The 6th Annual Health Econometrics Workshop

September 25–27, 2014

University of St. Michael's College
In the University of Toronto,
Toronto, Canada

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.

Journal of
APPLIED ECONOMETRICS

Research Article

APPLYING BETA-TYPE SIZE DISTRIBUTIONS TO HEALTHCARE COST REGRESSIONS

Andrew M. Jones[1,*], James Lomas[1,2] and Nigel Rice[1,2]

Article first published online: 8 JUL 2013
DOI: 10.1002/jae.2334

Issue

Journal of Applied Econometrics
Volume 29, Issue 4, pages 649–670, June/July 2014

SUMMARY

This paper extends the literature on modelling healthcare cost data by applying the generalised beta of the second kind (GB2) distribution to English hospital inpatient cost data. A quasi-experimental design, estimating models on a sub-population of the data and evaluating performance on another sub-population, is used to compare this distribution with its nested and limiting cases. While for these data the beta of the second kind (B2) distribution and generalised gamma (GG) distribution outperform the GB2, our results illustrate that the GB2 can be used as a device for choosing among competing parametric distributions for healthcare cost data. Copyright © 2013 John Wiley & Sons, Ltd.

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

Literature Context

- 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 ($\sqrt{}$-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 ($\sqrt{}$-link)
HH	Han and Hausman
FP	Foresi and Peracchi
CH	Chernozhukov, FernndezVal 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

2 Finite Mixtures of Gamma Models

GB2_LOG	generalised beta of the second kind (log-link)
GB2_SQRT	generalised beta of the second kind ($\sqrt{}$-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 ($\sqrt{}$-link)
HH	Han and Hausman
FP	Foresi and Peracchi
CH	Chernozhukov, FernndezVal 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

3 Semiparametric Approaches

GB2_LOG	generalised beta of the second kind (log-link)
GB2_SQRT	generalised beta of the second kind ($\sqrt{}$-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 ($\sqrt{}$-link)
HH	Han and Hausman
FP	Foresi and Peracchi
CH	Chernozhukov, FernndezVal 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

2 Quantile Regression Approaches

GB2_LOG	generalised beta of the second kind (log-link)
GB2_SQRT	generalised beta of the second kind ($\sqrt{}$-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 ($\sqrt{}$-link)
HH	Han and Hausman
FP	Foresi and Peracchi
CH	Chernozhukov, FernndezVal 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

Densities for Parametric Models

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

Survival Functions

$\mathbf{Pr(y > k|X)} =$
$1 - I_Z\,(p,q)^*$ where $z = \left(\frac{k}{\exp(X\beta)}\right)^a$
$1 - I_Z\,(p,q)^*$ where $z = \left(\frac{k}{(X\beta)^2}\right)^a$
if $\kappa > 0$: $1 - \Gamma\left(z;\kappa^{-2}\right)^{**}$ if $\kappa < 0$: $\Gamma\left(z;\kappa^{-2}\right)^{**}$ where $z = \kappa^{-2}\left(\frac{k}{\exp(X\beta)}\right)^{\kappa/\sigma}$
$\kappa > 0$: $1 - \Gamma\left(z;\kappa^{-2}\right)^{**}$ if $\kappa < 0$:$\Gamma\left(z;\kappa^{-2}\right)^{**}$ where $z = \kappa^{-2}\left(\frac{k}{\exp(X\beta)}\right)$
$1 - \Phi\left(\frac{\ln k - X\beta}{\sigma}\right)$
$\exp\left(-\left(\frac{k}{\exp(X\beta)}\right)^{\frac{1}{\sigma}}\right)$
$\exp\left(-\frac{k}{\exp(X\beta)}\right)$
$\sum_j^2 \pi_j\,(1 - \Gamma\,(z;\alpha_j))^{***}$ where $z = \frac{k}{\exp(X\beta_j)}$
$\sum_j^2 \pi_j\,(1 - \Gamma\,(z;\alpha_j))^{***}$ where $z = \frac{k}{(X\beta_j)^2}$

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

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

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

A Simpler Form of the Gamma Model

In the form in Table 2, the following results if $P = \kappa^{-2}$ and $\lambda = \dfrac{\kappa^{-2}}{\exp(\beta' x)}$.

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

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

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

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

$$\text{Prob}(y \geq k) = \int_x \int_k^\infty \frac{\lambda^P}{\Gamma(P)} \exp(-\lambda y) y^{P-1} dy dx$$

Approximate with $\dfrac{1}{N_c} \sum_{i=1}^{N_c} \int_k^\infty \dfrac{\lambda(x_i)^P}{\Gamma(P)} \exp(-\lambda(x)_i y) y^{P-1} dy$

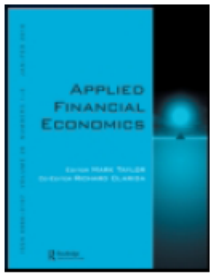
Finite Mixtures of (2) Gammas

$$f(y \mid x) = \pi \frac{\exp(-\beta_1' x)^{P_1}}{\Gamma(P_1)} \exp(-\exp(-\beta_1' x) y) y^{P_1 - 1} + (1-\pi) \frac{\exp(-\beta_2' x)^{P_2}}{\Gamma(P_2)} \exp(-\exp(-\beta_2' x) y) y^{P_2 - 1}$$

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

With "square root link," $\lambda = \left(\beta_j' x\right)^2$

Another Candidate: Skew Normal
W. D. Walls (2006) On Skewness in the Movies



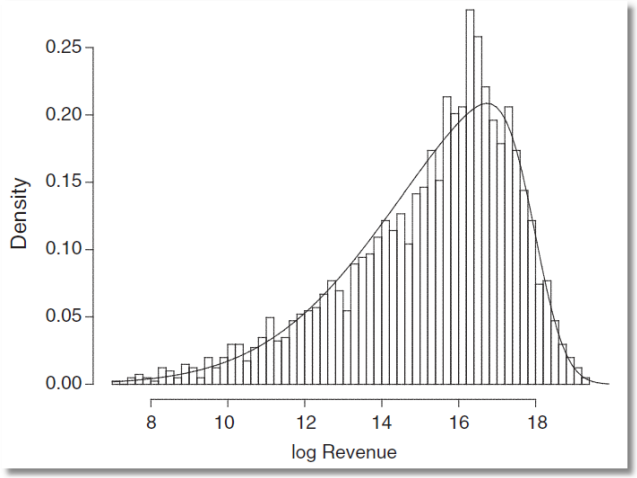
Applied Financial Economics
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/rafe20

Modelling heavy tails and skewness in film returns

W. D. Walls [a]
[a] Department of Economics, University of Calgary, Calgary, Alberta, Canada, T2N 1N4
Published online: 20 Aug 2006.

To cite this article: W. D. Walls (2005) Modelling heavy tails and skewness in film returns, Applied Financial Economics, 15:17, 1181-1188, DOI: 10.1080/0960310050391040

To link to this article: http://dx.doi.org/10.1080/0960310050391040



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

$$2\phi(z)\Phi(\alpha z) \qquad (1)$$

where ϕ and Φ denote the density and distribution functions, respectively, of a standard Normal $N(0, 1)$ variate.[3] 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;

$$\begin{aligned}\log \text{Revenue}_i = {} & \beta_0 + \beta_1 \log \text{Budget}_i \\ & + \beta_2 \log \text{Opening Screens}_i + \beta_3 \text{Sequel}_i \\ & + \beta_4 \text{Star}_i + \gamma_1' \text{Genre}_i + \gamma_2' \text{Rating}_i \\ & + \gamma_3' \text{Year}_i + \mu_i \qquad (3)\end{aligned}$$

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

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"

$$\mathrm{RIF}(y,q_\tau) = q_\tau + \frac{\tau - \mathbf{1}(y \le q_\tau)}{f_y(q_\tau)}$$

$f_y(q_\tau)$ = kernel density estimated for q_τ.
RIF is the LHS in an OLS regression
Subsequent steps same as for MM

Semiparametric Quantile Estimation

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.

Quasi - Monte Carlo Method

- 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 Prob(y > k|x) and average over x to obtain Prob(y>k).
  - Compare to observed empirical proportion of data that exceed k.
  - k=500,1000,1500,500,7500,10000.
- Construct estimate of Prob(y > k)/Sample Frequency(y > k).
  - Average across replications.
  - Variance across replications (average absolute deviation), s.d., and range

Data

- 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

N	6,164,114	
Mean	£2,610	
Median	£1,126	
Standard deviation	£5,088	
Skewness	13.03	
Kurtosis	363.18	
Minimum	£217	
Maximum	£604,701	
	% observations	% of total costs
> £500	82.96%	97.20%
> £1,000	55.89%	89.80%
> £2,500	27.02%	72.35%
> £5,000	13.83%	54.65%
> £7,500	6.92%	38.67%
> £10,000	4.09%	29.35%

Table 1: Descriptive statistics for hospital costs

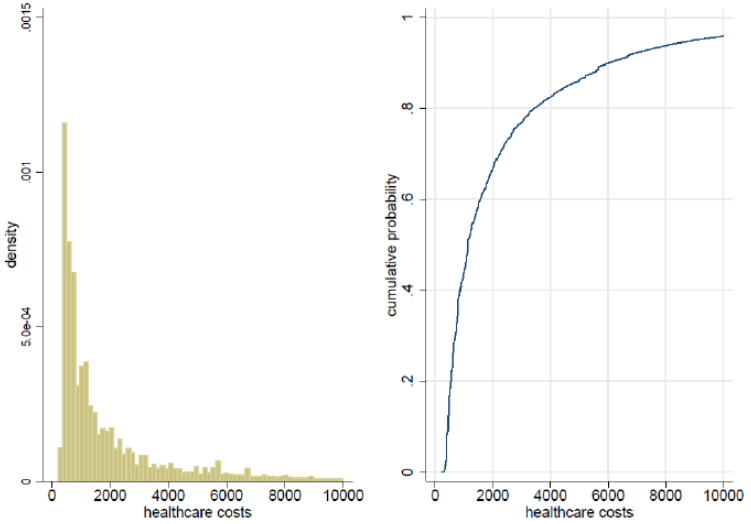


Figure 1: Empirical density and cumulative distribution of healthcare costs

The Cost Data

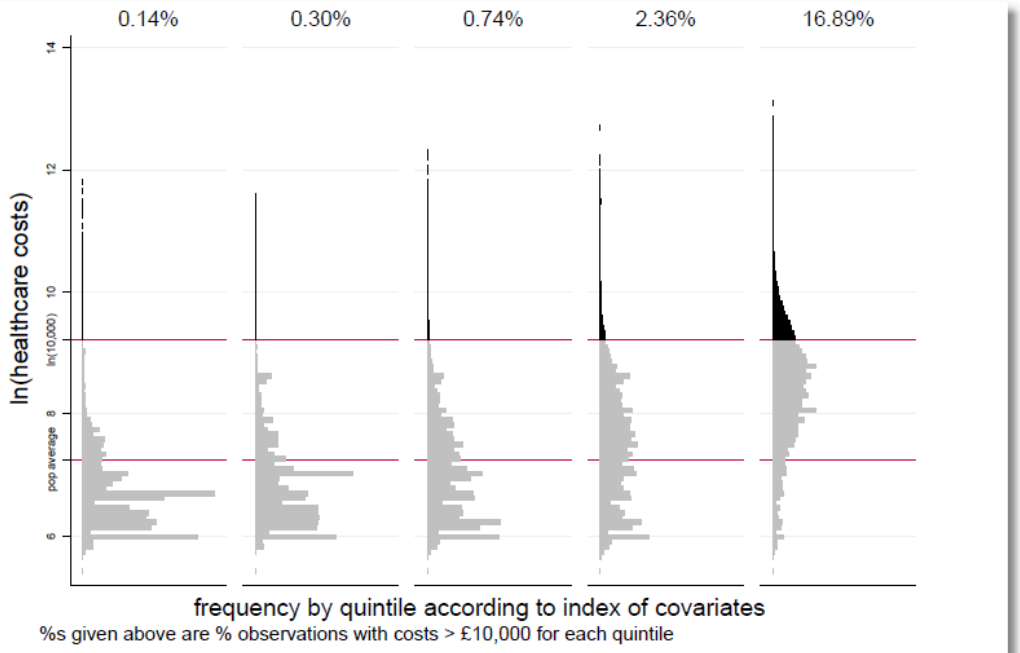


Figure 2: Empirical distribution of log-costs for each of the 5 quantiles of the linear index of covariates

Linear index computed by multiple regression, divided into 5 quintiles. Distribution of log costs in each index quintile.

Relationship of Skewness and Kurtosis in Parametric Models

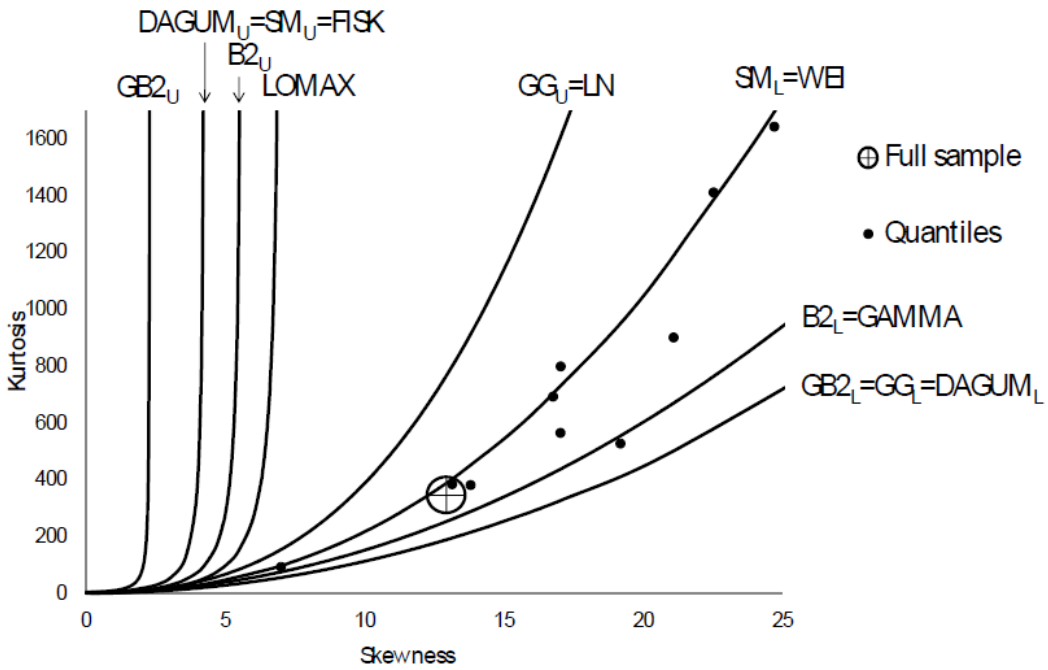


Figure 3: Kurtosis against skewness for each of the 10 quantiles of the linear index of covariates

Note: Taken from Jones et al. (2014) and adapted from McDonald et al. (2011). The dots shown on Figure 3 were generated as follows: the data were divided into ten subsets using the deciles of a simple linear predictor for healthcare costs using the set of regressors used in this paper. Figure 3 plots the skewness and kurtosis coefficients of actual healthcare costs for each of these subsets, the skewness and kurtosis coefficient of the full estimation sub-population (represented by the larger circle with cross) and theoretically possible skewness-kurtosis spaces and loci for parametric distributions considered in the literature.

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 ($3(5.088^4) > 363.18$ for the full sample, and more skewness (13.03)

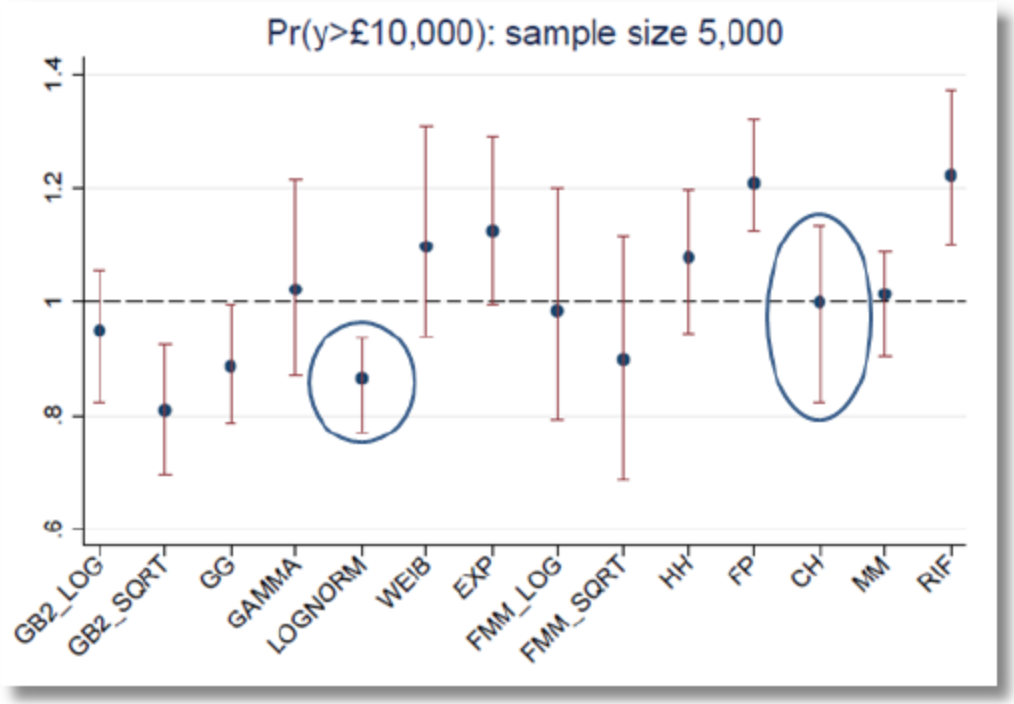
For the gamma model, kurtosis = 1.5 skewness^2.

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: $\frac{\text{estimated } P(y>k)}{\text{fraction of observations in validation set with } y>k}$

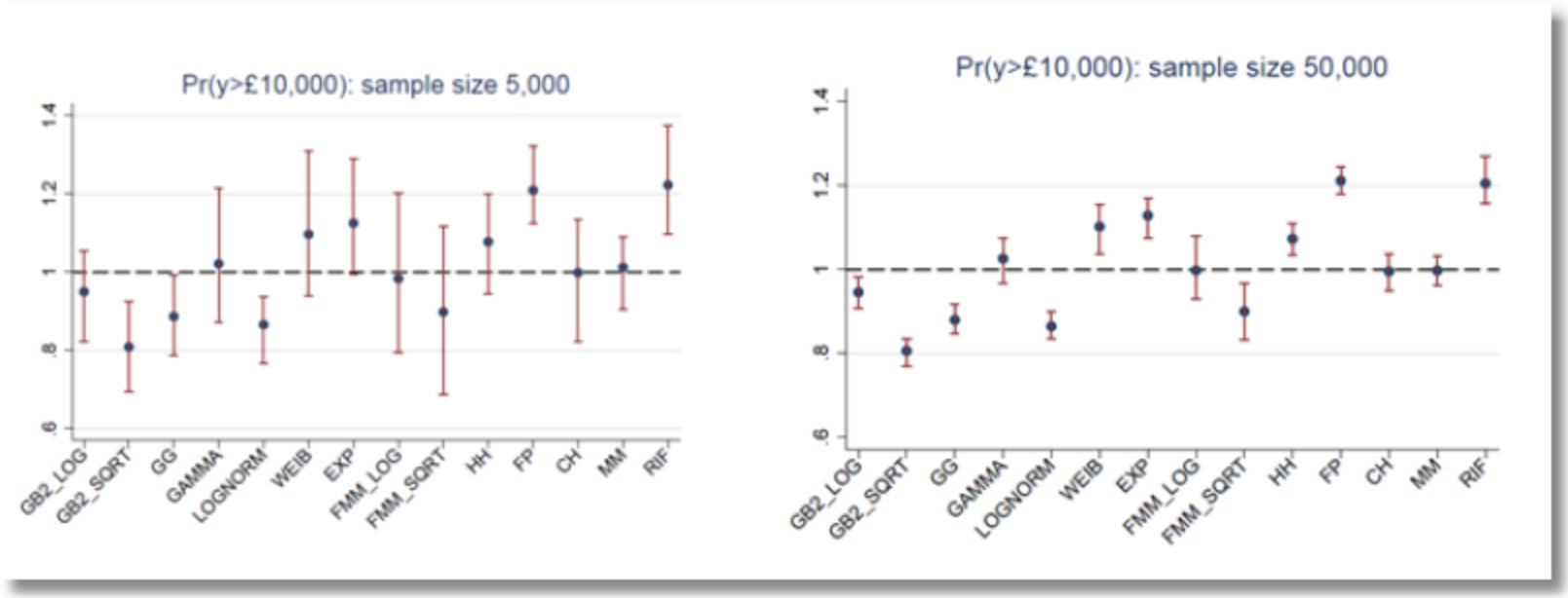
k	% observations in 'validation' set $> k$
£500	82.93%
£1,000	55.89%
£2,500	27.04%
£5,000	13.84%
£7,500	6.94%
£10,000	4.10%

Findings

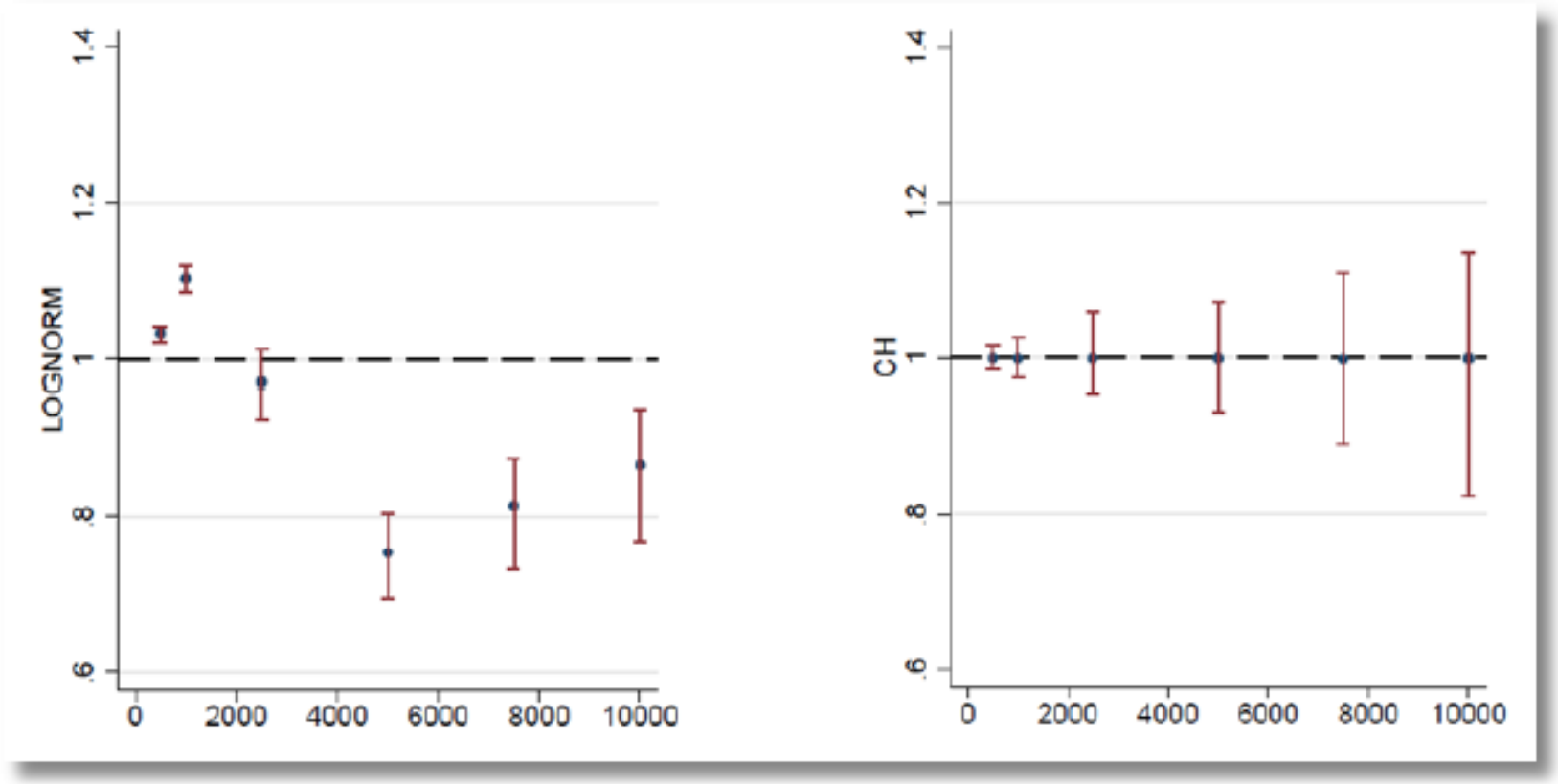


Method	Bias	Range
GB2_LOG	5th	6th
GB2_SQRT	12th	5th
GG	9th	4th
GAMMA	4th	11th
LOGNORM	11th	1st
WEIB	7th	12th
EXP	10th	9th
FMM_LOG	3rd	13th
FMM_SQRT	8th	14th
HH	6th	7th
FP	13th	3rd
CH	1st	10th
MM	2nd	2nd
RIF	14th	8th

Performance of Candidate Models



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