Describing, explaining or predicting mental health care costs: a guide to regression models

Methodological review

GRAHAM DUNN, MASSIMO MIRANDOLA, FRANCESCO AMADDEO and MICHELE TANSELLA

Background Analysis of the patterns of variation in health care costs and the determinants of these costs (including treatment differences) is an increasingly important aspect of research into the performance of mental health services.

Aims To encourage both investigators of the variation in health care costs and the consumers of their investigations to think more critically about the precise aims of these investigations and the choice of statistical methods appropriate to achieve them.

Method We briefly describe examples of regression models that might be of use in the prediction of mental health costs and how one might choose which one to use for a particular research project.

Conclusions If the investigators are primarily interested in explanatory mechanisms then they should seriously consider generalised linear models (but with careful attention being paid to the appropriate error distribution). Further insight is likely to be gained through the use of two-part models. For prediction we recommend regression on raw costs using ordinary least-square methods. Whatever method is used, investigators should consider how robust their methods might be to incorrect distributional assumptions (particularly in small samples) and they should not automatically assume that methods such as bootstrapping will allow them to ignore these problems.

Declaration of interest None.

Analysis of the pattern of variation in individuals' health care costs and the determinants of these costs (including treatment differences) is an increasingly important aspect of research into the performance of mental health services. Econometric modelling (Kennedy, 1998; Verbeek, 2000; Jones & O'Donnell, 2002; Wooldridge, 2003) is a rather specialised activity within mental health research and, for obvious reasons, is not covered (at least not in sufficient detail) by standard textbooks on medical statistics (e.g. health care costs get only two very brief entries in Armitage *et al*, 2002). The present review aims to fill this gap. It is our intention neither to give a detailed picture of how and when each of these methods has been used in the mental health literature nor to appraise the quality of any particular applications. Given the inevitable space limitations of such a review we will not dwell on many of the technical details but will give a brief summary of many of the methods that are available and indicate how and when they might be useful. One important area of health economics that might not appear to have much in common with health econometrics is the analysis of incremental cost-effectiveness ratios using data from randomised controlled trials. Incremental cost-effectiveness ratios cannot be analysed using regression-based methods but Hoch *et al* (2002) recently have illustrated how a new approach to cost-effectiveness analysis (based on the net benefit framework; Stinnett *et al*, 1998; Tambour *et al*, 1998) can lead to the effective use of econometric modelling.

Background

The purpose of this review is to enhance readers' ability to understand and appraise research papers and other reports on the prediction of mental health care costs, paying particular attention to the statistical methodology, in terms of choice of model, and to evaluation of the likely future performance of the chosen predictive model. Although we would not expect the typical reader of this journal to be fully aware of the technical pitfalls of analysing costs data, in our view it is vital that, as in the critical appraisal of other research evidence, readers are familiar with the main issues and how the authors' interpretations of the results of such studies might be misleading or mistaken. Whenever possible, we wish to be able to make our own judgements as to the quality of a piece of research rather than having to take the views of 'experts' on trust. Other topics, such as methods of patient selection and methodological problems concerning measurement of the actual costs of care for individual patients, are extremely important but we will not attempt to discuss these in detail here. Many of the problems concerning the selection of patients to study are similar to those that are the usual concerns of anyone wishing to make a critical appraisal of prognosis studies and we therefore refer readers to the relevant literature in this area (Sackett *et al*, 1991).

Our own interest in the appraisal of the validity of many past studies of health care costs and a recent review by Diehr *et al* (1999) have prompted us to question whether the methods currently available for modelling or predicting health care costs, other than ordinary least-squares regression of logged costs, are widely known in the mental health field. We are not aware of an elementary discussion of the relevant methodologies but there is a useful study illustrating most of the methodological problems in the context of analysis of variation in mental health care costs (Kilian *et al*, 2002). Although it covers most of the same ground as the present paper, the discussion by Kilian *et al* is technically more difficult than the one presented here. The goal of this review is to make these methods more widely accessible to non-specialists and, in particular, to the consumers of the resulting research findings.

Our intention is to describe and explain the competing methods as clearly as possible while keeping the technical details to the minimum necessary for this objective. We will use little mathematics, restricting most of it to the definition of the various indices of the predictive power of the competing models. We hope that the present review can be read and understood by clinicians and other mental health workers, although we would hope that it might also provide a good starting point for

https://doi.org/10.1192/bjp.183.5.398 Published online by Cambridge University Press

statisticians and health economists who do not have experience or specialist knowledge of econometric modelling.

DEFINING THE GOALS OF THE STUDY

In reading papers on the prediction of mental health costs, one of the striking conclusions made concerns the frequent lack of clarity in the authors' aims and this lack of clarity arises from the vague way in which they deal with the concept of prediction. 'Predictive power' refers to a model's ability to discriminate between patients and to account for their cost differences. Sometimes the authors are content simply to describe differences in the health care costs of different patient groups, usually also reporting the results of simple significance tests of group differences. This is often accompanied by the use of some sort of regression model that can be used to 'explain' or account for the variation of costs within and between these patient groups. Finally (but very rarely is this made explicit) is the aim of being able to predict or forecast the costs of future patients (either individually or collectively). More often than not, authors fail to distinguish between explanatory models and those used for forecasting, accordingly giving very little thought to which statistical technique or group of techniques might be optimal for a given goal. It is possible, and frequently likely, that authors have several related aims in the presentation and analysis of their data, but it would be very helpful for the reader if they could be more precise in explaining exactly what they are.

To summarise, goals need to be defined precisely and the statistical methods should be chosen to fulfil these goals. A model and corresponding fitting method might be optimal for one particular goal but not the most effective for another. The optimal choice of methodology should be dependent upon the authors' chosen (and explicitly stated) aims. A given statistical model might be good as an explanatory device but poor as a tool for forecasting, or vice versa. In practice, however, the choice of statistical model might not matter too much (i.e. the results of the analysis are fairly insensitive or robust to model choice) but, again, both the authors and readers of studies of health costs need to know whether this is likely to be the case.

DISTRIBUTIONAL PROPERTIES OF COST DATA

Cost data are virtually always highly positively skewed and (at least in the context of the investigations discussed here – this would not be true for the analysis of net benefits, for example) cannot have negative values (zero values are possible but, in practice, it is also unlikely that a patient will incur exactly zero cost). Another characteristic of this type of data is that the variance of the observations increases with their mean (an example of heteroscedasticity as opposed to homoscedasticity, the latter implying a constant variance). It is also possible to get what is called censoring (incomplete or variable follow-up). In this situation the data collection stops before some or all of the patients have incurred their full health care costs, so that all we know is that the observed cost is the minimum that has been incurred by a given patient but the exact amount is unknown. Censoring is not a problem that is unique to cost data. It is likely to be more familiar to readers in the context of the analysis of times to certain events (times to recovery, relapse or death are three common examples). Examples of censoring occur when patients are lost to follow-up prior to the end of the data collection period, or, if the cost of an episode of illness is the variable of interest, termination of the follow-up period prior to the end of the patient's episode of illness. Another possible example is an incomplete measurement of health care costs arising from one or two components of the cost incurred by a given patient being missing from the data file. Here, again, we know the minimum cost incurred for that patient (the sum of the non-missing component costs) but not the total (the sum of the non-missing and missing components). Discussion of censored data is beyond the scope of the present paper and we refer interested readers to Diehr *et al* (1999) and to discussions of survival analysis (see Armitage *et al*, 2002).

A given population (or sample) of patients can often be thought of as a mixture of two types. First, there are those who will incur little, if any, treatment costs: those that attend for assessment, advice or brief support but do not need access to long-term care. They may have only a very minor problem or one that is acute but from which they make a quick and full recovery. Second, there are patients who need varying but non-trivial amounts of treatment and long-term care. These are the patients who may incur quite modest yearly health care costs but need very expensive long-term care and support. Thus, the first question faced by the statistical modeller, whether interested in explanation or forecasting, is whether to try to take this heterogeneity of the patients (i.e. the group structure) into account. Do we use a one-part model or is it better to use a two-part model? Before trying to answer this question we first need to describe what the two types of model are. We also need a more general discussion on the choice of regression models.

ONE- OR TWO-PART MODELS?

At the first stage of a two-part model we try to discriminate between the two patient types, that is, we try to predict who will incur substantial costs (group A, say) as opposed to those who will cost little or nothing (group B). Typically, this will be carried out using a multiple logistic regression. At the second stage we drop group B patients from the analysis and then try to model the incurred costs in those patients who are in group A. Patient characteristics that distinguish groups A and B might, or might not, be the same as those that appear to be responsible for the variations in the costs of those in group B. If the aim is to predict (forecast) the total cost for a given patient, for example, then this is equal to the sum of two components. The first is the product of the probability of being in group A and the modelled (expected) cost if the patient is in group A. The second component is the product of the probability of being in group B and the modelled cost if the patient is in group B. In symbols, this is

$$E(\mathrm{Cost}|\mathrm{X}) = P(\mathrm{A}|\mathrm{X})\cdot E(\mathrm{Cost}|\mathrm{A,X}) \\ +P(\mathrm{B}|\mathrm{X})\cdot E(\mathrm{Cost}|\mathrm{B})$$

where $P(\ )$ is the modelled probability from stage 1, $E(\ )$ is the expected or predicted value from stage 2, | means 'given' or 'conditional upon' and X is an indicator of the observed characteristics of the patient; $E(\mathrm{Cost}|\mathrm{B})$ is simply the average cost for those patients in group B. If we were concerned with predicting treatment costs (as opposed to the total cost to the health service, say) and group B patients are those who do not receive treatment, then $E(\mathrm{Cost}|\mathrm{B})$ would be zero.

In a one-part model we use a single regression equation to model the costs for

https://doi.org/10.1192/bjp.183.5.398 Published online by Cambridge University Press

everyone in the data-set (i.e. we do not first separate groups A and B). The predicted cost for a given patient with characteristics X is then simply $E(\text{Cost}|\text{X})$.

We will assume that the investigator has a clear idea of how to distinguish ‘substantial’ from ‘little or nothing’ costs based on his or her knowledge of the population being sampled. But what if it is not at all obvious what the boundary between the two groups might be? What if we are convinced that the population is made up of the two groups A and B but have difficulty assigning group membership to many of the individual patients? It may not be at all clear what the cost cut-off should be in order to discriminate between the two. In this case we might wish to postulate a more subtle version of a two-part model in which group membership remains latent or hidden. This type of model is called a latent class or finite mixture model in the statistical literature. We do not pursue this idea further here but refer the interested reader to Deb & Holmes (2002) for an illustrative example and methodological discussion.

The two-part model (or possibly a model with more than two parts; see Duan *et al*, 1983) is conceptually much richer than the simpler one-part model. For this reason it is likely to provide more insight concerning the ways in which costs arise. Diehr *et al* (1999) comment:

> ‘When the goal is understanding the system, a two-part model seems best because it permits the investigator to distinguish factors that affect the propensity to use any services from factors that affect volume of utilisation once the person has entered the system. . . For understanding the effect of individual covariates on total costs, a one-part model is most useful because it generates a single regression coefficient for each variable and so can be interpreted easily’.

We will defer discussion on accuracy of forecasts until later. Before moving on, however, it should be noted that an intelligent data analyst is likely to make a decision concerning the use of a one-part or two-part model at least partly on the basis of his or her prior knowledge concerning the heterogeneity of the population of patients under study and also from the way the sample of patients for analysis has been chosen. The analyst may have deliberately selected a relatively homogeneous subsample of patients prior to any further statistical analyses.

Having chosen which of the two approaches to use, we are still faced with the problem of how to choose an appropriate regression model for either total costs (one-part model) or costs in those that enter the system (two-part model). This is the subject of the following section. Readers wishing to read more on two-part modelling are referred to Duan *et al* (1983, 1984), Mullahy (1998) and the review of Diehr *et al* (1999).

CHOICE OF REGRESSION MODEL

The simplest approach is to model the observed costs directly using multiple regression; the fitting is done using the amiliar ordinary least-squares algorithm. Multiple regression, however, assumes that the effects of the predictive factors are additive. Furthermore, ordinary least squares is not the optimal fitting method (in the sense of producing parameter estimates with maximum precision) when the distribution of the errors (differences between observed and modelled costs) has a non-constant variance (heteroscedasticity). The latter characteristics of the data, together with non-normality, will also invalidate tests of significance associated with the model-fitting process, and estimates of the standard errors and confidence intervals for the parameters. Ordinary least-squares modelling of raw cost data – based on invalid distributional assumptions – can (and does) also produce invalid (i.e. negative) estimates of costs for some patients. It is not surprising, then, that investigators might be tempted to use methods other than ordinary least-squares modelling of raw cost data (but see below).

If one takes logarithms of the observed cost data, this transformation usually will have two consequences: a considerable reduction in the skewness of the data, although complete symmetry is unlikely to be achieved in practice; and stability of the variance (i.e. the variability of the observed costs will not increase with their mean). Both of these consequences lead to better performance of ordinary least-squares regression methods. Examples of the use of this approach can be found in Amaddeo *et al* (1998) and Bonizzato *et al* (2000). The method is (usually) implicitly based on a multiplicative model for the actual costs (including a multiplicative error term). There is a problem if there are observed costs of zero (the logarithm of zero is undefined) but this is often remedied by adding a small constant (unity, for example) prior to the logarithmic transformation. The method seems to work satisfactorily in practice but one should always remember that the aim of the analysis is to evaluate our ability to predict actual costs and not their logarithms. Values of R^2 and other indices of concordance of observed and predicted values (see below) must be evaluated using the observed and predicted costs (not their logarithms). More importantly, investigators should be aware of the fact that, even though ordinary least-squares methods produce unbiased estimates of log-costs, the predicted actual costs (and also total costs derived from the individual predictions) will be biased. They will underestimate the true cost. However, bias-reduction methods are available (e.g. the non-parametric method called ‘smearing’; see Duan, 1983) so this underestimation is not a serious problem as long as it is recognised by the investigator.

If the investigator really believes that the relationship between the predictive factors and cost is multiplicative, then it is probably preferable to model this explicitly using an appropriate generalised linear model. In a generalised linear model, the familiar regression equation of the form $\alpha+\Sigma\beta_i x_i$ is called the ‘linear predictor’. But the linear predictor is not necessarily equated with the expected cost, as in multiple regression with the raw data, but via ‘link function’. So, for example, we could have a model in which the natural logarithm of the expected costs is equated with the linear predictor

$$\log_e[E\text{Cost}|\text{X})] = \alpha + \beta_i x_i$$

or, equivalently

$$\begin{aligned} E(\text{Cost}|\text{X}) &= \exp[\alpha + \Sigma\beta_i x_i] \\ &= \exp[\alpha]\exp[\beta_1 x_1]\exp[\beta_2 x_2]\ldots \end{aligned}$$

where ‘exp’ indicates exponentiation (taking antilogarithms). The form of the model in the final line should make it clear why it is a multiplicative model. This generalised linear model differs from a multiple regression with logged cost data, however. The first difference is that the errors are assumed to be additive. That is

$$\begin{aligned} \text{Observed cost} &= E(\text{Cost}|\text{X}) + \text{error} \\ &= \exp[\alpha + \Sigma\beta_i x_i] + \text{error} \end{aligned}$$

The second difference is that there is a more realistic assumption concerning the probability distribution of the observed costs (taking into account that they are non-negative, their high degree of skewness

https://doi.org/10.1192/bjp.183.5.398 Published online by Cambridge University Press

and their heteroscedasticity). The cost data are usually assumed to follow a gamma distribution (potentially highly skewed) and the model is fitted by a method called maximum likelihood rather than ordinary least squares. The key similarity between the ordinary least-squares model for costs and this generalised linear model, however, is that in both we are explicitly modelling the raw costs themselves and not some arbitrary transformation of them. A relatively non-technical discussion of generalised linear models can be found in Everitt & Dunn (2001). If the investigator chooses to use a one-part model to explain the variation in total costs, then the generalised linear model with a log link (i.e. a log-linear model) and gamma errors is likely to provide the most realistic description of the data. For this purpose, ordinary least squares using raw costs would seem to be unrealistic (in terms of both additive effects and the distribution of the errors) and, apart from its simplicity and familiarity, ordinary least squares using logged costs does not appear to have any obvious attractions. Again, we defer forecasting until later. Recent examples of the use of generalised linear models in the analysis of mental health care costs can be found in Byford *et al* (2001), Chisholm & Knapp (2002) and Knapp *et al* (2003).

One very natural extension of the above log-linear generalised linear model is through the use of an 'offset'. Suppose that each patient provided cost data for a different number of years (let this variable be called 'Years'). Instead of modelling total costs, suppose that we were also interested in modelling costs per year

$$E(\mathrm{Cost}|\mathrm{X})/\mathrm{Years} = \exp[\alpha + \Sigma\beta_i x_i]$$

or

$$\log_e[E(\mathrm{Cost}|\mathrm{X})] = \log_e[\mathrm{Years}] + \alpha + \beta_i x_i$$

We still have the same log-linear model for costs but it now has an extra term, $\log_e[\mathrm{Years}]$, which is a fixed known constant for each patient. In the language of the generalised linear model, an explanatory variable that has a regression coefficient fixed at unity (rather than it being estimated from the data) is called an offset. Its use perhaps will be more familiar (particularly to epidemiologists) in the context of log-linear modelling of disease rates using the so-called person-years method (see Armitage *et al*, 2002). There is a close link between the use of offsets (person-years) in this way and the survival models in which one handles incomplete follow-up data via censoring. This link is also relevant to the analysis of incomplete or censored cost data (see above).

ASSESSING THE MODEL'S PERFORMANCE

Here we need an index or statistic to measure the concordance (agreement) between predicted and observed costs. Note that we are not, or should not be, interested in the concordance between predicted and observed log-costs.

Perhaps the simplest index is the familiar Pearson product-moment correlation (R) between predicted and observed costs (Zheng & Agresti, 2000), but this is far from ideal. It is a measure of association rather than concordance and it is probably better to use Lin's concordance coefficient (R_c; Lin, 1989) or an intraclass correlation (R_i; Dunn, 1989). But both of these indices, as well as the product-moment correlation, are dependent on patient heterogeneity – they will increase with increases in the variability of the costs, irrespective of the accuracy of the predictions. Perhaps the most commonly used index for a multiple regression model is the 'coefficient of determination' or 'proportion of variance explained', R^2 (equivalent in this situation to the square of the product-moment correlation between prediction and observation) – usually obtained from the analysis of variance table. But, again, this is not particularly useful unless the aim is to discriminate between patients. Like the above correlations, it is dependent on the heterogeneity of the observed costs. Despite this potential disadvantage, however, they are obviously useful for comparison of the performance of various models for the same data. Problems only arise when we try to compare the performance of predictive models on different groups. Some authors prefer to use what is called the adjusted R^2, R^2_a, where

$$R^2_a = 1 - [(1 - R^2)(n - 1)/(n - p)]$$

and n is the number of patients in the sample and p is the number of estimated parameters (including the intercept term). The idea is that the adjusted R^2 provides a better estimate of the likely performance of the model on future data-sets. Draper & Smith (1998) comment that

> 'The value of this statistic for the latter purpose is, in our opinion, not high; R^2_a might be useful as an initial gross indicator, but this is all'

(see the section on cross-validation below). The use of R^2_a instead of R^2, however, may lead to less overfitting because R^2_a is a penalised goodness-of-fit index that is dependent on the number of estimated parameters (p) in addition to the proportion of the total sum of squares explained. Unlike R^2, which cannot decrease as p increases (i.e. when a variable is added, the explained sum of squares will either increase or stay the same), the value of R^2_a can actually decrease when extra variables are added to the model (as in the case of overfitting; Greene, 2000).

The accuracy of a model's predictions is probably best evaluated by a function of the differences between the predicted and observed costs. That is, by a function of $(c_o - c_p)$, where c_o is the observed cost for a given patient and c_p is the corresponding prediction: $E(\mathrm{Cost}|\mathrm{X})$. The three obvious choices are the residual mean square (RMS), root-mean-square error (RMSE) and the mean of the absolute error (MAE). A less familiar index is Theil's U-statistic (Theil, 1966; Greene, 2000).

The RMSE is the square root of the mean of the squared differences between the predicted and observed values of cost, MAE is the mean of the absolute value of the differences, and RMS is the residual sum of squares divided by the residual degrees of freedom as obtained from the relevant analysis of variance table. The square root of the RMS (i.e. the standard deviation of the residuals) is likely to be close but not identical to the RMSE. Theil's U-statistic is the square root of the sum of the squared deviations of the predicted from the observed costs divided by the square root of the sum of the squared predictions. Algebraically, the less familiar of these indices are defined as follows

$$\begin{aligned}
&\mathrm{RMSE} = \surd[\Sigma(c_o - c_p)^2/n] \\
&\mathrm{MAE} = \Sigma[c_o - c_p] \\
&U = \surd[\Sigma(c_o - c_p)^2/\Sigma c_o^2]
\end{aligned}$$

Some authors use

$$U = \Sigma(c_o - c_p)^2/\Sigma c_o^2$$

In all four cases the addition (indicated by the symbol Σ) is over all patients in the sample, and a value of zero for the index indicates perfect prediction. Index U, like the various correlation coefficients and

https://doi.org/10.1192/bjp.183.5.398 Published online by Cambridge University Press

R^2, is a scale-free measure of concordance and shares the same advantages and pitfalls.

One potential problem, whatever indicator of performance is used, is that if it is used naively it is likely to be overoptimistic. If the explanatory variables in the final model have been chosen using the same data as those used to assess the model's performance, then we are likely to have capitalised on chance associations between potential explanatory variables and the cost outcomes and inevitably will have produced a model that has been overfitted (Greene, 2000). A more realistic evaluation of the performance of the model ideally should be made by cross-validation using a data-set collected from a second, independent sample of patients. Unfortunately, however, we often do not have adequate resources within a particular research project to be able to collect such a data-set, and if we test our model on someone else's data it is unlikely that they will have collected exactly the same information using the same measurement procedures on a comparable sample of patients. A more realistic option is to split our original sample into two, develop the model on one of the subsamples (the so-called training set) and evaluate it using the second one (the validation set). This split-sample or internal approach to cross-validation is the one advocated by Diehr *et al* (1999) and illustrated in Kilian *et al* (2002).

One pitfall of the split-half approach is its inefficient use of the data. Unless we have a very large sample to start with, we are usually loath to use only half of the patients to develop the model and half to test it. Ideally, we would like to maximise the use of the data for both functions. One approach is to take the full sample of n patients and leave each of the patients out in turn. Each time, we derive a model from the $n-1$ remaining patients and test its performance on the one that has been left out. This 'leave-one-out' procedure in principle involves a separate analyses from which we can then produce an overall summary of the model's performance. In practice this will not be necessary, but the technical details are beyond the scope of the present discussion. The text by Mosteller & Tukey (1977) contains a nice introduction to cross-validation methods and Armitage *et al* (2002: p. 395) provides a brief discussion of variants of the leave-one-out method (see also Picard & Berk, 1990).

HOW ROBUST ARE THE STATISTICAL METHODS?

Returning to the simple ordinary least-squares multiple regression models for observed costs, how can we be confident that inferences based on such a model are safe? We know because of the skewness (non-normality) and heteroscedasticity of the data that ordinary least-squares regression is not optimal. How does this affect the model's parameter estimates, their standard errors, P values, confidence intervals, etc.? Safe statistical inference for these models rests on the assumption that on repeated sampling the parameter estimates would be normally distributed. This is likely to be the case for large samples but frequently we have doubts about whether our sample is large enough. In the context of the analysis of cost data from a randomised trial, Barber & Thompson (2000*a*,*b*; see also Desgagné *et al*, 1998) have advocated the use of distribution-free procedures based on a resampling procedure called the bootstrap (Efron & Tibshirani, 1993). They claim that the bootstrap will provide robust inferences that are not dependent on distributional assumptions. They conclude that

> 'such bootstrap techniques can be recommended either as a check on the robustness of standard parametric methods, or to provide the primary statistical analysis when making inferences about arithmetic means for moderately sized samples of highly skewed data such as costs' (Barber & Thompson, 2000*b*).

The use of bootstrapping now appears to be commonplace in health economics studies, but is it the panacea that many health economists appear to believe it is?

Barber & Thompson's claims concerning the robustness of the inferences based on the bootstrap have been challenged recently by O'Hagan & Stevens (2003). They point out that for highly skewed cost data obtained from small samples of patients the sample mean is not the ideal estimator of the required population mean. It is very sensitive to the presence of one or two stragglers with relatively high costs, and inferences based on bootstrapping the sample mean will be equally affected by this problem. They argue that even when the methods advocated by Barber & Thompson are technically valid (in terms of their large sample properties), in small samples they may lead to inefficient and even misleading inferences. We suspect that this is likely to be an even greater problem for ordinary least-squares-based multiple regression models. O'Hagan & Stevens agree with Barber & Thompson's assertion that we should be concentrating on inferences on untransformed costs (as do we in the present paper), but their main message is that it is important to apply statistical methods (in the present context, model-fitting procedures) that recognise the skewness in cost data.

O'Hagan & Stevens (2003) advocate parametric modelling with realistic error structures. This does not, however, rule out the use of bootstrapping. Having chosen the model-fitting procedure to cope with the distributional characteristics of the data, we can use bootstrapping to obtain standard errors, confidence intervals, etc. O'Hagan & Stevens pursue Bayesian methods, but a viable alternative might be the use of robust model-fitting procedures. These are methods that are not unduly influenced by outlying or extreme observations (Mosteller & Tukey, 1977; Berk, 1990). Note that robust fitting methods should not be confused with robust methods of standard error estimation (the bootstrap, for example) once we have got our best-fitting model. They are complementary and should not be seen as competitors. A recent health economics application of robust model-fitting methodology can be found in Hoch *et al* (2002).

DISCUSSION

Choice of model

It is our experience and that of others (Diehr *et al*, 1999; Kilian *et al*, 2002) that, as a method of prediction (forecasting), a one-part model involving ordinary least-squares on raw costs data consistently performs as well as, if not better than, ordinary least squares on logged costs or the more theoretically satisfying log-linear generalised linear model. The former sometimes produces negative cost estimates but this is not a serious problem. We provisionally follow Diehr *et al* (1999) in recommending the use of ordinary least-squares regression with raw costs for this purpose. However, the use of methods that pay more attention to the distribution of the costs data, or the use of robust model-fitting algorithms, is likely to produce improvements over the use of ordinary least squares. If research workers are primarily interested in explanatory modelling and if they think that their model should be multiplicative, then they should seriously consider the use of a generalised linear model with a logarithmic link

https://doi.org/10.1192/bjp.183.5.398 Published online by Cambridge University Press

function (i.e. a log-linear model) with an appropriately specified error distribution. But even more value as an explanatory tool might be the use of two-part models. Bootstrapping is a very useful all-purpose and distribution-free method of obtaining standard errors, confidence and *P* values, but its use should not replace the careful thought that should be given to the choice of the type of model to be fitted and the optimum model-fitting algorithm to be used once the model type has been chosen. Bootstrapping comes later.

Assessing the performance of the model

We do not recommend the use of standardised indices such as R^2 or Theil's U-statistic to compare the performance of a model when applied to *different* groups. The apparent lack of predictive value for patients in one particular group (group 1), for example, as opposed to that in another (group 2) may simply be a statistical artefact caused by the fact that there is less variability in the costs for the patients in group 2. The performance of the forecasts (as measured by root-mean-square error or mean absolute error) may, in fact, be better in group 2 than in group 1. The main advantage of R^2 and Theil's U-statistic is to compare the performance of competing models within the *same* group of patients. For comparison of the performance of models on different groups, we recommend the use of the root-mean-square error or mean absolute error. Finally, we stress the importance of cross-validation – how well will the model perform in a future sample?

CLINICAL IMPLICATIONS

- This paper provides a relatively non-technical introduction to statistical regression models for mental health cost data for research workers, clinicians and mental health workers.
- Different models are described according to the goal of a study. A given statistical model might be good as an explanatory device but poor as a tool for forecasting or vice versa.
- In the analysis of mental health cost data, clear and easily interpretable indices of the performance of a model are proposed. Clinicians and health managers are interested in indices that measure the difference between predicted and observed costs, rather than in their concordance or association.

LIMITATIONS

- We have not dwelt on many of the technical details but give only a brief summary. We refer interested readers to other publications.
- The methods of selecting explanatory variables are not discussed.
- We pay only limited attention to cost-effectiveness data.

GRAHAM DUNN, Biostatistics Group, School of Epidemiology & Health Sciences, University of Manchester, UK; MASSIMO MIRANDOLA, Dr.Psychol., FRANCESCO AMADDEO, MD, MICHELE TANSELLA, MD, Department of Medicine and Public Health, Section of Psychiatry, University of Verona, Italy

Correspondence: Professor Graham Dunn, Biostatistics Group, School of Epidemiology & Health Sciences, Stopford Building, Oxford Road, Manchester M13 9PT, UK. E-mail: g.dunn@man.ac.uk

(First received 4 July 2002, final revision 15 May 2003, accepted 3 June 2003)

ACKNOWLEDGEMENTS

We thank two anonymous referees for their thorough but positive criticism of the previous version of this manuscript.

REFERENCES

Amaddeo, F., Beecham, J., Bonizzato, P., *et al* (1998) The costs of community-based psychiatric care for first-ever patients. A case register study. *Psychological Medicine*, **28**, 173–183.

Armitage, P., Berry, G. & Matthews, J. N. S. (2002) *Statistical Methods in Medical Research* (4th edn). Oxford: Blackwell.

Barber, J. & Thompson, S. G. (2000*a*) Analysis and interpretation of cost data in randomised controlled trials: review of published studies. *BMJ*, **317**, 1195–1200.

__ & __ (2000*b*) Analysis of cost data in randomized trials: an application of the non-parametric bootstrap. *Statistics in Medicine*, **19**, 3219–3236.

Berk, R. A. (1990) A primer in robust regression. In *Modern Methods of Data Analysis* (eds J. Fox & J. S. Long), pp. 292–324. Newbury Park, CA: Sage Publications.

Bonizzato, P., Bisoffi, G., Amaddeo, F., *et al* (2000) Community-based mental health care: to what extent are service costs associated with clinical, social and service history variables? *Psychological Medicine*, **30**, 1205–1215.

Byford, S., Barber, J. A., Fiander, M., *et al* (2001) Factors that influence the cost of caring for patients with severe psychotic illness. Report from the UK 700 trial. *British Journal of Psychiatry*, **178**, 441–447.

Chisholm, D. & Knapp, M. (2002) The economics of schizophrenia care in Europe: the EPSILON study. *Epidemiologia e Psichiatria Sociale*, **11**, 12–17.

Deb, P. & Holmes, M. (2002) Estimates of use and costs of behavioural health care: a comparison of standard and finite mixture models. In *Econometric Analysis of Health Data* (eds A. M. Jones & O. O'Donnell), pp. 87–99. New York: John Wiley.

Desgagné, A., Castilloux, A.-M., Angers, J. F., *et al* (1998) The use of the bootstrap statistical method for the pharmacoeconomic cost analysis of skewed data. *Pharmacoeconomics*, **13**, 487–497.

Diehr, P., Yanez, D., Ash, A., *et al* (1999) Methods for analyzing health care utilisation and costs. *Annual Review of Public Health*, **20**, 125–144.

Draper, N. R. & Smith, H. (1998) *Applied Regression Analysis* (3rd edn). New York: John Wiley.

Duan, N. (1983) Smearing estimate: a non-parametric retransformation method. *Journal of the American Statistical Association*, **78**, 605–610.

__, Manning, W. G., Jr., Morris, C. N., *et al* (1983) A comparison of alternative models for demand for medical care. *Journal of Business and Economic Statistics*, **1**, 115–126.

__, __, __, *et al* (1984) Choosing between the sample selection model and the multi-part model. *Journal of Business and Economic Statistics*, **2**, 283–289.

Dunn, G. (1989) *Design and Analysis of Reliability Studies*. London: Edward Arnold.

Efron, B. & Tibshirani, R. J. (1993) *An Introduction to the Bootstrap*. London: Chapman & Hall.

Everitt, B. S. & Dunn, G. (2001) *Applied Multivariate Data Analysis* (2nd edn). London: Edward Arnold.

Greene, W. H. (2000) *Economic Analysis* (4th edn). Englewood Cliffs, NJ: Prentice Hall.

Hoch, J. S., Briggs, A. H. & Willan, A. (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.

https://doi.org/10.1192/bjp.183.5.398 Published online by Cambridge University Press

Jones, A. M. & O'Donnell, O. (2002) *Econometric Analysis of Health Data*. New York: John Wiley.

Kennedy, P. (1998) *A Guide to Econometrics* (4th edn). Oxford: Blackwell Publishers.

Kilian, R., Matschinger, H., Löffler, W., *et al* (2002) A comparison of methods to handle skew cost variables in the analysis of the resource consumption in schizophrenia treatment. *Journal of Mental Health Policy and Economics*, **5**, 21–31.

Knapp, M., Chisolm, D., Leese, M., *et al* (2003) Comparing patterns and costs of schizophrenia in five European countries: The EPSILON Study. *Acta Psychiatrica Scandinavica*, **105**, 42–54.

Lin, L. I.-K. (1989) A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, **45**, 255–268.

Mosteller, F. & Tukey, J. W. (1977) *Data Analysis and Regression*. Reading, MA: Addison Wesley.

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

O'Hagan, A. & Stevens, J. W. (2003) Assessing and comparing costs: how robust are the bootstrap and methods based on asymptotic normality? *Health Economics*, **12**, 33–49.

Picard, R. R. & Berk, K. N. (1990) Data splitting. *American Statistician*, **44**, 140–147.

Sackett, D. L., Haynes, R. B., Guyatt, G. H., *et al* (1991) *Clinical Epidemiology: A Basic Science for Clinical Medicine*. Boston, MA: Little Brown.

Stinnett, A. A. & Mullahy, J. (1998) Net health benefits: a new framework for the analysis of uncertainty in cost-effectiveness analysis. *Medical Decision Making*, **18** (Pharmacoeconomics special issue), S68–S80.

Tambour, M., Zethraeas, N. & Johannesson, M. A. (1998) A note on confidence intervals in cost-effectiveness analysis. *International Journal of Technology Assessment in Health Care*, **14**, 467–471.

Theil, H. (1966) *Applied Economic Forecasting*. Amsterdam: North Holland.

Verbeek, M. (2000) *A Guide to Modern Econometrics*. New York: John Wiley.

Wooldridge, J. M. (2003) *Introductory Econometrics* (2nd edn). Madison, OH: South-Western College Publishing.

Zheng, B. & Agresti, A. (2000) Summarising the predictive power of a generalised linear model. *Statistics in Medicine*, **19**, 1771–1781.

https://doi.org/10.1192/bjp.183.5.398 Published online by Cambridge University Press