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

A. M. Jones\textsuperscript{1}  J. Lomas\textsuperscript{2}  N. Rice\textsuperscript{1,2}

\textsuperscript{1}Department of Economics and Related Studies
University of York

\textsuperscript{2}Centre for Health Economics
University of York
Motivation

- Modelling costs
- A shift in emphasis

Empirical models

- Overview
- Parametric methods
- Distributional regressions

Data and methodology

- Data
- Methodology

Results

Discussion
Outline

1 Motivation
   - Modelling costs
   - A shift in emphasis

2 Empirical models
   - Overview
   - Parametric methods
   - Distributional regressions

3 Data and methodology
   - Data
   - Methodology

4 Results

5 Discussion
Why model costs?

- **cost-effectiveness analysis:**
  - populating decision models
  - obtaining precise treatment effects (trial and observational data)

- **risk-adjustment:**
  - budgets for healthcare bodies (health authorities, GP consortia)
  - insurance companies

- **attributable healthcare costs:**
  - health behaviours (smoking, alcohol consumption, obesity)

- **drivers of health expenditures:**
  - decomposition analysis over two time periods

- **disparities in utilisation of healthcare:**
  - e.g. related to ethnicity/gender/social class
Challenges of healthcare cost data

- Mass point at zero
  - Focus here on observations with strictly positive costs
- Non-negative
- Heteroskedastic
- Heavily skewed
- Leptokurtic (thick tail)
- Non-linear responses to covariates
Modelling healthcare costs

Histogram plot of outcome variable

- **Costs**
- **root(Costs)**
- **ln(Costs)**
Various approaches

- Linear regression
- Linear regression (with transformed dependent variable)
- Generalised Linear Models (GLM)
  - and extended estimating equations (EEE)
- Duration analysis approaches
- Finite mixture models (FMM)
- Conditional density approximation estimator (CDE)
Which to choose? - an empirical question

- Basu et al. (2006): it is unlikely that economic theory will provide any *a priori* “guidance about distributional characteristics and functional forms that may relate the outcome of interest to covariates”.

=> need for empirical comparisons to guide researchers
Comparative work to date


*But...* Mullahy (2009): Main focus has always been the conditional mean
'Beyond the mean’ - why?

- Mean is of course important, if interested in the total or if government has risk-bearing role (Arrow and Lind, 1970).
- ... But if analysis is restricted solely to the mean, then miss out on a lot of information (Bitler et al., 2006).
- Emphasis on identifying individuals or characteristics of individuals that lead to very large costs “target the high-end parameters of particular interest” including tail probabilities, $P(y > k)$ (Mullahy, 2009).
'Beyond the mean’ - existing techniques

- Methods developed for analysing features of the distribution beyond the mean, particularly in labour economics.
- These methods have been applied in health economics – see, *inter alia*, Cook and Manning (2009) and de Meijer et al. (2013).
Research question

- We want to generate models which can be applied to observations that are out-of-sample...
- ...to forecast $P(y > k)$ [for their given $X$ values]
- And see which econometric approach produces the best results!
Outline

1 Motivation
   - Modelling costs
   - A shift in emphasis

2 Empirical models
   - Overview
   - Parametric methods
   - Distributional regressions

3 Data and methodology
   - Data
   - Methodology

4 Results

5 Discussion
Two groups considered in this paper:

- **Parametric** [from cost regressions lit]: duration analysis models, 2 component gamma FMM.
- **Distributional** [from labour economics]:
  - Using cdf: Han and Hausman (1990), Foresi and Peracchi (1995) and Chernużhukov et al. (2013)
Excluded approaches

- Note! Do not include linear regression (OLS) or GLM approaches.
- These can generate $E(y|X)$ or $\text{Var}(y|X)$, but not the full distribution...
- ...and so can’t be used to give estimates of $P(y>k)$!
What we do for parametric methods - illustration (exponential distribution)
What we do for parametric methods - illustration (exponential distribution)

\[ P(y > k) = \int_k^\infty f(y) \, dy = 1 - F(k) = \exp(-\lambda k) \]
Exponential distribution - 1-parameter
GB2 distribution - 4-parameter

$b=1$, $p=0.5$, $q=2$, and $a$ varies

- $a=1$
- $a=2$
- $a=3$
- $a=4$
- $a=6$
- $a=8$
- $a=12$
- $a=16$

A. M. Jones, J. Lomas, N. Rice
Modelling healthcare costs
Generating a tail probability forecast

- For exponential, as illustration, \( \lambda = \frac{1}{\exp(X\beta)} \)
- Therefore \( P(y > k|X) = \exp\left(-\frac{k}{\exp(X\beta)}\right) \)
- We then average over [all] observations. In additional analysis, we average over different subsets of Xs (description later).
- Parametric methods considered: GB2 (log and sqrt link), generalised gamma (GG), gamma, log-normal (LOGNORM), Weibull (WEIB), exponential (EXP), 2-component gamma finite mixture model (log and sqrt link).

All divide up the distributions into discrete intervals.

Will describe and HH and FP as if distribution is divided into ten deciles, but actually we implement slightly differently (will return to this later).

HH is estimated by running ordered logit with decile number \([1,2,...,10]\) as dependent variable.
We implement Foresi and Peracchi (1995) by running a logit with dependent variable as 1 if decile number is 1, and 0 otherwise... Then a logit with dependent variable as 1 if decile number is 1 OR 2, and 0 otherwise and so on.

The results of each of these logit regressions are saved.

Each logit provides an estimate of $F(\text{decile}|X)$

Chernozhukov et al. (2013) is very similar but extended so that rather than using a fixed number of quantiles, instead do for each unique value of healthcare costs – computationally intensive! Alternative is to run LPM as opposed to logit (e.g. de Meijer et al. (2013)).
Notes regarding our implementation of HH, FP and CH

- Han and Hausman (1990) argue for using as many intervals as possible. With increasing sample sizes more can be used. In preliminary results, we found good convergence performance from using 33 intervals for $N_s = 5000$ and $N_s = 10000$, and 36 intervals for $N_s = 50000$.

- Foresi and Peracchi (1995) use 20 quantiles, we do the same.

- The number of intervals used in Chernozhukov et al. (2013) depends upon the number of unique values of the healthcare cost variable in each sample. We use the LPM method in order to speed up computation. Performance based on this approach was more versatile than performance using logit, since able to estimate with less variation in dependent variable. Where both able to estimate, very little difference in preliminary work.
Generating a tail probability forecast

- Each logit/LPM provides an estimate of $P(y < k^*|X)$, where $k^*$ represents one of the boundaries of the intervals generated using HH or FP, or any cost value observed in the sample when implementing CH.
- Where $k^* \neq k$, use two values of $k^*$ closest to $k$ and use weighted average of two.
- Compute $P(y > k|X) = 1 - P(y < k|X)$. We then average over [all] observations. N.b. could also average over sub-population of observations based on X values.
Estimating the quantile function

- Machado and Mata (2005) [and Melly (2005)] (MM): Estimate full range of quantiles $q_\tau$, where $\tau = 0.5$ corresponds to the median, $\tau = 0.99$ corresponds to 99th percentile etc. using quantile regressions. Save down coefficients.

- For a population [or sub-population] of out-of-sample observations, where wish to forecast distribution of health care costs, forecast a randomly chosen quantile for each observation.

- The predicted quantile represents a draw from the counterfactual distribution of healthcare costs.

- Simply calculate proportion of observations that is greater than $k$ for forecasted tail probability.
Firpo et al. (2009) (RIF): Exactly the same as MM, but use RIF regressions as opposed to quantile regressions to estimate $q_{\tau}$.

Calculate RIF using $RIF(y; q_{\tau}) = q_{\tau} + \frac{\tau - 1[y \leq q_{\tau}]}{f_y(q_{\tau})}$.

Use RIF as dependent variable in rescaled LPM.
Motivation

Empirical models

Data and methodology

Results

Discussion

Outline

1 Motivation
   - Modelling costs
   - A shift in emphasis

2 Empirical models
   - Overview
   - Parametric methods
   - Distributional regressions

3 Data and methodology
   - Data
   - Methodology

4 Results

5 Discussion
**Data**

- **Dependent variable:**
  - Produce annual healthcare cost for patients by summing the costs of all spells taking place in English public sector hospitals finishing in the financial year 2007-2008 (using Hospital Episode Statistics - HES)

- **Explanatory variables:**
  - Age and gender (including squared, cubed and interaction terms)
  - Morbidity markers, adapted from ICD10 chapters – hence indicate presence not severity of morbidity
Methodology

- Use very large dataset from administrative records (HES 2007-2008)
- Divide full set of observations (6,164,114), randomly, into two equally sized subsets: ‘Estimation’ set (3,082,057) ‘Validation’ set (3,082,057)
- From ‘estimation’ set randomly draw samples of size $N_s \in (5,000; 10,000; 50,000)$, with 100 replications
- Estimate models on sample
- Evaluate performance on full ‘validation’ set
Evaluation strategy

- For each model, and for each sample, calculate \( P(y > k) \) for every observation [and later also for subsets of population based on \( X \) values] in the ‘validation’ set and calculate average.
- Then compare this to observed proportion of observations with healthcare costs greater than ‘k’
  - using ratio: \( \frac{\text{estimated } P(y > k)}{\text{fraction of observations in validation set with } y > k} \)
Descriptive statistics

<table>
<thead>
<tr>
<th>$k$</th>
<th>% observations in ‘validation’ set &gt; $k$</th>
</tr>
</thead>
<tbody>
<tr>
<td>£500</td>
<td>82.93%</td>
</tr>
<tr>
<td>£1,000</td>
<td>55.89%</td>
</tr>
<tr>
<td>£2,500</td>
<td>27.04%</td>
</tr>
<tr>
<td>£5,000</td>
<td>13.84%</td>
</tr>
<tr>
<td>£7,500</td>
<td>6.94%</td>
</tr>
<tr>
<td>£10,000</td>
<td>4.10%</td>
</tr>
</tbody>
</table>
Outline

1 Motivation
   - Modelling costs
   - A shift in emphasis

2 Empirical models
   - Overview
   - Parametric methods
   - Distributional regressions

3 Data and methodology
   - Data
   - Methodology

4 Results

5 Discussion

A. M. Jones, J. Lomas, N. Rice
Modelling healthcare costs
Results for $k = 10,000$

![Graph showing probability of high healthcare costs]

<table>
<thead>
<tr>
<th>Method</th>
<th>Bias</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>GB2_LOG</td>
<td>5th</td>
<td>6th</td>
</tr>
<tr>
<td>GB2_SQRT</td>
<td>12th</td>
<td>5th</td>
</tr>
<tr>
<td>GG</td>
<td>9th</td>
<td>4th</td>
</tr>
<tr>
<td>GAMMA</td>
<td>4th</td>
<td>11th</td>
</tr>
<tr>
<td>LOGNORM</td>
<td>11th</td>
<td>1st</td>
</tr>
<tr>
<td>WEIB</td>
<td>7th</td>
<td>12th</td>
</tr>
<tr>
<td>EXP</td>
<td>10th</td>
<td>9th</td>
</tr>
<tr>
<td>FMM_LOG</td>
<td>3rd</td>
<td>13th</td>
</tr>
<tr>
<td>FMM_SQRT</td>
<td>8th</td>
<td>14th</td>
</tr>
<tr>
<td>HH</td>
<td>6th</td>
<td>7th</td>
</tr>
<tr>
<td>FP</td>
<td>13th</td>
<td>3rd</td>
</tr>
<tr>
<td>CH</td>
<td>1st</td>
<td>10th</td>
</tr>
<tr>
<td>MM</td>
<td>2nd</td>
<td>2nd</td>
</tr>
<tr>
<td>RIF</td>
<td>14th</td>
<td>8th</td>
</tr>
</tbody>
</table>
Results for $k = 10,000$

<table>
<thead>
<tr>
<th>Method</th>
<th>Bias</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>GB2_LOG</td>
<td>5th</td>
<td>6th</td>
</tr>
<tr>
<td>GB2_SQRT</td>
<td>12th</td>
<td>5th</td>
</tr>
<tr>
<td>GG</td>
<td>9th</td>
<td>4th</td>
</tr>
<tr>
<td>GAMMA</td>
<td>4th</td>
<td>11th</td>
</tr>
<tr>
<td>LOGNORM</td>
<td>11th</td>
<td>1st</td>
</tr>
<tr>
<td>WEIB</td>
<td>7th</td>
<td>12th</td>
</tr>
<tr>
<td>EXP</td>
<td>10th</td>
<td>9th</td>
</tr>
<tr>
<td>FMM_LOG</td>
<td>3rd</td>
<td>13th</td>
</tr>
<tr>
<td>FMM_SQRT</td>
<td>8th</td>
<td>14th</td>
</tr>
<tr>
<td>HH</td>
<td>6th</td>
<td>7th</td>
</tr>
<tr>
<td>FP</td>
<td>13th</td>
<td>3rd</td>
</tr>
<tr>
<td>CH</td>
<td>1st</td>
<td>10th</td>
</tr>
<tr>
<td>MM</td>
<td>2nd</td>
<td>2nd</td>
</tr>
<tr>
<td>RIF</td>
<td>14th</td>
<td>8th</td>
</tr>
</tbody>
</table>
## Results for $k = 10,000$

<table>
<thead>
<tr>
<th>Method</th>
<th>Bias</th>
<th>Range</th>
<th>Standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>GB2_LOG</td>
<td>6th</td>
<td>6th</td>
<td>6th</td>
</tr>
<tr>
<td>GB2_SQRT</td>
<td>13th</td>
<td>5th</td>
<td>3rd</td>
</tr>
<tr>
<td>GG</td>
<td>10th</td>
<td>4th</td>
<td>2nd</td>
</tr>
<tr>
<td>GAMMA</td>
<td>5th</td>
<td>12th</td>
<td>11th</td>
</tr>
<tr>
<td>LOGNORM</td>
<td>12th</td>
<td>1st</td>
<td>1st</td>
</tr>
<tr>
<td>WEIB</td>
<td>8th</td>
<td>13th</td>
<td>12th</td>
</tr>
<tr>
<td>EXP</td>
<td>11th</td>
<td>9th</td>
<td>8th</td>
</tr>
<tr>
<td>FMM_LOG</td>
<td>4th</td>
<td>14th</td>
<td>15th</td>
</tr>
<tr>
<td>FMM_SQRT</td>
<td>9th</td>
<td>15th</td>
<td>14th</td>
</tr>
<tr>
<td>HH</td>
<td>7th</td>
<td>7th</td>
<td>9th</td>
</tr>
<tr>
<td>FP</td>
<td>14th</td>
<td>3rd</td>
<td>4th</td>
</tr>
<tr>
<td>CH</td>
<td>2nd</td>
<td>10th</td>
<td>10th</td>
</tr>
<tr>
<td>MM</td>
<td>3rd</td>
<td>2nd</td>
<td>5th</td>
</tr>
<tr>
<td>RIF</td>
<td>15th</td>
<td>8th</td>
<td>7th</td>
</tr>
<tr>
<td>NAÏVE</td>
<td>1st</td>
<td>11th</td>
<td>13th</td>
</tr>
</tbody>
</table>
Results for different values of $k$
Results for different values of $k$
Results for different values of $k$
Results for different sample sizes
Results based on subsets of X

- Each method can predict probabilities for specific values of X.
- Results not based on quantiles, average over those observations with those values of X only.
- Results based on quantiles, look at draws from distribution where randomly chosen observation has those values of X only.
- Know what proportion of observations with given X values exceed certain costs in validation set.
- As illustration create an index based on X using a linear regression of y on X in estimation dataset. Divide observations into ten deciles according to index.
Results for deciles of X index

<table>
<thead>
<tr>
<th>Method</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>GB2 LOG</td>
<td>0.467%</td>
<td>0.520%</td>
<td>0.630%</td>
<td>0.79%</td>
<td>1.01%</td>
<td>1.28%</td>
<td>1.73%</td>
<td>2.71%</td>
<td>5.38%</td>
<td>24.47%</td>
</tr>
<tr>
<td>GB2 SQRT</td>
<td>0.338%</td>
<td>0.372%</td>
<td>0.509%</td>
<td>0.71%</td>
<td>0.99%</td>
<td>1.34%</td>
<td>1.90%</td>
<td>3.08%</td>
<td>5.81%</td>
<td>18.14%</td>
</tr>
<tr>
<td>GG</td>
<td>0.330%</td>
<td>0.377%</td>
<td>0.473%</td>
<td>0.61%</td>
<td>0.80%</td>
<td>1.06%</td>
<td>1.49%</td>
<td>2.45%</td>
<td>5.10%</td>
<td>23.68%</td>
</tr>
<tr>
<td>GAMMA</td>
<td>0.000%</td>
<td>0.001%</td>
<td>0.003%</td>
<td>0.01%</td>
<td>0.03%</td>
<td>0.13%</td>
<td>0.43%</td>
<td>1.57%</td>
<td>6.37%</td>
<td>33.36%</td>
</tr>
<tr>
<td>LOGNORM</td>
<td>0.041%</td>
<td>0.059%</td>
<td>0.096%</td>
<td>0.16%</td>
<td>0.25%</td>
<td>0.43%</td>
<td>0.79%</td>
<td>1.73%</td>
<td>4.88%</td>
<td>27.11%</td>
</tr>
<tr>
<td>WEIB</td>
<td>0.001%</td>
<td>0.002%</td>
<td>0.010%</td>
<td>0.03%</td>
<td>0.07%</td>
<td>0.26%</td>
<td>0.74%</td>
<td>2.33%</td>
<td>7.91%</td>
<td>33.64%</td>
</tr>
<tr>
<td>EXP</td>
<td>0.005%</td>
<td>0.015%</td>
<td>0.046%</td>
<td>0.11%</td>
<td>0.21%</td>
<td>0.56%</td>
<td>1.29%</td>
<td>3.23%</td>
<td>8.83%</td>
<td>31.87%</td>
</tr>
<tr>
<td>FMM LOG</td>
<td>0.014%</td>
<td>0.048%</td>
<td>0.132%</td>
<td>0.22%</td>
<td>0.32%</td>
<td>0.79%</td>
<td>1.49%</td>
<td>3.05%</td>
<td>6.91%</td>
<td>27.38%</td>
</tr>
<tr>
<td>FMM SQRT</td>
<td>0.015%</td>
<td>0.045%</td>
<td>0.178%</td>
<td>0.33%</td>
<td>0.44%</td>
<td>1.11%</td>
<td>2.00%</td>
<td>3.88%</td>
<td>7.72%</td>
<td>21.13%</td>
</tr>
<tr>
<td>HH</td>
<td>0.407%</td>
<td>0.463%</td>
<td>0.590%</td>
<td>0.75%</td>
<td>0.98%</td>
<td>1.30%</td>
<td>1.83%</td>
<td>3.02%</td>
<td>6.39%</td>
<td>28.48%</td>
</tr>
<tr>
<td>FP</td>
<td>0.341%</td>
<td>0.499%</td>
<td>0.584%</td>
<td>0.68%</td>
<td>1.09%</td>
<td>1.43%</td>
<td>2.37%</td>
<td>3.80%</td>
<td>7.69%</td>
<td>31.14%</td>
</tr>
<tr>
<td>CH</td>
<td>-3.258%</td>
<td>-2.481%</td>
<td>-2.052%</td>
<td>-1.01%</td>
<td>0.88%</td>
<td>2.09%</td>
<td>4.27%</td>
<td>6.86%</td>
<td>11.34%</td>
<td>24.34%</td>
</tr>
<tr>
<td>MM</td>
<td>0.006%</td>
<td>0.018%</td>
<td>0.046%</td>
<td>0.09%</td>
<td>0.21%</td>
<td>0.49%</td>
<td>1.13%</td>
<td>2.59%</td>
<td>6.69%</td>
<td>30.23%</td>
</tr>
<tr>
<td>RIF</td>
<td>0.159%</td>
<td>0.264%</td>
<td>0.378%</td>
<td>0.59%</td>
<td>1.16%</td>
<td>1.82%</td>
<td>3.49%</td>
<td>6.58%</td>
<td>13.08%</td>
<td>22.57%</td>
</tr>
</tbody>
</table>

Table 6: Forecasted frequencies of a cost exceeding £10,000, sample size 5,000, by decile of linear index of covariates
Outline

1 Motivation
   - Modelling costs
   - A shift in emphasis

2 Empirical models
   - Overview
   - Parametric methods
   - Distributional regressions

3 Data and methodology
   - Data
   - Methodology

4 Results

5 Discussion

A. M. Jones, J. Lomas, N. Rice
Modelling healthcare costs
Discussion

- Trade-off between bias and precision.
- Bias for parametric models is determined by $k$.
- Some generally more precise than others – LOGNORM good, GAMMA, FMM_SQRT etc bad. (note link function doesn’t appear that important)
- Precision increases for all methods with greater sample size.
- MM and RIF don’t seem to perform well in terms of either bias or precision.
- CH demonstrates potential - especially for larger sample sizes… As do HH and FP (although not for large $k$)...
- … but smoothing techniques are required for forecasting out-of-support.