Going Beyond the Mean in Healthcare Cost Regressions: A Comparison of Methods for Estimating the Full Conditional Distribution

Andrew Jones
James Lomas
Nigel Rice

Department of Economics and Centre for Health Economics
University of York, UK

Professor William Greene
Stern School of Business
Department of Economics
Continuation of earlier work that focused on the mean. Focused on prediction and “fit” as measured by the likelihood. Developed generalized beta of the second kind as a model.
Outline

- Context: Modeling Health Care Costs
- Candidate Models for Health Care Costs
- Empirical Methods
  - A ‘quasi - Monte Carlo study
  - Data
- Findings
- Discussion
Setting: Modeling Health Care Costs

Motivation

- Cost Effectiveness
  - Decision models
  - Estimation of Treatment effects
- Resource allocation
  - Attributable costs
  - Health behavior (smoking, obesity)
- Risk adjustment in insurance systems
  - Budget for healthcare providers
  - Reimbursement policies for insurance
- Ethnic/gender/demographic variation in utilization
Long history of studies of healthcare costs that mostly focus on the conditional mean
  - (log)linear regressions
  - Monte Carlo Studies
  - GLMs
    - Non-normal models of the distribution
    - Focused on the mean
  - Finite mixture models
  - Density approximations

Less attention to non- and semiparametric

Recommendations (e.g., Mullahy (2009)) to examine tail probabilities (upper quantiles)
Modeling Objective

- Interest in full distribution of healthcare costs, not just the mean. (“Beyond the mean…”)  
- Characterizing the DGP  
  - Familiar conditional mean – regression style models  
  - Variation and scedastic functions  
  - Higher moments: skewness, kurtosis  
  - Quantiles and tail behavior  
  - Predict tail probabilities out of sample  
- Theory does not help choose functional forms or approaches (regression, semiparametric, etc.)  
- Why beyond the mean?  
  - ID individuals that lead to large costs  
  - Examine x such that Prob(Cost > K |x) is small (tails)
Contributions of the Study

- **General**: Methodology for more detailed examination of distribution of a cost variable
  - Features of the distribution
  - Performance of different modeling approaches
  - Comparison of methods for fitting the full distribution
  - Devise a method of finding an out of sample prediction

- **Specific**: Characteristics of the distribution of a specific component of healthcare costs in UK
Main Approach of the Study

- Familiar Regression Style Approach
  - Nonlinear
  - GLM style modeling
  - Parameterize location and scale parameters
  - Heteroscedasticity?
    - A side result, and only an accident of GLMs
    - Not a focus of the study
- Examination of Higher Moments: GLMs don’t seem to work well with heavy tailed data; they focus on the conditional mean function.
- Examine quantiles of distribution through survival function
Features of the Candidate Models

- Respect nonnegativity of costs
- Familiar conditional moments
  - Nonlinearity of conditional mean and responses
  - Higher Conditional Moments: skewness, kurtosis
- Conditional quantiles
- Survival function
14 Candidates for Modeling Costs

- 7 Parametric models.
  - Not including normal
  - Variants of least squares are not among the estimators.
  - Generally not GLM. Generalized linearity is not an objective
- 2 finite mixture (of gamma) models
- 3 semiparametric approaches to approximating the quantiles of the distribution
- 2 quantile regressions methods
7 Parametric Models

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

Table 2: Key for method labels
<table>
<thead>
<tr>
<th>GB2_LOG</th>
<th>generalised beta of the second kind (log-link)</th>
</tr>
</thead>
<tbody>
<tr>
<td>GB2_SQRT</td>
<td>generalised beta of the second kind (√-link)</td>
</tr>
<tr>
<td>GG</td>
<td>generalised gamma (log-link)</td>
</tr>
<tr>
<td>GAMMA</td>
<td>two-parameter gamma (log-link)</td>
</tr>
<tr>
<td>LOGNORM</td>
<td>log-normal (log-link)</td>
</tr>
<tr>
<td>WEIB</td>
<td>Weibull (log-link)</td>
</tr>
<tr>
<td>EXP</td>
<td>exponential (log-link)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>FMM_LOG</th>
<th>two-component finite mixture of gamma densities (log-link)</th>
</tr>
</thead>
<tbody>
<tr>
<td>FMM_SQRT</td>
<td>two-component finite mixture of gamma densities (√-link)</td>
</tr>
<tr>
<td>HH</td>
<td>Han and Hausman</td>
</tr>
<tr>
<td>FP</td>
<td>Foresi and Peracchi</td>
</tr>
<tr>
<td>CH</td>
<td>Chernozhukov, FernndezVal and Melly (linear probability model)</td>
</tr>
<tr>
<td>MM</td>
<td>Machado and Mata - Melly (log-transformed outcome)</td>
</tr>
<tr>
<td>RIF</td>
<td>recentered-influence-function regression (linear probability model)</td>
</tr>
</tbody>
</table>

Table 2: Key for method labels
3 Semiparametric Approaches

<table>
<thead>
<tr>
<th>Method</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>GB2_LOG</td>
<td>Generalised beta of the second kind (log-link)</td>
</tr>
<tr>
<td>GB2_SQRT</td>
<td>Generalised beta of the second kind (√-link)</td>
</tr>
<tr>
<td>GG</td>
<td>Generalised gamma (log-link)</td>
</tr>
<tr>
<td>GAMMA</td>
<td>Two-parameter gamma (log-link)</td>
</tr>
<tr>
<td>LOGNORM</td>
<td>Log-normal (log-link)</td>
</tr>
<tr>
<td>WEIB</td>
<td>Weibull (log-link)</td>
</tr>
<tr>
<td>EXP</td>
<td>Exponential (log-link)</td>
</tr>
<tr>
<td>FMM_LOG</td>
<td>Two-component finite mixture of gamma densities (log-link)</td>
</tr>
<tr>
<td>FMM_SQRT</td>
<td>Two-component finite mixture of gamma densities (√-link)</td>
</tr>
<tr>
<td>HH</td>
<td>Han and Hausman</td>
</tr>
<tr>
<td>FP</td>
<td>Foresi and Peracchi</td>
</tr>
<tr>
<td>CH</td>
<td>Chernozhukov, Fernández-Val and Melly (linear probability model)</td>
</tr>
<tr>
<td>MM</td>
<td>Machado and Mata - Melly (log-transformed outcome)</td>
</tr>
<tr>
<td>RIF</td>
<td>Recentered-influence-function regression (linear probability model)</td>
</tr>
</tbody>
</table>

Table 2: Key for method labels
## 2 Quantile Regression Approaches

<table>
<thead>
<tr>
<th>Label</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>GB2_LOG</td>
<td>Generalised beta of the second kind (log-link)</td>
</tr>
<tr>
<td>GB2_SQRT</td>
<td>Generalised beta of the second kind (√-link)</td>
</tr>
<tr>
<td>GG</td>
<td>Generalised gamma (log-link)</td>
</tr>
<tr>
<td>GAMMA</td>
<td>Two-parameter gamma (log-link)</td>
</tr>
<tr>
<td>LOGNORM</td>
<td>Log-normal (log-link)</td>
</tr>
<tr>
<td>WEIB</td>
<td>Weibull (log-link)</td>
</tr>
<tr>
<td>EXP</td>
<td>Exponential (log-link)</td>
</tr>
<tr>
<td>FMM_LOG</td>
<td>Two-component finite mixture of gamma densities (log-link)</td>
</tr>
<tr>
<td>FMM_SQRT</td>
<td>Two-component finite mixture of gamma densities (√-link)</td>
</tr>
<tr>
<td>HH</td>
<td>Han and Hausman</td>
</tr>
<tr>
<td>FP</td>
<td>Foresi and Peracchi</td>
</tr>
<tr>
<td>CH</td>
<td>Chernozhukov, FernandezVal and Melly (linear probability model)</td>
</tr>
<tr>
<td>MM</td>
<td>Machado and Mata - Melly (log-transformed outcome)</td>
</tr>
<tr>
<td>RIF</td>
<td>Recentered-influence-function regression (linear probability model)</td>
</tr>
</tbody>
</table>

Table 2: Key for method labels
# Densities for Parametric Models

| Model     | $f(y | X) = $ |
|-----------|-------------|
| GB2_LOG   | $a y^{a p - 1} \frac{1}{\exp(X\beta)^{a p} B(p, q)[1 + (\frac{y}{\exp(X\beta)})^a]^{(p+q)}}$ |
| GB2_SQRT  | $a y^{a p - 1} \frac{1}{(X\beta)^{2 a p} B(p, q)[1 + (\frac{y}{(X\beta)^2})^a]^{(p+q)}}$ |
| GG        | $\frac{\kappa}{\sigma y \Gamma(\kappa - 2)} \left( \kappa - 2 \left( \frac{y}{\exp(X\beta)} \right)^{\kappa/\sigma} \right)^{\kappa - 2} \exp\left(-\kappa - 2 \left( \frac{y}{\exp(X\beta)} \right)^{\kappa/\sigma} \right)$ |
| GAMMA     | $\frac{1}{y \Gamma(\kappa - 2)} \left( \kappa - 2 \left( \frac{y}{\exp(X\beta)} \right)^{\kappa/\sigma} \right)^{\kappa - 2} \exp\left(-\kappa - 2 \left( \frac{y}{\exp(X\beta)} \right)^{\kappa/\sigma} \right)$ |
| LOGNORM   | $\frac{1}{\sqrt{2\pi} y \sqrt{2\pi}} \exp\left(-\frac{(\ln y - X\beta)^2}{2\sigma^2} \right)$ |
| WEIB      | $\frac{1}{\sigma y} \left( \frac{y}{\exp(X\beta)} \right)^{\frac{1}{\sigma}} \exp\left(-\left( \frac{y}{\exp(X\beta)} \right)^{\frac{1}{\sigma}} \right)$ |
| EXP       | $\frac{1}{\exp(X\beta)} \left( \frac{-y}{\exp(X\beta)} \right)$ |
| FMM_LOG   | $\sum_j^2 \pi_j \frac{y^{\alpha_j}}{\gamma(\alpha_j) \exp(X\beta_j)^{\alpha_j}} \exp\left(-\left( \frac{y}{\exp(X\beta_j)} \right)\right)$ |
| FMM_SQRT  | $\sum_j^2 \pi_j \frac{y^{\alpha_j}}{\gamma(\alpha_j)(X\beta_j)^{2\alpha_j}} \exp\left(-\left( \frac{y}{(X\beta_j)^2} \right)\right)$ |
Survival Functions

\[
\Pr(y > k|X) =
\]

\[
1 - I_Z(p, q)^* \text{ where } z = \left(\frac{k}{\exp(X\beta)}\right)^\alpha
\]

\[
1 - I_Z(p, q)^* \text{ where } z = \left(\frac{k}{(X\beta)^2}\right)^\alpha
\]

\[
\text{if } \kappa > 0: \quad 1 - \Gamma\left(z; \kappa^{-2}\right)^* \quad \text{if } \kappa < 0: \quad \Gamma\left(z; \kappa^{-2}\right)^* \text{ where } z = \kappa^{-2}\left(\frac{k}{\exp(X\beta)}\right)^{\kappa/\sigma}
\]

\[
\kappa > 0: \quad 1 - \Gamma\left(z; \kappa^{-2}\right)^* \quad \text{if } \kappa < 0: \Gamma\left(z; \kappa^{-2}\right)^* \text{ where } z = \kappa^{-2}\left(\frac{k}{\exp(X\beta)}\right)^\alpha
\]

\[
1 - \Phi\left(\frac{\ln k - X\beta}{\sigma}\right)
\]

\[
\exp\left(-\left(\frac{k}{\exp(X\beta)}\right)^{\frac{1}{\sigma}}\right)
\]

\[
\exp\left(-\frac{k}{\exp(X\beta)}\right)
\]

\[
\sum_j^2 \pi_j \left(1 - \Gamma(z; \alpha_j)\right)^*** \text{ where } z = \frac{k}{\exp(X\beta_j)}
\]

\[
\sum_j^2 \pi_j \left(1 - \Gamma(z; \alpha_j)\right)^*** \text{ where } z = \frac{k}{(X\beta_j)^2}
\]

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

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

***where \(\Gamma(z; \alpha_j) = \frac{1}{\Gamma(\alpha_j)} \int_0^z t^{(\alpha_j-1)} \exp(-t) dt\).
In the form in Table 2, the following results if $P = \kappa^{-2}$ and $\lambda = \frac{\kappa^{-2}}{\exp(\beta' x)}$.

\[
f(y \mid x) = \frac{\lambda^P}{\Gamma(P)} \exp(-\lambda y) y^{P-1}, \quad y \geq 0, \quad P > 0, \quad \lambda = \exp(-\beta' x)
\]

\[
E[y \mid x] = \frac{P}{\lambda} = P \exp(\beta' x), \quad V(y \mid x) = \frac{P}{\lambda^2} = \frac{1}{P} \left( E[y \mid x] \right)^2
\]

\[
S(k \mid x) = \text{Prob}(y \geq k \mid x) = \int_{k}^{\infty} \frac{\lambda^P}{\Gamma(P)} \exp(-\lambda y) y^{P-1} dy
\]

Incomplete gamma integral. (There is literature and software available.)

\[
\text{Prob}(y \geq k) = \int_{x}^{\infty} \int_{k}^{\infty} \frac{\lambda^P}{\Gamma(P)} \exp(-\lambda y) y^{P-1} dy dx
\]

Approximate with

\[
\frac{1}{N_c} \sum_{i=1}^{N_c} \int_{k}^{\infty} \frac{\lambda(x_i)^P}{\Gamma(P)} \exp(-\lambda(x_i) y) y^{P-1} dy
\]
Finite Mixtures of (2) Gammas

\[
f(y | x) = \pi \frac{\exp(-\beta'_1 x)^{P_1}}{\Gamma(P_1)} \exp(-\exp(-\beta'_1 x) y) y^{P_1 - 1} \\
+ (1 - \pi) \frac{\exp(-\beta'_2 x)^{P_2}}{\Gamma(P_2)} \exp(-\exp(-\beta'_2 x) y) y^{P_2 - 1}
\]

With "log link," \( \lambda = \exp\left(-\beta'_j x\right) \)

With "square root link," \( \lambda = \left(\beta'_j x\right)^2 \)
The skew-normal distribution

Azzalini (1985, 1986) defines a continuous random variable $Z$ to have a skew-Normal distribution, denoted $SN(0, 1, \alpha)$, if it has density function

$$2\phi(z)\Phi(\alpha z)$$

where $\phi$ and $\Phi$ denote the density and distribution functions, respectively, of a standard Normal $N(0, 1)$ variate. The skew-Normal distribution is essentially a Normal distribution that has been augmented by the addition of a shape parameter $\alpha \in (-\infty, +\infty)$ that quantifies the skewness of the distribution;

$$\log \text{Revenue}_i = \beta_0 + \beta_1 \log \text{Budget}_i + \beta_2 \log \text{Opening Screens}_i + \beta_3 \text{Sequel}_i + \beta_4 \text{Star}_i + \gamma_1 \text{Genre}_i + \gamma_2 \text{Rating}_i + \gamma_3 \text{Year}_i + \mu_i$$

Another Candidate: Skew Normal

Quantile Regression Methods

• Machado and Mata (2005), Melly (2005) [MM]
  Quantile Regression Method for log of costs.
  Estimated quantile functions used to construct the
  counterfactual distribution

• Firpo et al. (2009) [RIF]
  Also quantile regression based,
  “Recentered Influence Function Regressions”

\[
\text{RIF}(y,q_\tau) = q_\tau + \frac{\tau - 1(y \leq q_\tau)}{f_y(q_\tau)}
\]

\(f_y(q_\tau)\) = kernel density estimated for \(q_\tau\).
RIF is the LHS in an OLS regression
Subsequent steps same as for MM
Partitioning the distribution into discrete intervals.

- Han and Hausman (1990): Partition range of cost into intervals. Ordered logit using the decile number as LHS gives estimates of the CDF. Used 33 or 38 intervals. HH recommended 10. (HH)

- Foresi and Peracchi (1995) use separate logit for each cell. (See Long and Freese – “parallel regressions” thing) Each logit provides an estimate of the CDF. Used 20 intervals. (FP)

- Chernozhukov et al. (2013) Fit a logit for each unique cost value. Provides a continuous CDF. What if every cost value is different? Better use neighborhoods. Used LPM instead of logit to save time. (CH)

Each logit or LPM from these methods gives an estimate of $P(y<k^*|x)$. $K^*$ might not be the $k$ of interest. Use the $k^*$ closest to $k$ and weighted average of two nearest. Compute survival as $1 - CDF$, then average over sub-populations based on the index of $x$ values.
• Split sample into an estimation half and a validation half
• Samples are drawn from the estimation half
  • 300 subsamples drawn with replacement,
  • $N_1 = 5000$, $N_2 = 10000$, $N_3 = 50000$, 100 samples of each size
• 14 models fit with each of the 300 subsamples
  • Construct $F(y|x)$ based on the split of the index variable.
  • Construct counterfactual $F(y|x)$ based on validation set
  • For specific values of $k$, obtain $\text{Prob}(y > k|x)$ and average over $x$ to obtain $\text{Prob}(y > k)$.
  • Compare to observed empirical proportion of data that exceed $k$.
  • $k = 500, 1000, 1500, 500, 7500, 10000$.
• Construct estimate of $\text{Prob}(y > k)/\text{Sample Frequency}(y > k)$.
  • Average across replications.
  • Variance across replications (average absolute deviation), s.d., and range
UK Hospital Episode Statistics 2007-2008 financial year
- Not mental or maternity
- Costs
- 6.164M observations
- 24 Morbidity markers (dummy variables)
- Age, sex
The Data

<table>
<thead>
<tr>
<th>N</th>
<th>6,164,114</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>£2,610</td>
</tr>
<tr>
<td>Median</td>
<td>£1,126</td>
</tr>
<tr>
<td>Standard deviation</td>
<td>£5,088</td>
</tr>
<tr>
<td>Skewness</td>
<td>13.03</td>
</tr>
<tr>
<td>Kurtosis</td>
<td>363.18</td>
</tr>
<tr>
<td>Minimum</td>
<td>£217</td>
</tr>
<tr>
<td>Maximum</td>
<td>£604,701</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>% observations</th>
<th>% of total costs</th>
</tr>
</thead>
<tbody>
<tr>
<td>£500</td>
<td>82.96%</td>
</tr>
<tr>
<td>£1,000</td>
<td>55.89%</td>
</tr>
<tr>
<td>£2,500</td>
<td>27.02%</td>
</tr>
<tr>
<td>£5,000</td>
<td>13.83%</td>
</tr>
<tr>
<td>£7,500</td>
<td>6.92%</td>
</tr>
<tr>
<td>£10,000</td>
<td>4.09%</td>
</tr>
</tbody>
</table>

Table 1: Descriptive statistics for hospital costs

Figure 1: Empirical density and cumulative distribution of healthcare costs
Linear index computed by multiple regression, divided into 5 quintiles. Distribution of log costs in each index quintile.

Figure 2: Empirical distribution of log-costs for each of the 5 quantiles of the linear index of covariates.
Skewness and kurtosis in 10 quintiles of linear index and full sample

Locus of possible skewness and kurtosis combinations based on 8 distributions programmed in McDonald et al. (2011)

Full sample shows more kurtosis $(3(5.088) > 363.18$ for the full sample, and more skewness $(13.03)$

For the gamma model, kurtosis $= 1.5 \text{ skewness}^2$. 
Findings

- For each model, and for each sample, calculate $P(y > k)$ for every observation [could have chosen a sub-set 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: $\frac{\text{estimated } P(y > k)}{\text{fraction of observations in validation set with } y > k}$

<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>
Findings

**Graph:**
- `Pr(y>£10,000): sample size 5,000`

**Table:**

<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>
Performance of Candidate Models

Pr(y>£10,000): sample size 5,000

Pr(y>£10,000): sample size 50,000
Performance of Best Candidates by Quantile
Conclusions from the Study

• General conclusions?
  • Lognormal seems to be a good model
  • Gamma and generalizations seem not to be
  • Nonparametric methods perform well, but are less useful than parametric for the purpose here
  • Semiparametric estimators HH, FP, CH do not allow out of sample estimation.

• Bias vs. Precision
  • Estimators appear to be consistent
  • MM and RIF don’t look good by any measure
  • CH looks good, but fails the usefulness test

• What did we learn about the specific cost variable?
Some Points to (re)Consider

- Best model seems to depend on k and N.
- Is there any generality here?
- Why those parametric forms? Are there others, e.g., skew t?
- CH, HH and FP cannot extrapolate beyond the observed data. (Bayesian?) How are these methods useful?
- Choose a preferred model and analyze costs in detail
  - Quantiles
  - Partial effects and drivers