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

A. M. Jones[1] J. Lomas[2] N. Rice[1,2]

[1]Department of Economics and Related Studies
University of York

[2]Centre for Health Economics
University of York

Outline

Outline

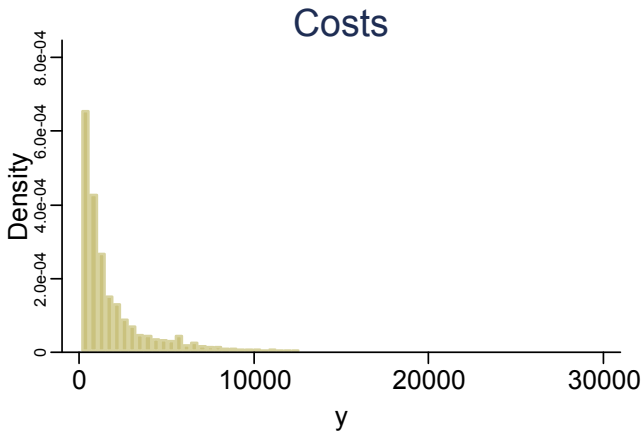
Why model costs?

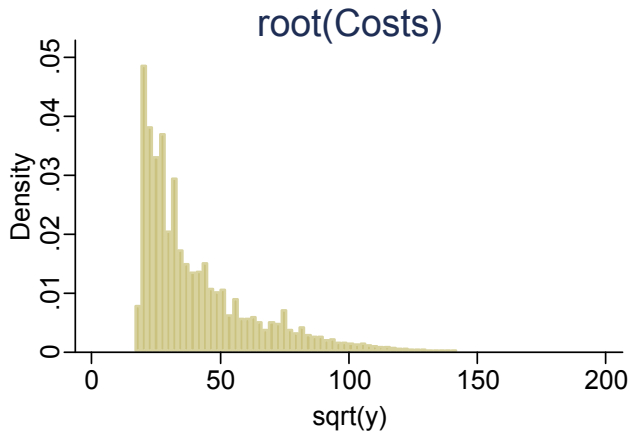
- cost-effectiveness analysis:
  - populating decision models
  - obtaining precise treatment effects (trial and observational data)
- risk-adjustment:
  - budgets for healthcare bodies (health authorities, GP consortia)
  - insurance companies
- attributable healthcare costs:
  - health behaviours (smoking, 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

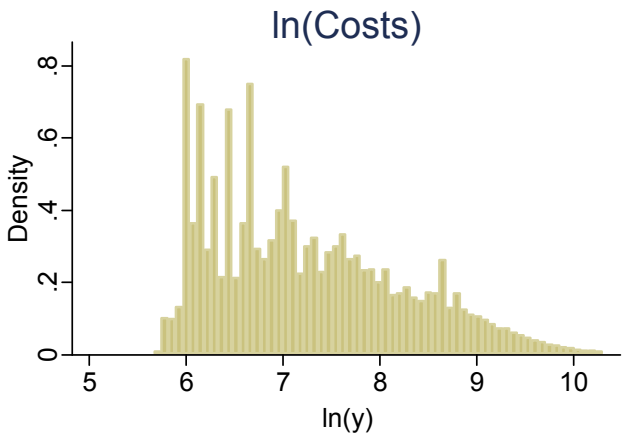
Challenges of healthcare cost data

- 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

Histogram plot of outcome variable







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

- Monte Carlo studies: Basu et al. (2004), Gilleskie and Mroz (2004), Manning et al (2005)
- Studies using cross-validation: Veazie et al (2003), Buntin and Zaslavsky (2004), Basu et al (2006), Hill and Miller (2010)
- Quasi-Monte Carlo studies: Deb and Burgess (2003), Jones et al (2013), Jones et al (2013)

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).
- Very interesting discussion of 'mean-based evaluation' in Vanness and Mullahy (2006).

'Beyond the mean' - existing techniques

- Methods developed for analysing features of the distribution beyond the mean, particularly in labour economics.
- Fortin N, Lemieux T, Firpo S. 2011. Decomposition methods in economics. Very useful source. Note we don't decompose! But use approaches for fitting a counterfactual distribution.
- 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

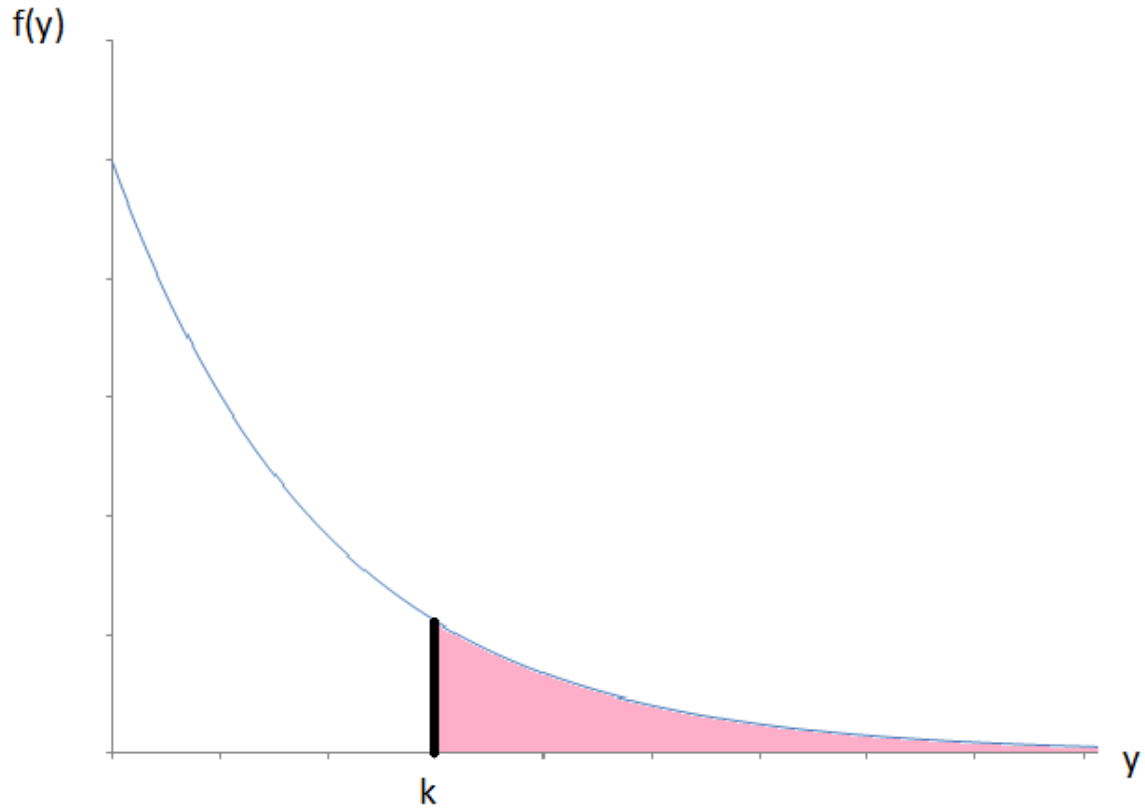
Included approaches

- 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 Chernuzhukov et al. (2013)
    - Using quantile function: Machado and Mata (2005) [+Melly (2005)] and Firpo et al. (2009): Recentred Influence regression (RIF regression)

Excluded approaches

- Note! Do not include linear regression (OLS) or GLM approaches.
- These can generate E(y|X) or Var(y|X), but not the full distribution...
- ...and so can't be used to give estimates of P(y>k)!

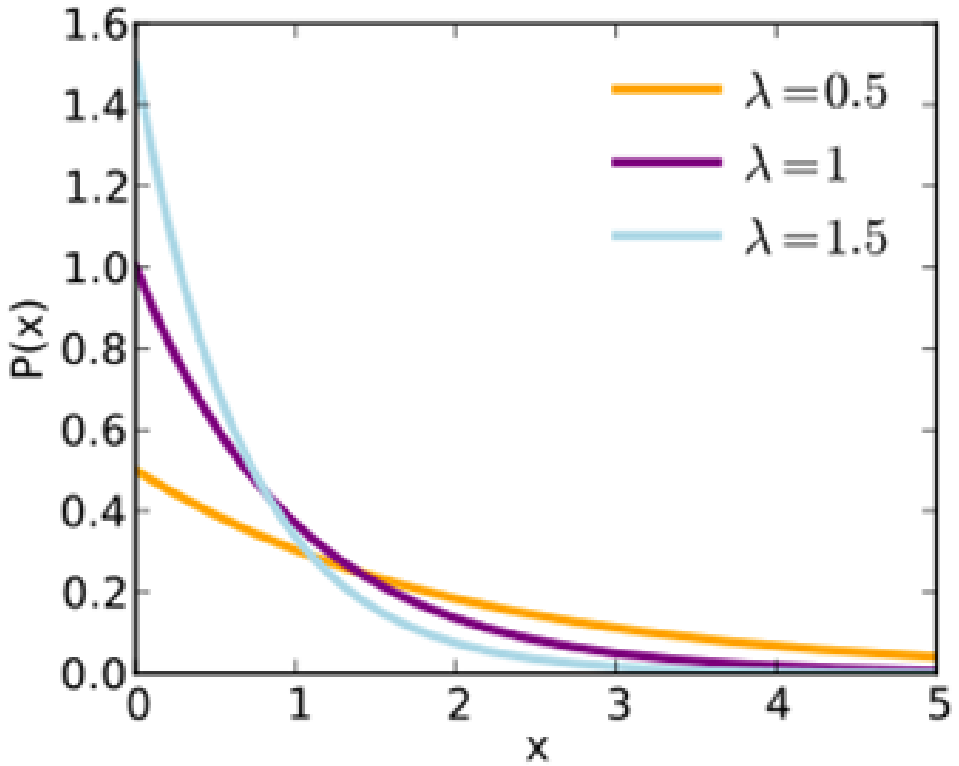
What we do for parametric methods - illustration (exponential distribution)



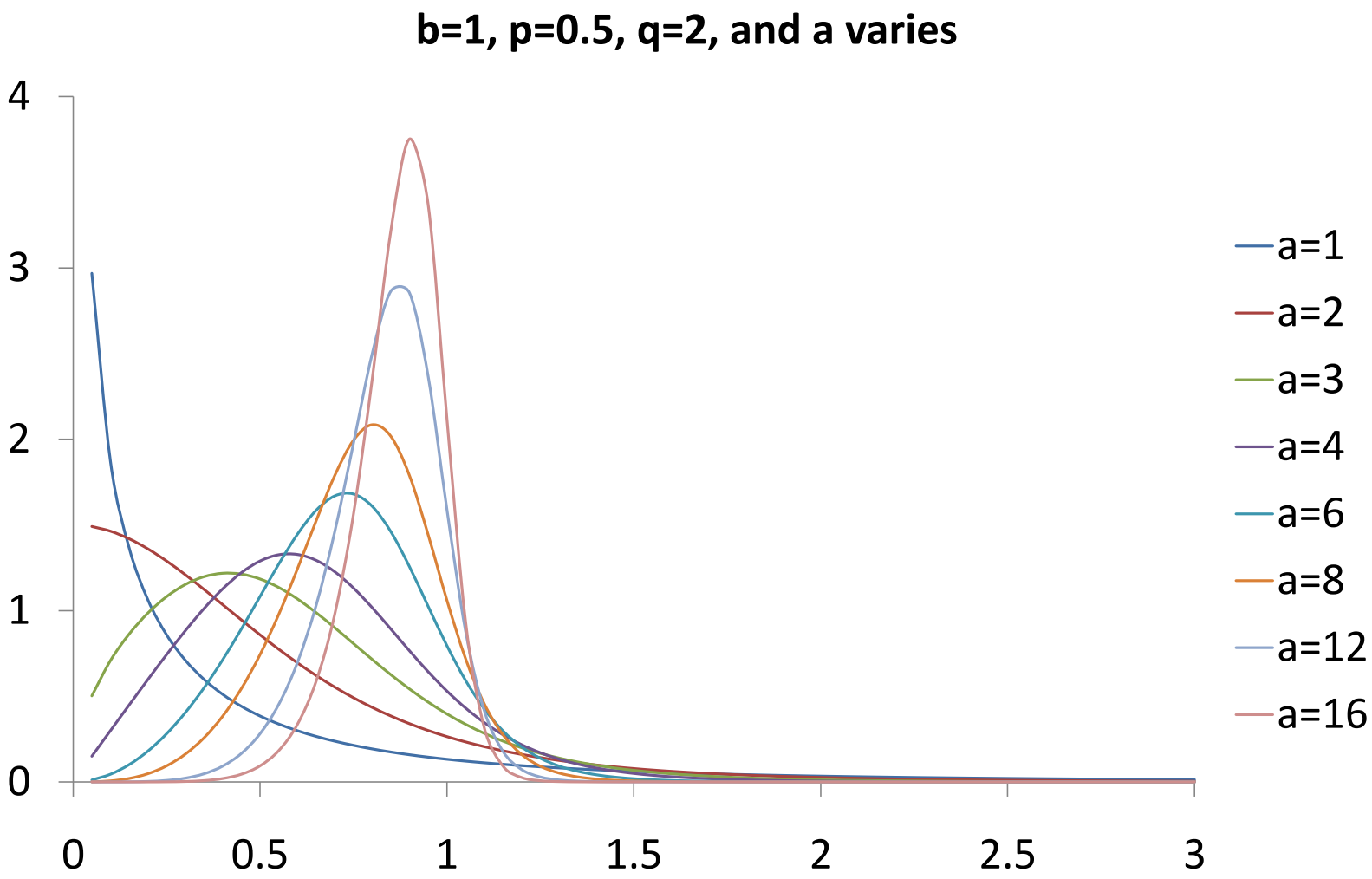
What we do for parametric methods - illustration (exponential distribution)

- $P(y > k) = \int_k^\infty f(y)dy = 1 - F(k) = \exp(-\lambda k)$

Exponential distribution - 1-parameter



GB2 distribution - 4-parameter



Generating a tail probability forecast

- For exponential, as illustration, $\lambda = \frac{1}{\exp(X\beta)}$
- Therefore $P(y > k|X) = \exp(-\frac{k}{\exp(X\beta)})$
- 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).

Estimating the cumulative distribution function

- Three approaches considered: Han and Hausman (1990) (HH), Foresi and Peracchi (1995) (FP) and Chernozhukov et al. (2013) (CH).
- 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.

FP and CH

- 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 $F(decile|X)$
- 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_τ, 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.

Estimating the quantile function

- Firpo et al. (2009) (RIF): Exactly the same as MM, but use RIF regressions as opposed to quantile regressions to estimate q_τ.
- 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.

Outline

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, apapted 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

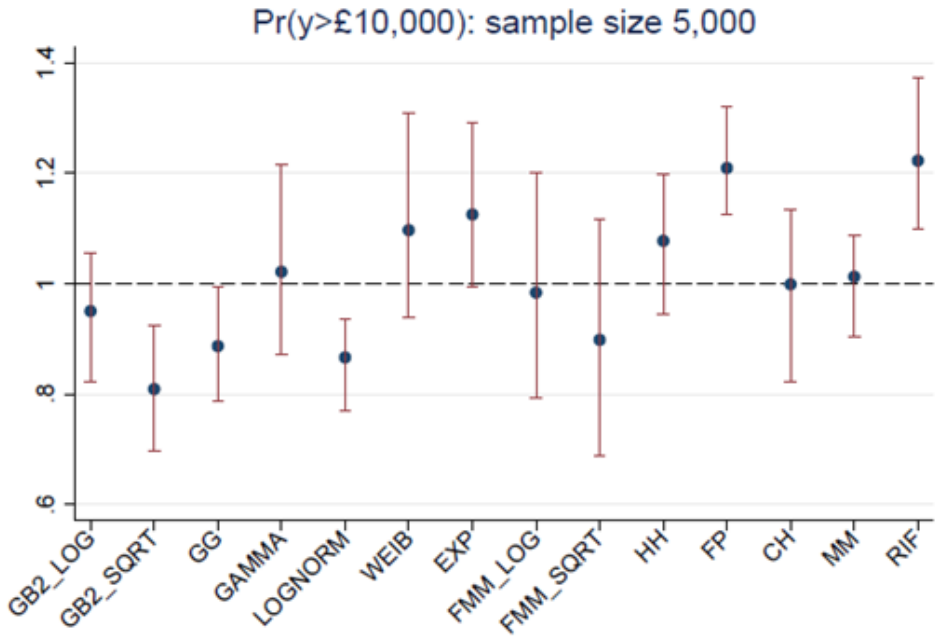
- For each model, and for each sample, calculate P(y>k) for every observation [and later also for subsets of population based on X values] in the 'validation' set and calculate average.
- Then compare this to observed proportion of observations with healthcare costs greater than 'k'
  - using ratio: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

Descriptive statistics

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%

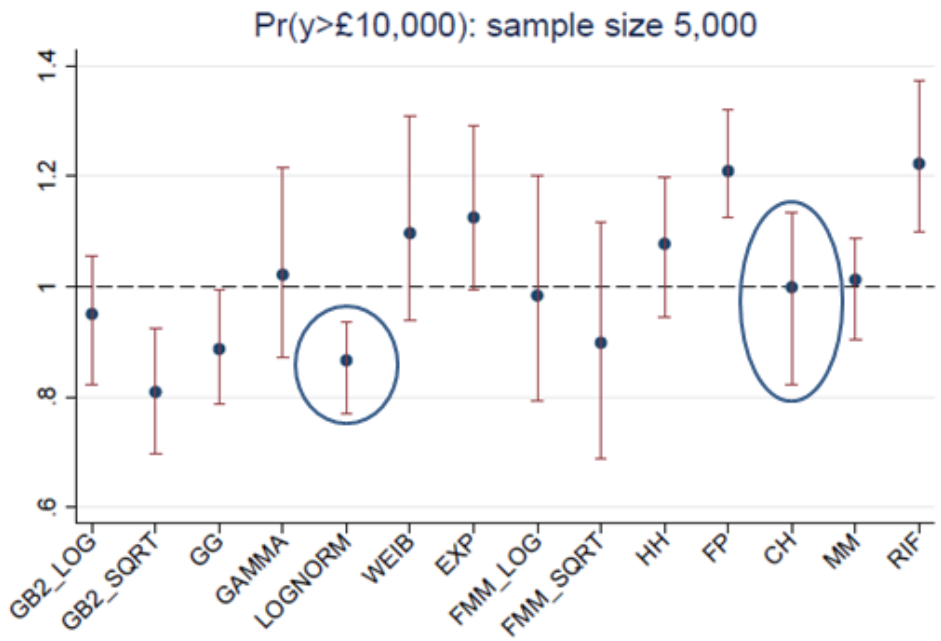
Outline

Results for k = 10,000



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

Results for k = 10,000

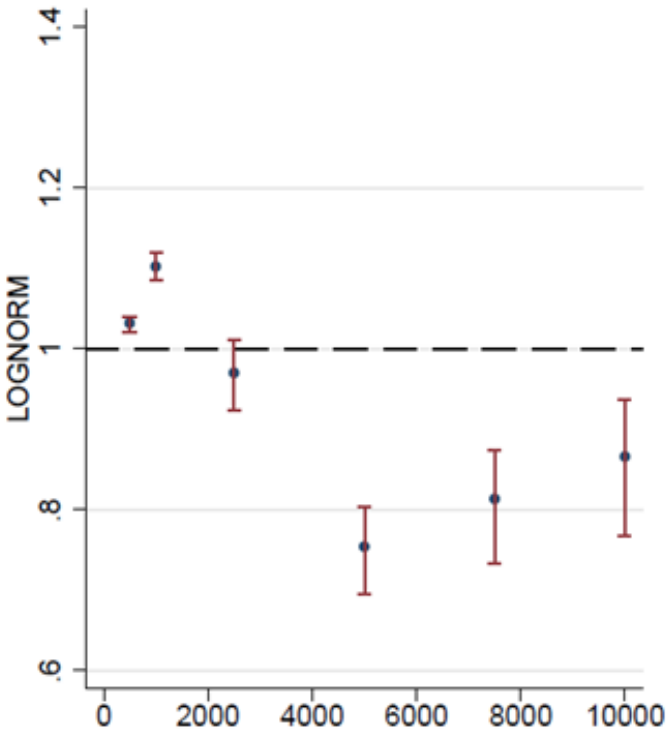


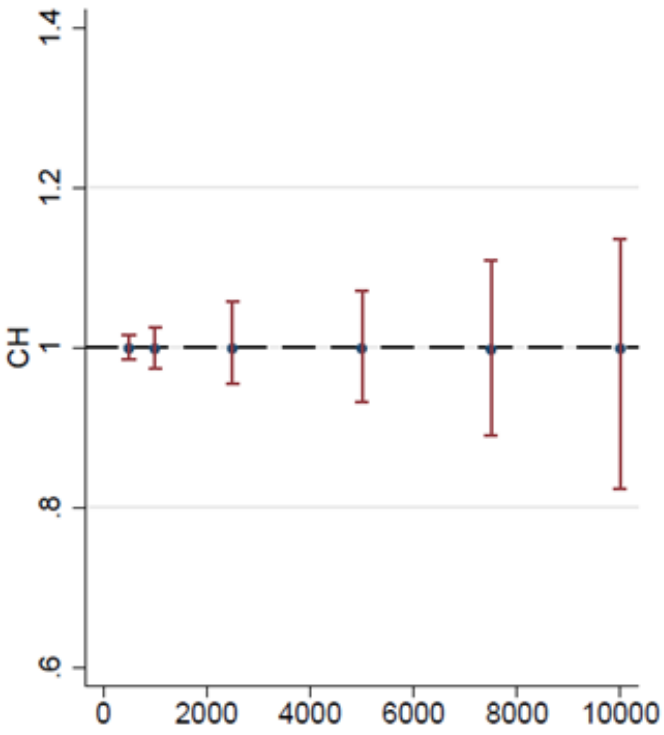
Method	Bias	Range
GB2_LOG	5th	6th
GB2_SQRT	12th	5th
GG	9th	4th
GAMMA	4th	11th
LOGNORM	11th	1st
WEIB	7th	12th
EXP	10th	9th
FMM_LOG	3rd	13th
FMM_SQRT	8th	14th
HH	6th	7th
FP	13th	3rd
CH	1st	10th
MM	2nd	2nd
RIF	14th	8th

Results for k = 10,000

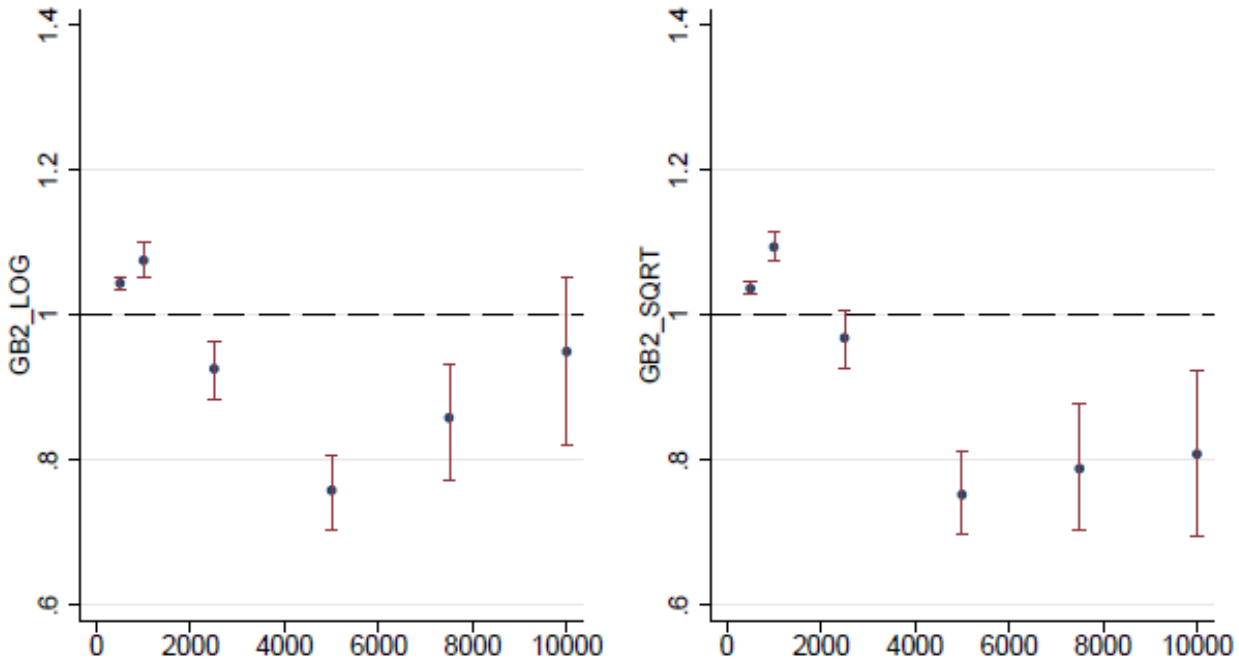
Method	Bias	Range	Standard deviation
GB2_LOG	6th	6th	6th
GB2_SQRT	13th	5th	3rd
GG	10th	4th	2nd
GAMMA	5th	12th	11th
LOGNORM	12th	1st	1st
WEIB	8th	13th	12th
EXP	11th	9th	8th
FMM_LOG	4th	14th	15th
FMM_SQRT	9th	15th	14th
HH	7th	7th	9th
FP	14th	3rd	4th
CH	2nd	10th	10th
MM	3rd	2nd	5th
RIF	15th	8th	7th
NAÏVE	1st	11th	13th

Results for different values of k

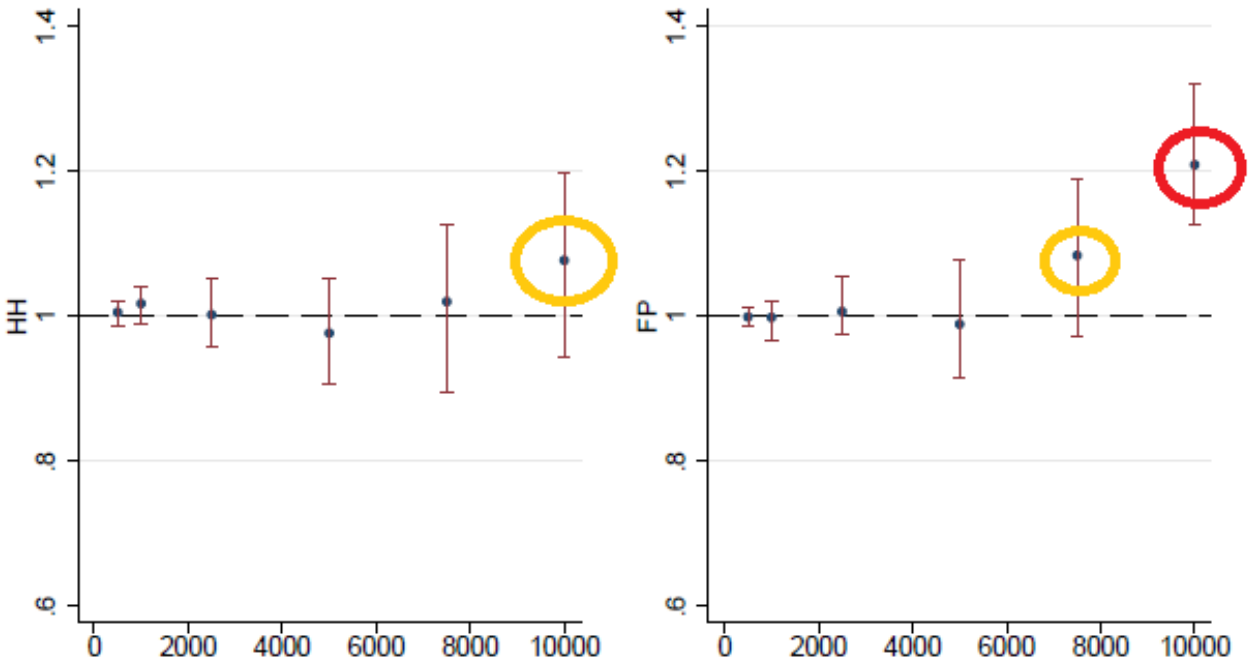




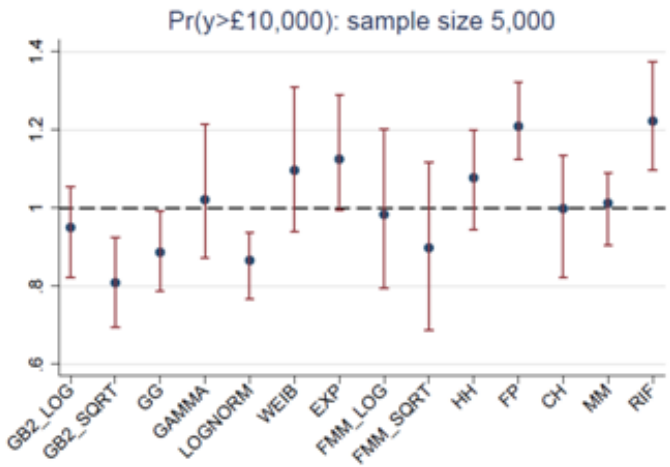
Results for different values of k

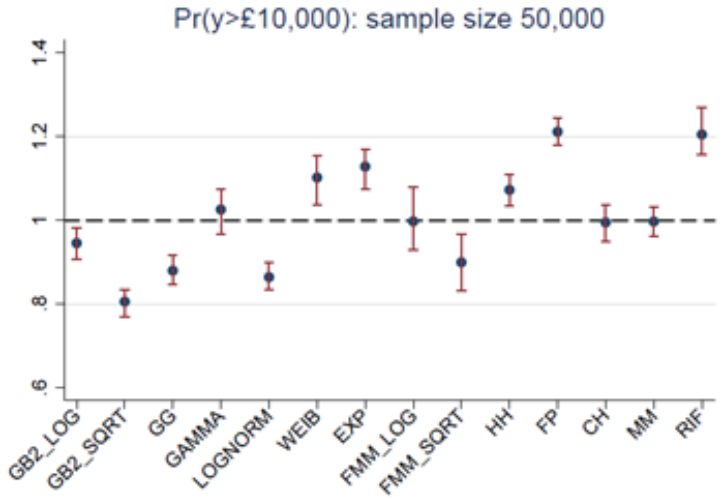


Results for different values of k



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 illiustration 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

	Decile number									
Method	1	2	3	4	5	6	7	8	9	10
Observed	0.107%	0.164%	0.287%	0.30%	0.52%	0.97%	1.76%	3.04%	6.82%	27.05%
GB2_LOG	0.467%	0.520%	0.630%	0.79%	1.01%	1.28%	1.73%	2.71%	5.38%	24.47%
GB2_SQRT	0.338%	0.372%	0.509%	0.71%	0.99%	1.34%	1.90%	3.08%	5.81%	18.14%
GG	0.330%	0.377%	0.473%	0.61%	0.80%	1.06%	1.49%	2.45%	5.10%	23.68%
GAMMA	0.000%	0.001%	0.003%	0.01%	0.03%	0.13%	0.43%	1.57%	6.37%	33.36%
LOGNORM	0.041%	0.059%	0.096%	0.16%	0.25%	0.43%	0.79%	1.73%	4.88%	27.11%
WEIB	0.001%	0.002%	0.010%	0.03%	0.07%	0.26%	0.74%	2.33%	7.91%	33.64%
EXP	0.005%	0.015%	0.046%	0.11%	0.21%	0.56%	1.29%	3.23%	8.83%	31.87%
FMM_LOG	0.014%	0.048%	0.132%	0.22%	0.32%	0.79%	1.49%	3.05%	6.91%	27.38%
FMM_SQRT	0.015%	0.045%	0.178%	0.33%	0.44%	1.11%	2.00%	3.88%	7.72%	21.13%
HH	0.407%	0.463%	0.590%	0.75%	0.98%	1.30%	1.83%	3.02%	6.39%	28.48%
FP	0.341%	0.499%	0.584%	0.68%	1.09%	1.43%	2.37%	3.80%	7.69%	31.14%
CH	−3.258%	−2.481%	−2.052%	−1.01%	0.88%	2.09%	4.27%	6.86%	11.34%	24.34%
MM	0.006%	0.018%	0.046%	0.09%	0.21%	0.49%	1.13%	2.59%	6.69%	30.23%
RIF	0.159%	0.264%	0.378%	0.59%	1.16%	1.82%	3.49%	6.58%	13.08%	22.57%

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

Outline

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.