Regression models for analyzing costs and their determinants in health care: an introductory review

DARIO GREGORI, MICHELE PETRINCO, SIMONA BO, ALESSANDRO DESIDERI, FRANCO MERLETTI AND EVA PAGANO

Laboratories of Epidemiological Methods and Biostatistics, Department of Environmental Medicine and Public Health, University of Padova, Padova, Italy, Unit of Cancer Epidemiology, University of Torino and CPO Piemonte, Torino, Italy, Department of Internal Medicine, University of Torino, Torino, Italy, and Unit of Cardiology, Castelfranco Veneto Hospital, Padova, Italy

Address reprint requests to: Dario Gregori, Department of Environmental Medicine and Public Health, Via Loredan 18, 35121 Padova, Italy. Tel: +39-049-8275384; Fax: +39-02-700445089; E-mail: dario.gregori@unipd.it

Accepted for publication 12 March 2011

Abstract

Objectives. This article aims to describe the various approaches in multivariable modelling of healthcare costs data and to synthesize the respective criticisms as proposed in the literature.

Methods. We present regression methods suitable for the analysis of healthcare costs and then apply them to an experimental setting in cardiovascular treatment (COSTAMI study) and an observational setting in diabetes hospital care.

Results. We show how methods can produce different results depending on the degree of matching between the underlying assumptions of each method and the specific characteristics of the healthcare problem.

Conclusions. The matching of healthcare cost models to the analytic objectives and characteristics of the data available to a study requires caution. The study results and interpretation can be heavily dependent on the choice of model with a real risk of spurious results and conclusions.

Keywords: cost analysis, skewness, zero-cost, censoring

Introduction

The analysis of costs in clinical and public health care has become a standard part of both experimental and epidemiological research. This is motivated by the growing interest in controlling public expenditure, in view of adopting interventions or treatments on the basis of their cost-effectiveness [1]. The healthcare system is obviously very concerned about cost evaluation and control, and predictive models are needed to understand how costs behave as a function of given patients’ or centres’ characteristics. Recent papers have examined healthcare delivery in terms of interest of this paper, namely, the understanding of how costs are related to a set of given covariates [2, 3]. The relationship of costs to the underlying features of healthcare delivery can be investigated in very traditional fields, like drug utilization [4], or in even more qualitative outcomes, like quality of life or anxiety or depression scores [5]. The conceptualization of the cost process for a single patient or for a healthcare institution in terms of a stochastic phenomenon [6] is perhaps the major explanation for the growth of a vast body of research on appropriate statistical methods and models in this area. Put simply, cost analysis is mainly devoted to:

(i) finding appropriate methods to get the most accurate estimate of the mean costs of treating the disease (clinical setting);
(ii) testing for differences among two or more groups of treated patients (experimental setting); and
(iii) identifying the patients’/structure characteristics influencing costs and getting an estimate of expected costs, at a fixed point in time during follow-up or disease progression, for specific types of patients (cost profiling in observational settings).

Statistical methods appropriate for comparing new treatments or interventions in terms of their cost-effectiveness have been largely investigated [7–44] and agreed on [45, 46]. The development of methods suitable for modelling patients’ costs in terms of clinical or socio-demographic covariates has been more recent and less conclusive. This is mainly due to the lack of interest in the cost-prediction exercises, in spite of their great importance for activities planning and optimization purposes [47–50].
This article aims to describe some of the main criticisms in multivariable modelling of healthcare costs data and to synthesize the respective methodological approaches proposed in the literature. The described methods are applied to two datasets: one, an experimental setting and the other, an observational setting, respectively, referring to a problem of costs in acute myocardial infarction healthcare delivery and to a diabetes cohort of patients facing repeated hospitalization.

The diabetes cohort and the COSTAMI study

The ‘diabetes cohort’ is a retrospective study analysing the repeated hospitalization in a cohort of diabetic patients [51, 52].

A cohort of all type 2 diabetic patients, resident in the Piedmont Region, attending the Diabetic Clinic of the San Giovanni Battista Hospital in Turin (Piedmont Region, Italy) during 1995 and alive on 1 January 1996 was identified (n = 3892). Mortality and hospitalization follow-up was carried out up to 30 June 2000. The patients were included in the study if they had at least one hospitalization in the subsequent years of follow-up. The final dataset included 2550 patients. A total of 4816 ordinary and 2183 daily hospitalizations were observed during the 4.5 years of follow-up. About 13% of the sample (342 patients) died during the follow-up. Demographic data (age, gender) and clinical data relative to the year 1995 (years of diabetes and number of co-morbidities) were utilized for the present analysis. The costs (in Euros) for daily and ordinary hospitalizations were calculated referring to the Italian DRGs (Diagnosis Related Groups) system. These data were analysed in an earlier paper [52], where all details on the cohort can be found, and are here used as support for some methodological comments.

The COSTAMI (COST of strategies After Myocardial Infarction) study [53, 54] compared an early discharge strategy based on stress echocardiography (EDS) with standard care based on clinical evaluation (CE) and post-discharge exercise electrocardiography (ECG). Patients were randomized either to an early discharge strategy (Day 3–4) after a stress EDS or to CE with a traditional discharge after 7–9 days of observation. Direct medical costs were collected for the 458 enrolled patients during a 1-year follow-up. The main analysis [53] was done fitting a model for costs as a function of the strategy, adjusted for the presence of known factors increasing the likelihood of events in the follow-up: gender, age, ejection fraction (EF), presence of antero-lateral ECG modifications at admission and presence of diabetes. The distribution and effects of all the variables, as well as further details about the general structure of the trial and detailed information on the observed variables are illustrated in the main paper [53].

Models for costs

The estimation of total medical cost is not straightforward, particularly when the goal of the analysis is to relate costs to a specific pattern of covariates. Indeed, accurate cost estimation is problematic when cost records are incomplete, because censoring could lead to biased estimates of costs, unless appropriately accounted in the analysis [55].

The main problems in analysing costs are the following (for a short definition of the terms used, please refer to Table 1):

- (i) an asymmetry of the distribution (Fig. 1), due to a minority of subjects with high medical cost compared with the rest of the population;
- (ii) a possible large mass of observations with zero-cost;
- (iii) a common presence of censored observations not satisfying the condition of being independent and not informative—this condition is needed because the individuals still under observation must be representative of the population at risk in each group, otherwise...

---

**Table 1** Glossary of the main statistical terms used in the paper

<table>
<thead>
<tr>
<th>Term</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Skewness</td>
<td>Skewness is a measure of the asymmetry of the distribution of a variable. It can be positive or negative or zero. A negative skewness (a positive the opposite) indicates that the tail on the left side of the distribution is longer than the right side and the main part of the values (including the median) lie to the right of the mean. A zero value indicates that the values are relatively evenly distributed on both sides of the mean, and it is a characteristic of the normal distribution.</td>
</tr>
<tr>
<td>Heteroschedasticity</td>
<td>Heteroschedasticity occurs when the error variance is not homogenous. It results in inefficient estimators and biased standard errors, rendering the t-tests and confidence intervals unreliable.</td>
</tr>
<tr>
<td>Censoring</td>
<td>Censoring occurs when the value of an observation is only partially known. A typical example is the survival process, where survival is usually known only up to the end of the follow-up time, commonly shorter than survival if no final events occur. In this case the survival time is said to be (right) censored.</td>
</tr>
<tr>
<td>Zero-costs</td>
<td>Zero-costs indicate that according to the definition of cost adopted in the study (for example hospitalization costs), no actual costs have been recorded for that patient (e.g. because no hospitalizations occurred) and thus to the cost variable is given a value of zero.</td>
</tr>
</tbody>
</table>
the observed failure rate in each group will be biased; and

(iv) frequent violation of the assumption of proportional hazards, particularly when costs are accumulated at different rates [56]. This is clearly seen in the COSTAMI data accumulation process (Fig. 2).

With reference to these characteristics and particularly to the presence of censoring, several studies in the literature [55, 57] have proposed using survival models such as the Kaplan–Meier and the Cox regression model, based on the conceptual similarity between costs and time, both being continuous non-decreasing variables. However, the assumptions behind the survival models listed above are often violated in cost estimations.

The same definition of censoring is sometimes not homogeneous in the literature on regression models for cost data (Table 2). Apart from the misleading approach of not considering censoring in the analysis when present, two main definitions are adopted. In the first case, the costs are accumulation up to a pre-specified point in time. In this case, the death of the patient is seen as a failure in observing the cost accumulation. In the second case, a more standard definition is adopted, when people are ending the follow-up, thus interrupting their cost accumulation, without events [55].

The three main issues (skewness, zero-cost and censoring) highlighted above and regarding the models for cost data can be used to classify methods and models used to deal with them (Table 2).

**Studies with skewness-only problem**

Without censoring and zero-costs, the issue in modelling healthcare costs is basically an exercise of model fitting for highly skewed data. In cost analysis, things are complicated by the fact that additivity is a property of the model that is commonly considered essential: costs are indeed added together to build up the total expenditure. In this sense, the mean cost is a more meaningful measure than other more robust alternatives, like median or geometric mean cost. Nevertheless, the estimate of the mean cost, possibly as a function of a set of covariates, must be as robust as possible to outliers or asymmetries in the data distribution. In our diabetes data, the effect of skewness can be largely understood. As shown in Fig. 1, the cost distribution is very
The resulting estimate of its centre is 3913 € when it is based on the median, but it is dragged to the right by about 100% (up to 7013 €) when the same estimate is based on the mean. This is something to be borne in mind when considering the perhaps simplest model, the linear model, fitted either via ordinary least squares (OLS) or maximum likelihood (ML), which assumes the following form for the costs:

\[ c_i = \sum_{j=1}^{h} \beta_j x_j + e_i, \]

where \( c_i \) is the cost observed for each patient, and \( x_j \) are the \( j \) covariates and \( \beta_j \) are the corresponding regression coefficients estimated with the least squares method. Residuals are commonly assumed to be normally distributed. To reduce skewness in the residuals, several transformations of the response variable have been proposed in the literature. These transformations are commonly able to gain a reasonable normalization effect even in presence of highly skewed data. Notice that the approach of transforming the costs requires in any case a back-transformation at the moment of interpreting results, causing several additional problems [58], partially avoided using approaches like the ‘smearing’ estimator [59].

In the case of the Box–Cox transformation [60], which is one of the most used within the uniparametric transforms, with a parameter \( \lambda \)

\[
\frac{c_i^\lambda - 1}{\lambda} = \sum_{j=1}^{h} \beta_j x_j + e_i, \quad \text{if} \; \lambda \neq 0, \\
\log(c_i) = \sum_{j=1}^{h} \beta_j x_j + e_i, \quad \text{if} \; \lambda = 0,
\]

Although very powerful in obtaining a final symmetric distribution as shown in Fig. 3, where several choices of the \( \lambda \)

\begin{table}
\centering
\begin{tabular}{|l|l|l|l|}
\hline
Analysis & Censoring definition & Caveats \\
\hline
Administrative & Cost till death [76] & Only dead patients have complete follow-up history & Cost and survival are closely related \\
Loss at follow-up & Cost till death & Only dead patients have complete follow-up history & Possible informative censoring \\
Death censoring & Cost up to a pre-specified time [55] & Only patients alive at the end of follow-up are uncensored & Informative censoring \\
No-censoring & Observed costs & & Downward bias in cost estimation \\
\hline
\end{tabular}
\caption{Definitions of censoring as appearing in the literature regarding cost data}
\end{table}
Regression models for healthcare costs

parameter are fitted, normality is still assumed. In addition, when the back-transformation is performed, the bias eventually occurring

$$\frac{\partial \sigma_i^2(c_i)}{\partial c_i},$$
can be proven to be a function of the variance function. Thus, if heterogeneity is present, additional efficiency and inference problems are present on the transformed scale.

A particular case of the Box–Cox transform is the lognormal model, where the transformation is the $\ln$ function. Supposing to have two treatments $j = 0, 1$, it happens that $E(C_j) = \exp(\beta_1 + 0.5\sigma^2)$. Thus, a test of the null hypothesis of equality of means $H_0$: $\gamma_1 - \gamma_2 = 0$ is perhaps a test for the geometric means. This is clearly shown by observing that the null hypothesis on the expected values of the transformed costs $E(C)$ is

$$H_0 : \exp(\gamma_1 + 0.5\sigma^2) = \exp(\gamma_2 + 0.5\sigma^2),$$

which implies $H_0$: $\gamma_1 - \gamma_2 = 0$ only if $\sigma^2_1 = \sigma^2_2$. Again, this ends up with a change in the meaning of the test, because the test of interest (the equivalence of the mean costs in the original scale) is equivalent to the test performed on the transformed scale only in case of homogeneity of variances in the treatment groups.

To understand how the interpretation of coefficients changes, consider the simple case of a linear model relating costs with the number of co-morbidities, i.e. $\epsilon = \beta_1 + \beta_2N$, where $\epsilon$ are the costs and $N$ the number of co-morbidities. In this case, the estimates of the coefficients are 2155.15 (SE 1028.50) for the intercept and 3913 (SE 1028.50) for the slope. These coefficients are directly interpretable as ‘mean effects’. This is no longer true when the model is written as $\log(\epsilon) = \beta_1 + \beta_2N$, because in this case the slope coefficient is not an average effect.

A common solution to the issues of outliers is the so-called threshold model, where the probability of having costs in excess of a given threshold is modelled using commonly a logistic function. Two different cut-off points in the costs distribution are generally used: the median $q_2$ and the third quartile $q_{3}$, leading to the model

$$P(\epsilon > q_{3,5}) = \frac{1}{1 + \exp\left(-\sum_{j=1}^{\delta} \beta_j x_j\right)}.$$

The model estimates nothing but the probability of having a cost greater than the median (third quartile) as a function of the observed covariates. Although it does not require normality and it can also work for heavily skewed distributions, it does not give an estimate of the mean costs and the conclusions are sensitive to the threshold chosen, which is sample dependent when based on percentiles. Obviously, no information is provided for the behaviour of the covariates in affecting the cost distribution: again, estimating the simple model for the number of co-morbidities $P(\epsilon > 3913) = 1/1 + e^{(\beta_1 + \beta_2N)}$, where 3913 is the median cost, yields to the final estimate of $-2.102 + 0.539N$ and finally, for a person with two co-morbidities, to the estimate $P(\epsilon > 3913) = 0.735$. A vastly better solution, proven by simulation to be highly preferable [61], is based on the generalized linear model (GLM) [62], which solves the open issue in the response transformation that in general

$$g^{-1}(E(\epsilon)) \neq E(g^{-1}(\epsilon)),$$

by the transformation of the expectation [63]

$$g(E(\epsilon)) = \sum \beta_j x_j,$$

where the distribution for the response is usually taken to be Gamma() and the link function is usually taken: (i) for additive effects, as the identity function $I()$, and (ii) for multiplicative models as the log(), avoiding in this case the bias in the back-transformation [64]. This approach is highly flexible and can accommodate a variety of distributions maintaining linearity of the effects on the response scale. In addition, the choice of transformation can be completely parameterized [65], moving toward a data-driven modelling strategy.

Studies with zero-costs

The issue of presenting a possible large mass of observations with zero-costs is based both on model inadequacies and on conceptual issues. From the model side, the application of standard models, in particular the OLS-based ones, can lead to predictions based on a negative model. From the conceptual point of view, it could be highly questionable that two populations (one with zero, the other with positive costs) have the same behaviour with respect to the covariates.

A common solution widely adopted in the applied literature is to address the ‘model’ issue is to add a positive constant $k$ to the costs, modelling therefore the log($\epsilon + k$). Again, from the big advantage of being easy to implement, this solution has the drawbacks of having poor behaviour and an arbitrariness in the choice of $k$, of aggravating the problem of back-transformation [66], and of not taking into account different behaviours of patients with zero-costs.

To address the latest issue, an elegant approach is based on the concept of latent variable, where an underlying variable $z$ is introduced which can be modelled as $z = \beta_0 + x\beta + \epsilon$, but it is only observable as a binary outcome $y = 1$, if $z > 0$, and $y = 0$ if $z \leq 0$. To add flexibility in the modelling strategy, the Tobit model [67] does not require the observable variable to have a binary form, modelling the observed variable $x = \max(0, z)$, where $z = x\beta + u$, $u \sim \text{Norm}(0, \sigma^2)$. The Tobit model uses ML to estimate both $\beta$ and $\epsilon$ for this model. Table 3 shows the estimates for the diabetes cohort.

It is important to realize that $\beta$ estimates the effect of $x$ on $z$, the latent variable, not on the observed cost process $\epsilon$. In addition, unless the latent variable $z$ is of interest, it is just not possible to interpret the coefficient directly, because
\[ E(\epsilon|x) = \Phi(x\beta) + \phi(x\beta), \] so that the effect on the observed costs is very hard to interpret, being \( \partial E(\epsilon|x)/\partial x = \beta \Phi(x\beta). \) In addition, Tobit models have been shown to be very sensitive to departures from normality or homoscedasticity \[68\], which can be easily detected by a simple residuals plot, like that shown in Fig. 4, where normality is shown to hold.

A class of models that explicitly takes into account the different nature of the populations, the one with positive and the other with zero-cost, is the so-called mixed model \[69\], where the conditional expectation is partitioned as

\[ E(\epsilon|x) = P(\epsilon > 0)E(\epsilon|\epsilon > 0). \]

Thus, expectation is split in two parts, the first modelling the probability of any use or expenditure, based on the full sample and usually modelled via a logit or probit GLM, and the second modelling the actual level of expenditure conditionally to \( \epsilon > 0 \), usually via an appropriate model in the linear or GLM class. Estimates of mean costs thus are obtained using the smearing estimator, taking the mean of the exponentiated residuals \[66\]

\[ E(\epsilon|x) = \Phi(x\alpha) \exp(x\beta) \frac{1}{\varepsilon} \sum \exp(\ln(\epsilon) - x\beta). \]

The different effects of the same covariates on the two model parts are clearly seen in the diabetes cohort, where the first part has been modelled through a standard logit model and the second via a linear OLS model (Table 5).

The marginal effect on costs for a given covariate \( x_h \) can be calculated computing the partial derivatives of the marginal expectation

\[
\frac{\partial E(\epsilon)}{\partial x_h} = \frac{\partial [P(\epsilon > 0)E(\epsilon|\epsilon > 0)]}{\partial x_h} \\
= \left( \frac{P(\epsilon > 0) \frac{\partial E(\epsilon|\epsilon > 0)}{\partial x_h}}{P(\epsilon > 0)} \right) + \left( E(\epsilon|\epsilon > 0) \frac{\partial P(\epsilon > 0)}{\partial x_h} \right).
\]

In the case of the diabetes cohort, we have that \( P(\epsilon > 0) = 0.54 \) and \( E(\epsilon|\epsilon > 0) = 7509.82 \), implying that for the

---

**Table 3** Models used in cost-analyses

| Skewness | Zero-cost | Censoring | Mean estimation \( E(\epsilon|x) \) |
|----------|-----------|-----------|----------------------------------|
| Original scale models |
| OLS (\( c \)) | X | X | X |
| Tobit/adjusted Tobit | | X | X |
| GLM (gamma, log-gamma) | X | O | X |
| Transformed response |
| OLS log(\( c + k \)) | | O | O |
| Threshold logit models | X | X | X |
| Survival models |
| Parametric (Weibull) | X | X | X |
| Semiparametric (Cox Proportional hazard) | X | X | O |
| Non-parametric additive regression \[80\] | X | X | X |
| Mixed models | X | X | X |
| Weighted regression \[73, 86\] | X | X | X |

X = condition satisfied, O = condition partially satisfied, blank = condition not satisfied.

**Table 4** Tobit fit to diabetes cohort

<table>
<thead>
<tr>
<th>Value</th>
<th>Std. Error</th>
<th>( \tau )</th>
<th>( P )</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>5510.8</td>
<td>1474.4643</td>
<td>3.737</td>
</tr>
<tr>
<td>Age</td>
<td>16.94</td>
<td>22.2739</td>
<td>0.761</td>
</tr>
<tr>
<td>Gender (M vs. F)</td>
<td>-62.85</td>
<td>424.3257</td>
<td>-0.148</td>
</tr>
<tr>
<td>Years of diabetes</td>
<td>50.48</td>
<td>25.0192</td>
<td>2.018</td>
</tr>
<tr>
<td>Number of co-morbidities [21]</td>
<td>2134.09</td>
<td>605.1603</td>
<td>3.526</td>
</tr>
<tr>
<td>Log(scale)</td>
<td>9.09</td>
<td>0.0167</td>
<td>544.008</td>
</tr>
</tbody>
</table>

Figure 4 Residuals plot for the Tobit model for diabetes cohort (Table 4).
covariate ‘Year of diabetes’, the effect on the marginal expectation can be computed recalling (Table 5) that $\hat{b}_{\text{logit}} = 0.025$ and $\hat{b}_{\text{OLS}} = 49.83$, obtaining that the costs are increasing on average by €208 per each additional year of diabetes history. Note that two-part models can suffer from the problem of back-transformation if the model part for the continuous response is based on a transformation function, e.g. a logarithm transform [70]. To avoid this, the substitution of the OLS model part with a Gamma model is a valid option [71].

### Table 5 Mixed model fit to diabetes cohort

<table>
<thead>
<tr>
<th></th>
<th>Value</th>
<th>Std. Error</th>
<th>$t$-value</th>
<th>$p$-value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Logit model</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Intercept)</td>
<td>-2.17</td>
<td>0.23</td>
<td>-9.47</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Age</td>
<td>0.03</td>
<td>0.00</td>
<td>8.53</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Gender</td>
<td>0.11</td>
<td>0.07</td>
<td>1.60</td>
<td>0.05</td>
</tr>
<tr>
<td>Years of diabetes</td>
<td>0.02</td>
<td>0.00</td>
<td>5.84</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Number of co-morbidities $\geq 1$</td>
<td>0.69</td>
<td>0.11</td>
<td>6.45</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td><strong>OLS model</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Intercept)</td>
<td>5125.61</td>
<td>1428.88</td>
<td>3.59</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Age</td>
<td>28.02</td>
<td>21.33</td>
<td>1.31</td>
<td>0.09</td>
</tr>
<tr>
<td>Gender</td>
<td>483.89</td>
<td>413.26</td>
<td>1.17</td>
<td>0.12</td>
</tr>
<tr>
<td>Years of diabetes</td>
<td>49.83</td>
<td>24.24</td>
<td>2.06</td>
<td>0.02</td>
</tr>
<tr>
<td>Number of co-morbidities $\geq 1$</td>
<td>2596.41</td>
<td>566.67</td>
<td>4.58</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>

### Table 6 Estimates of the mean costs for the diabetes cohort according to the various definition of censoring adopted (Table 2)

<table>
<thead>
<tr>
<th></th>
<th>Mean estimate</th>
<th>SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lin estimate (administrative censoring)</td>
<td>5856</td>
<td>249</td>
</tr>
<tr>
<td>Cox estimate (death censoring at 4 years)</td>
<td>33896</td>
<td>1249</td>
</tr>
<tr>
<td>No-censoring estimatea</td>
<td>4488.18</td>
<td>129.44</td>
</tr>
</tbody>
</table>

*aGLM with Gamma distribution and log link function.

### Studies with censored data

To adjust for censoring, the basic idea is to weight the costs for the inverse of the probability of being alive, mimicking the basic Horvitz–Thompson estimator, where each observation is weighted by the probability of being observed [72].

Thus, the basic estimator is [73]

$$E(\hat{c}_i) = \frac{1}{n} \sum_{i=1}^{n} \frac{\delta_i M_{t_i}}{K(T_i)},$$

where $\delta$ is the censoring indicator, $M(t)$ is the cumulative cost up to time $t$ and $K(t)$ is the Kaplan–Meier estimate.

Bang–Tsaiatis proposed an improved version accounting for cost-history lost due to censoring, allowing $M(t)$ and $K(t)$ to be estimated in each of the $K$ intervals, defined as in Lin [74]

$$E(\hat{c}_i) = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{K} \delta_i \left\{ \frac{M_i(t_j) - M_i(t_{j-1})}{K_j(T_j)} \right\}.$$

The two estimators have been shown to be equivalent [75, 76]. Improved versions of the basic estimators have been proposed in recent works using bootstrap approximation of confidence interval, which has been shown to have much better coverage accuracy than the normal approximation in the presence of a heavily skewed distribution [77]. When there is light censoring on medical costs ($< 25\%$) the bootstrap confidence interval based on the simple weighted estimator is again preferred due to its simplicity and good coverage accuracy. For heavily censored cost data (censoring rate $> 30\%$) with larger sample sizes ($n > 200$), the bootstrap confidence interval based on the partitioned estimator has superior performance in terms of both efficiency and coverage accuracy [77]. Clearly, the adopted definition of censoring (Table 2) affects estimates and their precision. In the diabetes cohort (Table 6), mean costs are highly overestimated when Dudley’s definition is used. On the contrary, ignoring censoring is also a source of bias, underestimating actual average costs (Table 6).

A popular approach to modelling censored cost data were to fit survival models on cost data on the basis of the well-known idea of shifting the application of survival techniques.
from a ‘time’ to a ‘costs’ framework, using the actual similarities between the two domains: (i) the non-negative values and (ii) the non-decreasing behaviour. After the first paper introducing the use of the Cox model [55], several criticisms have been raised pointing out the failure of such a model in accounting for non-proportionality in the cost-accumulation process in the presence of censoring [56]. Non-proportionality occurs when the risk of observing costs greater than any given value does not increase linearly with covariates’ value. Anyway, in using the survival models in the cost analysis, we define the cost function as $S(c) = p(c > c)$, where $c_1, \ldots, c_n$ are the costs for each subject, that is the probability that the cost is greater than a certain cost $c$.

The parametric survival model assuming Weibull distribution and the Cox regression model relate the hazard at each cumulative cost $c_i$ to the covariates:

$$\lambda(c_i|x_h) = \lambda_0(c) \exp \left( \sum_{j=1}^{b} \beta_j x_j \right),$$

where $\lambda(c_i|x_h)$ is the hazard rate of a cost $c_i$ for an individual $i$ with a covariates vector $x_h$, and $\lambda_0(c)$ is the baseline risk function. In the Weibull regression model the baseline $\lambda_0(c)$ is assumed to be distributed according to the Weibull distribution, and in the Cox model no assumptions are made about the baseline function. The interpretation varies accordingly: in the case of diabetes data, the estimate of the coefficient for the number of co-morbidities (0.4829, SE 0.0051) indicates that the risk of accumulating further costs is equal to increasing 1.62 times for each additional co-morbidity.

More flexible models have been introduced, as in the non-parametric additive regression approach, where the risk of having an excess of costs (the hazard function) is not modelled using a fixed functional form. One of the most popular models in this class, the Aalen model [78, 79], assumes that the variables interact in an additive manner on the risk function. In this case, the model gives an estimate of how the risk of having costs greater than any given cost increases or decreases in association with the covariates. In the Aalen model the hazard function can be expressed as:

$$\lambda(c_i|x_h) = \lambda_0 + \sum_{j=1}^{b} \alpha_j(c) x_j(c),$$

where the hazard rate $\lambda(c_i|x_h)$ is a linear combination of the variables $x_j(c)$ and the weights $\alpha_j(c)$ are $b$ functions estimated from the data. The slope of each $b$ cumulative regression functions demonstrates the weight of each covariate on the hazard function, while the costs are on the x-axis. In the case in which the covariate has no effect, the function should be a straight line near zero [80, 81] (see Figure 5). In such models, the interpretation of the regression parameters changes and is no longer related to a (possibly transformed) response, as with reference to the diabetes cohort (Table 7). The regression parameters represent the effect of a given covariate on the probability of interrupting the cost-accumulation process after having reached a given cost $c$. Note that in the diabetes cohort also the standard Grambsh–Therneau test for proportionality is failing.

Finally, survival models can also be embedded in a two-part model to account at the same time for zero-costs and censoring [69].

**Discussion**

Several regression models for costs have been proposed in recent years, and now the menu of available approaches is much broader. However, there is no unique model that is able to deal with all the problems that can arise in the analysis of cost data. Therefore, the final choice depends on the type and design of the study.

On the other hand, adopting an approach based on deep modelling of the response variable, with the aim of reducing bias in the cost estimates, is also questionable, because it tends to produce results likely to be model dependent. Recent work has been done to deal with such an issue from the perspective of study design [82, 83], reducing in particular heteroskedasticity by more cautious study design and planning. Nevertheless, the strongest tool to avoid an excess of modelling is the proper use of model diagnostic tools, both as exploratory methods for understanding the degree and direction of skewness, and as way of checking if the fitted model is efficient and robust in statistical terms. Our paper shows indeed that several strategies can be taken to approach the issue of modelling costs. The evidence provided is of course anecdotal, being based on a simple set of two studies, but we believe that these examples are general enough to allow an easy translation of our thoughts to other situations and problems. We offer the following general recommendations. In case of no censoring and no zero-costs, a vast body of literature exists now in favour of the log-gamma GLM model [61]: this allows retention of all the benefits of log or Box–Cox transformation, avoiding troubles in the interpretation of coefficients and avoiding back-transformation issues. To avoid such problems in a different perspective, the survival models and the Cox model have been widely discussed in the literature. Unfortunately, they have been proven to be biased in the

<table>
<thead>
<tr>
<th>Table 7 Effects on cost-hazard for different covariates (95% confidence intervals) in the diabetes cohort</th>
</tr>
</thead>
<tbody>
<tr>
<td>Censored</td>
</tr>
<tr>
<td>---</td>
</tr>
<tr>
<td>Age</td>
</tr>
<tr>
<td>Gender (M vs. F)</td>
</tr>
<tr>
<td>Years</td>
</tr>
<tr>
<td>Number of co-morbidities ≥ 1</td>
</tr>
</tbody>
</table>

Grambsh–Therneau test for proportionality 11.7, $P = 0.01$. Coefficients have been re-parameterized to show a direct effect on costs of covariates. *Coefficients estimated via a Cox model both considering censoring and ignoring it.
presence of even small departures from the underlying assumptions. Survival models are, however, still of interest for a best-fitting strategy, as shown by some recent works, which, using an additive approach to cost data modelling, have a fit comparable with the gamma model with an increased flexibility in modelling extreme costs [80, 81, 84]. Regarding the zero-cost issue, the two-part or even the three-part model is perhaps the most informative, even if the issue of best fit for each part of the model is crucial. Inefficient estimates can heavily affect the marginal effects of covariates also. Without entering in the huge amount of work done with reference to the appropriate treatment of censored data, the Lin-based estimators are outperforming the simpler solutions based on survival techniques.

In any case, the paper clearly shows that alternatives exist to the very simple linear regression model [85], and, given such model's weaknesses, the biases it may introduce and its inability to cope efficiently with the challenges provided by the characteristics of the cost variables, its still large use in the applied literature on cost estimation in health care is hardly justified.

Funding

The study was financially supported by an unrestricted grant of the Compagnia di San Paolo (Torino).

References


