each of the intervals and the probability of costs lying in each interval is a function of the covariates.

The approach begins by dividing the support of *y* into a fixed number (*K*) of discrete intervals, or bins; these may be chosen to contain an equal number of observations, such as deciles, or they may reflect features of the distribution such as a mass point at zero. The estimator focuses on an approximation to the conditional expectation of some function of costs *h(y)*. This takes the form of a weighted average:

$$E[h(y_i) \mid x_i] = \int h(y_i) f(y_i \mid x_i) dy \approx \sum_k h^*(k) p[y_{k-1} \le Y < y_k \mid x_i]$$

where *h*(k)* is an approximation of the function of interest within the *k*th interval. The general formulation of the conditional expectation nests the conditional mean of costs, where *h(.)* is simply an identity. In practice, Gilleskie and Mroz (2004) choose to use the sample mean within each interval to implement the approximation, which does not allow for heterogeneity within the intervals, but local regressions could be used instead. This may be a particular problem with the open-ended interval at the top end of the distribution that contains the high cost cases (Basu and Manning, 2009)

The heart of the approach is estimation of the conditional probabilities of belonging to each interval which are then used as the weights in the averaging. They suggest that this should be estimated by a discrete hazard specification implemented using logit models on an expanded version of the data. A separate logit could be estimated for each interval but they adopt a pooled logit model that smooths over the intervals using higher order polynomials in the regressors. The model is estimated for a given number of partitions of the support of *y*. To choose the appropriate number of partitions Gilleskie and Mroz (2004) suggest selecting the value that maximises a

penalised log-likelihood and, on the basis of Monte Carlo experiments, indicate that 10-20 intervals will usually be sufficient.[10] Standard errors are obtained by bootstrapping the whole procedure.

6. Comparing model performance

6.1 Evidence from the literature

There is a rich literature that compares the performance of methods of estimating health care costs (see for example, Basu et al., 2004; Basu et al., 2006; Buntin and Zaslavsky, 2004; Deb and Burgess 2007; Duan et al., 1983; Gilleskie and Mroz, 2004; Hill and Miller 2009; Manning and Mullahy, 2001; Manning et al., 2005; Montez-Roth et al., 2006; Veazie et al., 2003). These studies include classical Monte Carlo analyses, with hypothetical cost data drawn randomly from specified parametric distributions, along with studies of empirical datasets that use a quasi-Monte Carlo design, with estimation and forecast samples drawn from the data. The former allow the performance of estimators to be assessed against known parameter values. The latter allow the predictive performance to be assessed when the models are confronted with the idiosyncracies of the distribution of actual cost data, rather than textbook parametric distributions, although the findings may then be specific to particular measures of costs and specific groups of people. A general finding of these studies is that the appropriate specification varies from application to application, for example, depending on whether the costs relate to elderly or non-elderly patients and whether total health care costs or specific costs such as prescription drug spending are being modelled. Table 1 illustrates the range of methods spanned by some recent published studies.

[10] The estimation routine is described as being a maximum likelihood procedure but the properties of the estimator, with respect to the sample size and number of intervals are not derived. Although the approach is not based on explicit distributional assumptions it does use explicit, logit, functional forms. So, compared to some other semiparametric estimators, predictions can be computed for counterfactual values of the regressors. This is used to compute numerical derivatives of the expected values.

Table 1: Coverage of methods in some recent comparative studies

	Basu, Manning & Mullahy (2004 HEc)	Manning, Basu & Mullahy (2005, JHE)	Basu, Arondeker & Rathouz (2006 HEc)	Deb & Burgess (2007)	Hill & Miller (2009, HEc)
OLS on y			■	■	■
OLS on ln(y) + Duan	■	■	■	■	■
OLS on $\sqrt{y}$				■	
Box-Cox					
GLM log-gamma	■		■		■
GLM linear-gamma & quadratic-gamma				■	
EEE			■		■
Weibull	■	■			
Generalized gamma		■			■
Cox PH	■				
FMM gamma				■	

One of the most comprehensive published comparisons of methods is provided by Hill and Miller (2009). They compare many of the models for positive expenditures that have been discussed above: linear OLS; OLS on log costs with smearing: GLMs using a log link and Poisson or gamma distributions; the standard generalized gamma model (GGM), without additional heteroskedasticity, and the extended estimating equations model (EEE). The GGM and EEE are the most flexible approaches and are not nested within each other, but they both share the gamma model as a common special case. Hill and Miller's empirical analysis is based on the first eight waves of the US Medical Expenditure Panel Survey (MEPS) spanning the years 1996-2003. They regress medical expenditures on measures of chronic conditions and socioeconomic characteristics from the previous wave of data. To encompass different shapes of cost distributions the analysis uses two groups of people, elderly people who are eligible for Medicare and non-elderly people who have insurance, and two measures of costs, total health care expenditure and expenditures on prescription drugs. This gives four sub-samples and the shape of the distribution of costs differs across the samples. The comparison of models uses cross validation, in the style of the Copas test, with repeatedly grouped balanced half-samples (RGBHS) that take account of the complex survey design of MEPS. This allows estimation and validation on 1024 half-samples and the models are compared in terms of model fit and out-of-sample predictions.

Hill and Miller's findings echo earlier work which shows that different functional forms (link functions) work better with different sub-samples and that it is not the case that one specification dominates. The log link works well for total expenditures among the non-elderly but not for the elderly. While for prescription drugs a square root link gives a better fit for both elderly and non-elderly. Bias is measured by the mean prediction error (MPE) and predictive accuracy is measured by the mean absolute prediction error (MAPE). The log transformed OLS model performs poorly, leading to substantial over-predictions and has the worst fit for all four distributions. The best performing models are linear OLS, Poisson regression and the EEE model. Linear OLS and EEE have less over-fitting, while the GGM and OLS on logs are much more likely to over-fit the data. The performance of the GGM and standard gamma model deteroriates when a log link is not appropriate, as is the case for three out of the four empirical distributions.

The MEPS data has a relatively small sample size. In contrast Deb and Burgess (2007) make use of 3 million observations from claims data for the US Department of Veterans Affairs (VA) for financial year 2000. This allows them to assess the role of sample size in determining the comparative performance of different methods[11]. They use a quasi-Monte Carlo approach, dividing the data into estimation and prediction groups, each with 1.5m observations. The estimation group is then randomly sampled, with replacement, to give estimation samples of five different sizes ranging from 10,000 to 500,000. Twenty samples are generated for each sample size. Predictions are computed using the full prediction group. These are evaluated using the mean prediction error (MPE) that indicates overall bias; the mean absolute prediction error (MAPE) that indicates the ability of the models to predict individual costs; and the absolute deviations of the MAPE (ADMAPE), based on deviations across the experimental replications. The models control for diagnostic groups and comorbidities and are estimated with and without trimming of the top 5 per cent of the cost data[12]. The results from the multiple simulations are combined and summarised using response surface regressions.

As in Hill and Miller (2009), and other recent studies, the log regression model performs poorly across the board in terms of bias (MPE) and predictive accuracy (MAPE). Linear and square root regressions exhibit negligible bias on the untrimmed prediction samples. When the data is trimmed of the top 5 per cent of costs finite mixtures of gammas, with 2 or 3 components, do better than the regression models. Comparison of the different sample sizes suggests that the linear and square root regressions converge on the asymptotic values of the MPE for sample sizes of 20-

[11] Montez-Roth et al. (2006) also use VA data and compare sample sizes ranging from 5,000 to 500,000, with a focus on expenditures by patients with diagnoses for mental health problems and substance abuse. Their comparison of linear, square root and log models suggests that the square root transformation works best for predictive accuracy with these data.

[12] Note that trimming only one end of the distribution will not be mean-preserving.

30,000, while the finite mixture models converge with samples of 30-40k[13]. When the focus shifts to the MAPE, the 2-component FMM dominates, whether or not the data is trimmed. This specification also does best in terms of the DAMAPE, which captures the variability across replications, but the other best performing models – the square root regression and the gamma model - give similar results and linear OLS is not far behind.

7. An empirical application

To illustrate the performance of the various specifications discussed above, this section presents an empirical application that draws on an easily accessible dataset. This is taken from *Microeconometrics using Stata* by Cameron and Trivedi (2009) and the dataset is available through their web page. The original source is the US Medical Expenditure Survey (MEPS), which is a set of surveys of families and individuals, their medical providers and employers across the US. The surveys collect data on the use of health services (e.g. frequency and cost) and whether individuals hold health insurance. The particular subset of data is taken from the MEPS sample used in Chapter 3 (p.71) of Cameron and Trivedi (2009), available as the Stata dataset mus03data.dta[14]. Cameron and Trivedi describe the data as follows:

"We analyze medical expenditure of individuals aged 65 years and older who qualify for health care under the U.S. Medicare program. … Medicare does not cover all medical expenses. For example, copayments for medical services and expenses of prescribed pharmaceutical drugs were not covered for the time period studies here. About half of eligible individuals therefore purchase supplementary insurance in the private market that provides insurance coverage against various out-of-pocket expenses." (p71)

Total annual health care expenditures are measured in US dollars and this is the outcome variable in the cost regressions. Sociodemographic and health-status measures are also available together with insurance status. Following Cameron and Trivedi (2009) a simple additive specification of the linear index is used that includes indicators of supplementary private insurance, physical limitations, activity limitations, the number of chronic conditions, age, gender and household income as regressors. It is important to note that the simple comparison of models presented hear uses the same linear index in each specification. In empirical applications a richer specification will typically be used with many more covariates and with polynomials and interaction terms, perhaps using a fully saturated model as a starting point if sufficient data is available (Manning *et al.*, 1987a). A fuller and fairer

[13] Note that the FMM performs poorly on the MPE criterion in the empirical case study presented in Section 7 which uses a much smaller sample of around 3,000 observations from MEPS.

[14] The MEPS has a complex survey design that involves over-sampling of specific groups. However sample weights and other design variables are not included in this subset and, purely for the purposes of this empirical illustration, it is treated here as if it was a simple random sample.

comparison of models may entail using different specifications of the regressors for each model, so that the best specification of one model is compared with the best specification of another[15]. For example, Veazie et al. (2003) discuss the case where a linear specification, $x_i'\beta$, is appropriate on the square root scale, then the appropriate specification on the levels scale would be a quadratic function of $x_i'\beta$.

The models presented here are the ones most commonly used in the health economics literature and some of the recent innovations: OLS estimates for linear regression of actual costs; OLS estimates for regressions on log and square root transformations, using the Duan smearing estimator; the ECM model, estimated by NLLS and using the Poisson ML estimator; the generalized gamma model estimated by ML, including the specification with additional heteroskedasticity; the generalised beta of the second kind (GB2); four variants of the GLM, one with a square root link and gamma distribution and three with a log link but with gamma, log-normal and Poisson distributions; the extended estimating equations model (EEE); and a finite mixture model (FMM) with a two-component gamma mixture. All of the models are estimated in Stata, using built-in and user-written commands[16].

The sample is made up of 2,955 individuals who have positive annual medical expenditures (109 cases with zero costs are excluded). The mean cost is $7,290, with a minimum of $3 and a maximum of $125,610. The interquartile range of $6,064 is quite tight compared to the overall range and the distribution of costs is distinguished by a very heavy right-hand tail. The skewness statistic is 4.1 (compared to 0 for symmetric data) and kurtosis is 25.6 (compared to 3 for normal data). As expected for heavily skewed data, the median cost, $3334, is less than half the mean cost.

Estimates for linear regression on the level of costs show evidence of a high degree of heteroskedasticity. For example, the Breusch Pagan test gives a F statistic of 74.1 and the White test statistic is 104.0[17]. Although, it is notable that the use of Huber-White robust estimates does little to change the magnitude of the standard errors in this application. The estimated residuals from the linear model inherit the shape of the distribution of costs and are highly non-normal, with a skewness statistic of 4.1 and a kurtosis statistic of 26.4. Individual residuals can be very large and range from -17,311 to 113,095. Also using the linear model does lead to some negative predicted costs. Specification tests for the linear model, along with the other models, are discussed below.

As well as making the distribution of costs more symmetric, the logarithmic transformation shrinks the range of variation in the dependent variable. When linear regression is applied to the log of costs the adjusted R^2 goes from 0.11 for the levels

[15] I am grateful to Will Manning for this observation.
[16] The discrete conditional density estimator is not included in this exercise: at the time of writing, no standard command or user-written program for this method is available in the public domain.
[17] This is the F test version of the Breusch-Pagan statistic that drops the assumption of normality.

model to 0.23 for the log model. Heteroskedasticity is less severe than on the levels scale but does not disappear: the Breusch-Pagan F statistic is 33.2. Using the log model requires retransformed estimates to predict costs[18]. The estimate of the standard Duan smearing factor is 2.0. A similar retransformation process is applied to the estimates of the square root regressions (see Veazie et al., 2003). The final transformed regression approach used here is to estimate the Box-Cox model. This suggests a transformation that is close to the log transformation with an estimated value of λ equal to 0.076, although this estimate is statistically significantly different from 0. However this standard Box-Cox model does not allow for heteroskedasticity (unlike the EEE model).

The exponential conditional mean (ECM) model is estimated by nonlinear least squares (using the `nl` command) and Poisson regression (`poisson`). Extensions that allow for the mean to be proportional to the exponential function and capture the shape of the distribution using parametric hazard functions are estimated for exponential (`streg, dist(exp)`), Weibull (`dist(w)`) and generalized gamma (`dist(gamma)`) distributions. The latter can also be estimated using Anirban Basu's user-written code (`gengam2`). This provides tests of all of the nested special cases all of which are rejected, although the lognormal distribution performs best: these are the standard gamma (chi squared equals 359.85), lognormal (16.24), Weibull (258.27) and exponential (412.52). The estimated value of κ is 0.2 and the estimate of σ is 1.19. Although the value of κ is greater than 0.1 the generalized gamma model is also estimated with additional heteroskedasticity. All of the special cases of this variant of the model are rejected, with the lognormal again performing best. To complement the generalized gamma model another flexible size distribution is estimated; this is the generalised beta of the second kind (GB2) which is estimated by ML using Stephen Jenkin's program `gbfit2` (Jenkins, 2009)[19].

The generalized linear model (GLM) framework is used to estimate a set of models; the first has a square root link and gamma variance (`glm, link(power 0.5) family(gamma)`) and the others all have log links, coupled with a gamma distribution (`glm, link(log) family(gamma)`), a Poisson distribution (`family(poisson)`), and a lognormal distribution (`family(normal)`). The link test rejects the log link but does not reject the square root link. The modified Park tests for these specifications always reject specific integer values of υ, although values of 1 (Poisson) and 2 (gamma) perform best. These GLM specifications are nested with the extended estimating equations (EEE) model of Basu and Rathouz (2005) which is estimated by Anirban Basu's program `pglm`. The estimate of the Box-Cox parameter for the link function is 0.563, suggesting a square root rather than a log transformation, and the estimate of υ_2 is 1.67, between the Poisson and gamma distributions.

[18] The heteroskedastic smearing uses predictions from a regression of the exponentiated residuals on the fitted values of the linear index, having confirmed that all of the predictions have positive values.

[19] Note that this program uses a linear rather than an exponential specification to introduce the regressors so the version estimated here is not an ECM.

The finite mixture model is estimated for a two-component gamma specification, using Partha Deb's program `fmm`. This divides the sample into two classes with membership probabilities (π) of 0.75 and 0.25. The predicted costs for the first group average $2,956 and range from $694 to $23,978. While those for the second group are higher, averaging $15,868 and ranging from $5,101 to $91,232, suggesting a smaller group of heavy users of health care.

Table 2 summarises some specification tests. P-values are reported for the Pregibon link test (computed, where applicable, using `linktest` and the Pearson test, which is related to the Hosmer-Lemeshow approach, and tests whether the correlation coefficient between the prediction error and the fitted values, on the raw cost scale, equals 0. The Copas test is implemented by v-fold cross validation. The sample is split into equal groups of size v and predictions for those observations are based on the estimates of the model computed for the rest of the sample[20]. It is notable that the regression on log costs, the exponential conditional mean models and the glms with a log link all perform poorly according to the link test. The Copas test indicates that both versions of the generalized gamma specification suffer from over-fitting, although performance is improved by allowing for additional heteroskedasticity. Over-fitting seems to be less of a problem with the generalized beta of the second kind. The GLM log-gamma model, one of the more widely used empirical specifications, also performs poorly with these data according to the Copas test.

Table 3 presents measures of goodness of fit within the estimation sample and measures of predictive performance based on the cross validation approach. For the estimation sample the measures of goodness of fit include the R^2 from a regression of actual costs on the predicted values on the raw scale, as well as the related measure of root mean squared error (RMSE):

$$RMSE = \sqrt{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2 \Big/ n}$$

and the mean absolute prediction error (MAPE):

$$MAPE = \sum_{i=1}^{n} abs\left(y_i - \hat{y}_i\right) \Big/ n$$

For the cross validation estimates the RMSE and MAPE, which measure precision of the predictions, are augmented by the mean prediction error (MPE), which captures bias within the forecast sample:

$$MPE = \sum_{i=1}^{n}\left(y_i - \hat{y}_i\right) \Big/ n$$

The three models which perform best on each criterion are highlighted in bold.

Ordinary Least Squares estimation of the linear regression model, which is based on an estimator that maximises the R-squared, does best on this specific criterion within the estimation sample. The EEE model and the GLM model with square root link and gamma distribution have a similar performance to OLS. The generalized gamma

[20] Here the sample is split into 100 groups with either 29 or 30 observations in each group.

model performs worst on this criterion. The same pattern is reflected in the RMSE for both the estimation and forecast samples. Turning attention to the MAPE, which captures the precision of the predictions in terms of the level of costs, OLS no longer dominates and EEE and the GLM model do better. The finite mixture of gammas does even better in terms of MAPE. But this is offset by a large degree of bias in the forecast sample, indicated by the MPE. The bias is small for linear regression, the square root transformed regression, Poisson regression (ML and GLM) and the EEE model. The bias is substantial for the log transformed regression, the generalized beta of the second kind and the FMM.

The results illustrate that there may be a trade-off between bias and precision of forecasts, most starkly in the case of the FMM estimator. It is notable that the simple linear model, estimated by OLS, performs quite well across all of the criteria, a finding that has been reinforced for larger datasets than the one used here.

Table 2: Specification Tests

	Link test p value, within sample	***Pearson test p value, within sample***	***Copas test, v-fold cross validation***
OLS on y	**0.133**		**0.974 (0.608)**
OLS on ln(y)	0.000	0.000	0.528 (0.000)
OLS on $\sqrt{y}$	**0.712**	**0.855**	1.210 (0.001)
ECM - NLLS	-	**0.350**	**1.002 (0.968)**
ECM – Poisson-ML	0.000	**0.158**	0.897 (0.035)
Generalized gamma	0.000	0.000	0.590 (0.000)
Gen gamma + het	0.000	0.004	0.841 (0.013)
Generalized beta 2	-	**0.832**	**0.974 (0.621)**
GLM sqrt-gamma	**0.633**	**0.343**	**0.934 (0.178)**
GLM log-gamma	0.000	0.000	0.759 (0.000)
GLM log-normal	0.001	**0.350**	**1.002 (0.968)**
GLM log-poisson	0.000	**0.158**	0.897 (0.035)
EEE	-	**0.690**	**0.955 (0.371)**
FMM gamma	-	**0.935**	**0.963 (0.489)**

Notes:

i) The results for the Copas tests with v-fold cross validation are all based on 100 groups of size 29/30. The figures reported are the slope coefficient and the p value for the test of the null hypothesis that this coefficient equals 1.

ii) Numbers in bold indicate that the model was not rejected by the specification test at a 5% level of statistical significance.

Table 3: Measures of goodness of fit

	R^2	*RMSE*		*MAPE*		*MPE*
		(1)	*(2)*	*(1)*	*(2)*	
OLS on y	**0.116**	**11270**	**11307**	6225	6244	**-1.41**
OLS on ln(y)	0.095	11499	11906	6329	6639	-721.3
OLS on $\sqrt{y}$	0.114	11283	11338	**6181**	6252	**-1.12**
ECM - NLLS	0.113	11296	11353	6267	6294	-102.6
ECM – Poisson-ML	0.110	11312	11362	6196	6220	**-3.36**
Generalized gamma	0.093	11769	11714	6429	6452	-403.5
Gen gamma + het	0.106	11354	11395	6221	6245	-39.0
Generalized beta 2	0.110	11319	11337	6409	6423	-432.1
GLM sqrt-gamma	**0.115**	**11281**	**11311**	6185	**6203**	-29.3
GLM log-gamma	0.106	11390	11432	6254	6276	-147.0
GLM log-normal	0.113	11295	11353	6267	6294	-102.6
GLM log-poisson	0.110	11312	11362	6196	6220	**-3.36**
EEE	**0.116**	**11274**	**11310**	**6179**	**6200**	-7.06
FMM gamma	0.106	11395	11433	**5775**	**5793**	1132.7

Note: R^2 denotes the R-squared from a regression of actual costs on the predicted values; RMSE is the root mean squared prediction error, on the cost scale, where (1) is for the estimation sample and (2) is for the cross validation predictions; MAPE is the mean absolute prediction error; MPE is the mean prediction error (bias) for the cross validation predictions.

8. Further reading

This chapter has focused on estimating and predicting health care costs using regression models and microdata. A comprehensive guide to microeconometric methods in general is provided by:

Cameron, A. C. and P. K. Trivedi (2005). *Microeconometrics.* Cambridge: Cambridge University Press.

This has a companion text which shows how the techniques can be implemented in Stata, with many empirical examples, including the use of the MEPS data on health care expenditures:

Cameron, A. C. and P. K. Trivedi (2009). *Microeconometrics Using Stata.* College Station Texas: Stata Press.

Models for health care costs are often based on health survey data. The issues associated with survey design, sampling, nonresponse and imputation, and inference with complex surveys are discussed in depth by:

Korn, E. L. and B. I. Graubard (1999). *Analysis of Health Surveys.* New York: John Wiley & Sons Inc.

Parametric models for health care costs draw on the theory of size distributions such as the lognormal and generalized gamma. These and other size distributions are given a comprehensive treatment in:

Kleiber, C. and S. Kotz (2003). *Statistical Size Distributions in Economics and Actuarial Sciences.* New York: John Wiley & Sons Inc.

A classic text for generalized linear models is:

McCullagh, P. and J. A. Nelder (1989). *Generalized Linear Models. Second Edition*. Boca Raton: Chapman and Hall.

The application of GLMs in Stata is described in:

Hardin, J. W. and J. M. Hilbe (2007). *Generalized Linear Models and Extensions. Second Edition.* College Station Texas: Stata Press.

References

Ai, C. and E. C. Norton (2000). 'Standard errors for the retransformation problem with heteroscedasticity.' *Journal of Health Economics,* 19: 697-718.

Ash, A. S., R. P. Ellis, G. Pope, M. S. John, J. Z. Ayanian, D. W. Bates, H. Burstin, L. I. Iezzoni, E. McKay and W. Yu (2000). 'Using diagnoses to describe populations and predict costs.' *Health Care Financing Review,* 21: 7-28.

Atella, V., F. Brindisi, P. Deb and F. C. Rosati (2004). 'Determinants of access to physician services in Italy: a latent class seemingly unrelated probit approach.' *Health Economics,* 13: 657-68.

Atella, V., F. Peracchi, D. Depalo, and C. Rossetti (2006). 'Drug compliance, co-payment and health outcomes: evidence from a panel of Italian patients.' *Health Economics,* 15: 875-92.

Bago d'Uva, T. (2006). 'Latent class models for utilisation of health care.' *Health Economics,* 15: 329-43.

Basu, A., B. V. Arondekar and P. J. Rathouz (2006). 'Scale of interest versus scale of estimation: Comparing alternative estimators for the incremental costs of a comorbidity.' *Health Economics,* 15: 1091-107.
Basu, A. and W. G. Manning (2009). 'Issues for the next generation of health care cost analyses.' *Medical Care,* 47: S109-S114.

Basu, A., W. G. Manning, and J. Mullahy (2004). 'Comparing alternative models: log vs Cox proportional hazard?' *Health Economics,* 13: 749-65.

Basu, A. and P. J. Rathouz (2005). 'Estimating marginal and incremental effects on health outcomes using flexible link and variance function models.' *Biostatistics,* 6: 93-109.

Blough, D. K., C. W. Madden and M. C. Hornbrook (1999). 'Modeling risk using generalized linear models.' *Journal of Health Economics,* 18: 153-71.

Box, G. E. P. and D. R. Cox (1964). 'An analysis of transformations.' *Journal of the Royal Statistical Society B,* 26: 211-252.

Breusch, T. S. and A. R. Pagan (1979). 'Simple test for heteroscedasticity and random coefficient variation.' *Econometrica,* 47: 1287-1294.

Briggs, A., R. Nixon, S. Dixon, and S. Thompson (2005). 'Parametric modelling of cost data: some simulation evidence.' *Health Economics,* 14: 421-28.

Buntin, M. B. and A. M. Zaslavsky (2004). 'Too much ado about two-part models and transformation?: comparing methods of modeling Medicare expenditures.' *Journal of Health Economics,* 23: 525-42.

Cameron, A. C. and P. K. Trivedi (2009). *Microeconometrics Using Stata.* College Station Texas: Stata Press.

Cantoni, E. and E. Ronchetti (2006). 'A robust approach for skewed and heavy-tailed outcomes in the analysis of health care expenditures.' *Journal of Health Economics,* 25: 198-213.

Carroll, R J. and D. Rupert (1988). *Transformations and Weighting in Regression.* New York: Chapman and Hall.

Chalkley, M. and C. Tilley (2006). 'Treatment intensity and provider remuneration: dentists in the British national health service.' *Health Economics,* 15: 933-46.

Chaze, J. P. (2005). 'Assessing household health expenditure with Box-Cox censoring models.' *Health Economics,* 14: 893-907.

Chiou, J-M. and H-G. Müller (1998). 'Quasi-likelihood regression with unknown link and variance functions.' *Journal of the American Statistical Association,* 93: 1376-1387.

Conway, K. S. and P. Deb (2005). 'Is prenatal care really ineffective? Or, is the 'devil' in the distribution?' *Journal of Health Economics,* 24: 489-513.

Copas, J. B. (1983). 'Regression, prediction and shrinkage.' *Journal of the Royal Statistical Society B,* 45: 311-354.

Cox, D. R. (1972). 'Regression models and life tables.' *Journal of the Royal Statistical Society B,* 34: 187-200.

Deb, P. (2001). 'A discrete random effects probit model with application to the demand for preventive care.' *Health Economics,* 10: 371-83.

Deb, P. and J. F. Burgess Jr. (2007). 'A quasi-experimental comparison of statistical models for health care expenditures.' Mimeo.

Deb, P. and A. M. Holmes (2000). 'Estimates of use and costs of behavioural health care: a comparison of standard and finite mixture models.' *Health Economics,* 9: 475-89.

Deb, P. and P. K. Trivedi (1997). 'Demand for medical care by the elderly: a finite mixture approach.' *Journal of Applied Econometrics,* 12: 313-36.

Deb, P. and P. K. Trivedi (2002). 'The structure of demand for health care: latent class versus two-part models.' *Journal of Health Economics,* 21: 601-25.

Dranove, D., D. Kessler, M. Mcclellan, and M. Satterthwaite (2003). 'Is more information better? The effects of 'report cards' on health care providers.' *Journal of Political Economy,* 111: 555-88.

Duan, N. (1983). 'Smearing estimate: a nonparametric retransformation method.' *Journal of the American Statistical Association,* 78: 605-10.

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

Dusheiko, M., H. S. E. Gravelle and R. Jacobs (2004). 'The effect of practice budgets on patient waiting times: allowing for selection bias.' *Health Economics,* 13: 941-58.

Dusheiko, M., H. S. E. Gravelle, R. Jacobs and P. C. Smith (2006). 'The effect of financial incentives on gatekeeping doctors: evidence from a natural experiment.' *Journal of Health Economics,* **25**: 449-78.

Dusheiko, M., H. S. E. Gravelle, N. Yu and S. Campbell (2007). 'The impact of budgets for gatekeeping physicians on patient satisfaction: evidence from fundholding.' *Journal of Health Economics,* 26: 742 - 62.

Ellis, R. P. and P. G. Mookin (2008). 'Cross-validation methods for risk adjustment models.' Mimeo, Boston University.

Farsi, M. and G. Ridder (2006). 'Estimating the out-of-hospital mortality rate using patient discharge data.' *Health Economics,* 15: 983-95.

Gilleskie, D. B. and T. A. Mroz (2004). 'A flexible approach for estimating the effects of covariates on health expenditures.' *Journal of Health Economics,* 23: 391-418.

Godfrey, L. G. (1978). 'Testing for multiplicative heteroscedasticity.' *Journal of Econometrics,* 8: 227-236.

Gourieroux, C. S., A. Monfort and A. Trognon (1984). 'Pseudo maximum likelihood methods: theory.' *Econometrica,* 52: 680-700.

Gravelle, H. S. E., M. Sutton, S. Morris, F. Windmeijer, A. Leyland, C. Dibben and M. Muirhead (2003). 'Modelling supply and demand influences on the use of health care: implications for deriving a needs based capitation formula.' *Health Economics,* 12: 985-1004.

Hill, S. C. and G. E. Miller (2009). 'Health expenditure estimation and functional form: application of the generalized gamma and extended estimating equation models.' *Health Economics,* in press, DOI: 10.1002/hec.1498.

Ho, V. (2002). 'Learning and the evolution of medical technologies: the diffusion of coronary angioplasty.' *Journal of Health Economics,* 21: 873-85.

Hosmer, D. W. and S. Lemeshow (1980). ‘Goodness of fit tests for the multiple logistic regression model.’ *Communications in Statistics – Theory and Methods*, 9: 1043-1069.

Hosmer, D. W. and S. Lemeshow (1995). *Applied Logistic Regression. Second edition.* New York: Wiley.

Jenkins, S. P. (2009). 'Distributionally-sensitive inequality indices and the GB2 income distribution.' *The Review of Income and Wealth*, 55: 392-398.

Jiménez-Martin, S., J. M. Labeaga and M. Martínez-Granado (2002). 'Latent class versus two-part models in the demand for physician services across the European Union.' *Health Economics*, 11: 301-21.

Jones, A. M. (2000). 'Health Econometrics'. In Culyer, A. J. and J. P. Newhouse (eds), *Handbook of Health Economics.* Amsterdam: Elsevier.

Jones, A. M. (2007). *Applied Econometrics for Health Economists: A Practical Guide.* Oxford: Radcliffe Medical Publishing.

Jones, A. M. (2009). 'Panel data methods and applications to health economics'. in Mills, T. C. and K. Patterson (eds) *Palgrave Handbook of Econometrics. Volume 2*. London: Palgrave MacMillan.

Jones, A. M., N. Rice, T. Bago d'Uva, and S. Balia (2007). *Applied Health Economics.* London: Routledge.

Koenker, R. (1981). 'A note on studentizing a test for heteroscedasticity'. *Journal of Econometrics*, 17: 107-112.

Lee, M.-C. and A. M. Jones (2004). 'How did dentists respond to the introduction of global budgets in Taiwan? An evaluation using individual panel data.' *International Journal of Health Care Finance and Economics,* 4: 307-26.

Lee, M.-C. and A. M. Jones (2006). 'Heterogeneity in dentists' activity in Taiwan: an application of quantile regression.' *Empirical Economics*, 31: 151-64.

Manning, W. (1998). 'The logged dependent variable, heteroscedasticity, and the retransformation problem.' *Journal of Health Economics,* 17: 283-95.

Manning, W. (2006). 'Dealing with skewed data on costs and expenditure.' In Jones, A.M. (ed) *The Elgar Companion to Health Economics*. Cheltenham: Edward Elgar.

Manning, W. G., A. Basu and J. Mullahy (2005). 'Generalized modeling approaches to risk adjustment of skewed outcomes data.' *Journal of Health Economics,* 24: 465-88.

Manning, W.G., N. Duan and W.H. Rogers (1987a). 'Monte Carlo evidence on the choice between sample selection and two-part models.' *Journal of Econometrics*, 35: 59-82.

Manning, W., J. P. Newhouse, N. Duan, E. Keeler, A. Leibowitz and M. S. Marquis (1987b). 'Health insurance and the demand for medical care: evidence from a randomized experiment.' *American Economic Review*, 77: 251-77.

Manning, W. G. and J. Mullahy (2001). 'Estimating log models: to transform or not to transform?' *Journal of Health Economics*, 20: 461-94.

Martin, S., N. Rice, R. Jacobs and P. C. Smith (2007). 'The market for elective surgery: joint estimation of supply and demand.' *Journal of Health Economics*, 26: 263 - 85

Montez-Roth, M., C. L. Christiansen, S.L. Ettner, S. Loveland and A. K. Rosen (2006). 'Performance of statistical models to predict mental health and substance abuse cost.' *BMC Medical Research Methodology*, 6: 53. DOI: 10.1186/1471-2288-6-53.

Mullahy, J. (1998). 'Much ado about two: reconsidering retransformation and the two-part model in health econometrics.' *Journal of Health Economics*, 17: 247-81.

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.

Park, R. E. (1966). 'Estimation with heteroscedastic error.' *Econometrica*, 34: 888.

Parker, S. C. (1999). 'The generalised beta as a model for the distribution of earnings.' *Economics Letters*, 62: 197-200.

Pregibon, D. (1980). 'Goodness of link tests for generalized linear models.' *Applied Statistics*, 29: 15-24.

Propper, C., S. Burgess and K. Green (2004). 'Does competition between hospitals improve the quality of care: hospital death rates and the NHS internal market.' *Journal of Public Economics*, 88: 1247-72.

Propper, C., B. Croxson and A. Shearer (2002). 'Waiting times for hospital admissions: the impact of GP fundholding.' *Journal of Health Economics*, 21: 227-52.

Propper, C., J. Eachus, P. Chan, N. Pearson and G. D. Smith (2005). 'Access to health care resources in the UK: the case of care for arthritis.' *Health Economics*, 14: 391-406.

Ramsey, J. B. (1969). 'Tests for specification errors in classical linear least squares regression analysis.' *Journal of the Royal Statistical Society B*, 31: 350-370.

Rice, N., P. Dixon, D. Lloyd and D. Roberts (2000). 'Derivation of a needs based capitation formula of allocating prescribing budgets to health authorities and primary care groups in England: regression analysis.' *British Medical Journal*, 320: 284 -88.

Seshamani, M. and A. Gray (2004). 'Ageing and health care expenditure: the red herring argument revisited.' *Health Economics*, 13: 303-14.

Smith, P. C., N. Rice and R. Carr-Hill (2001). 'Capitation funding in the public sector.' *Journal of the Royal Statistical Society A*, 164: 217-257.

Van de Ven, W. and R. P. Ellis (2000). 'Risk adjustment in competitive health plan markets.' In A. J. Culyer and J. P. Newhouse (eds), *Handbook of Health Economics*. Amsterdam: Elsevier.

Veazie, P. J., W. G. Manning and R. L. Kane (2003). 'Improving risk adjustment for Medicare capitated reimbursement using nonlinear models.' *Medical Care*, 41: 741-752.

Wedderburn, R. W. M. (1974). 'Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method.' *Biometrika*, 61: 439-447.

Welsh, A. H. and X. H. Zhou (2006). 'Estimating the retransformed mean in a heteroscedastic two-part model.' *Journal of Statistical Planning and Inference*, 136: 860-81.

White, H. (1980). 'A heteroscedasticity-consistent covariance matrix estimator and a direct test for heteroscedasticity.' *Econometrica*, 48: 817-838.

Zhou, X-H. H. Lin and E. Johnson (2009). 'Non-parametric heteroscedastic transformation models for skewed data with an application to health care costs.' *Journal of the Royal Statistical Society B*, 70: 1029-1047.