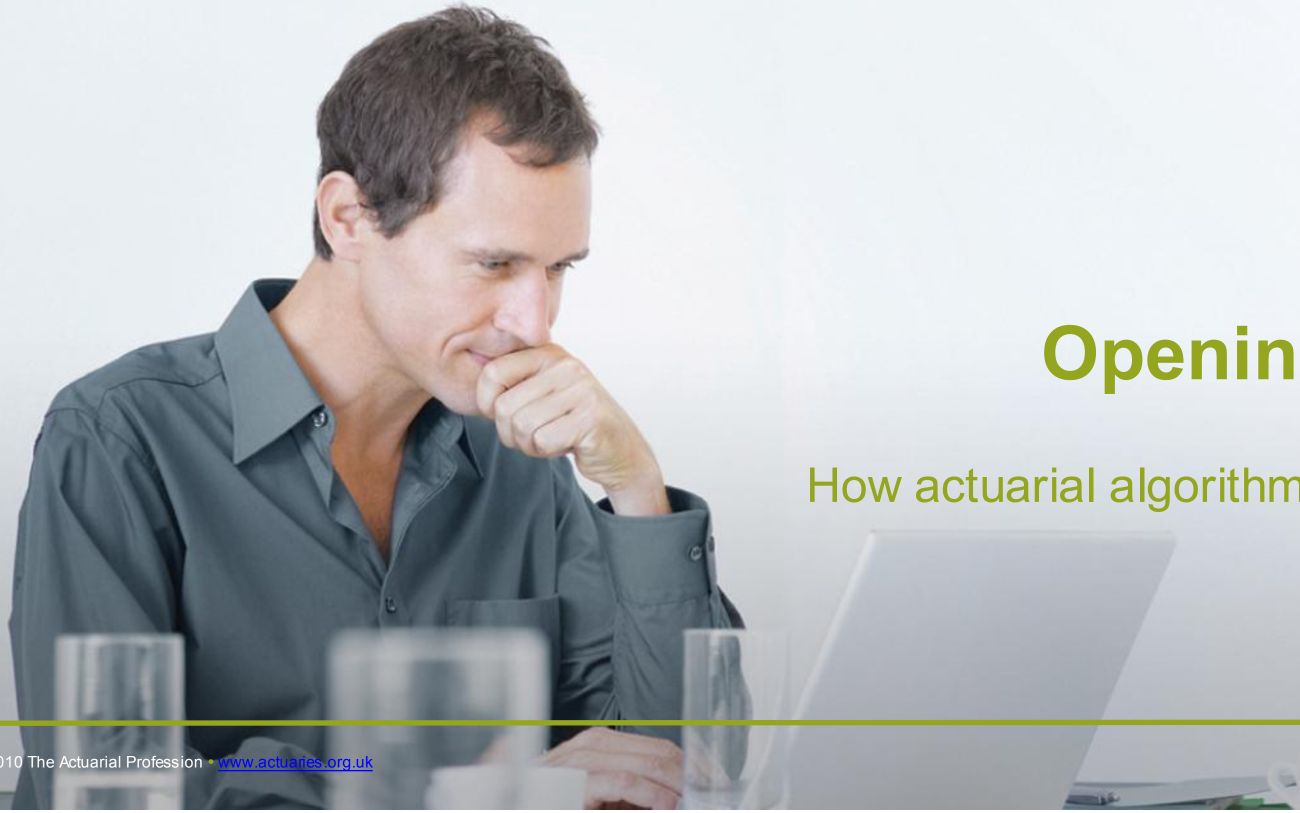
The Actuarial Profession
making financial sense of the future

GIRO conference and exhibition 2010
James Tanser and Richard Bland



Opening the Black Box

How actuarial algorithms work and sometimes fail

12-15 October 2010

Opening the Black Box

- Why GLMs go wrong
- Identifying the bottlenecks
- Expensive options
- Simulation
- Bootstrapping

The well-known GLM formula

$$E[\underline{Y}] = \underline{\mu} = g^{-1}(\mathbf{X}.\underline{\beta} + \underline{\xi})$$

$$\mathrm{Var}[\underline{Y}] = \phi.V(\underline{\mu}) \,/\, \underline{\omega}$$

- Maximum Likelihood Estimation – maximise the log likelihood of $\underline{\xi}$ with respect to $\underline{\beta}$

What can go wrong?

- What if the Newton-Raphson doesn't work because the partial differential matrix is singular?
  - $\underline{\beta}_{n+1} = \underline{\beta}_n - \mathbf{H}^{-1}.\underline{s}$
  - $\underline{s}$ is the first differential of the log likelihood
  - **H** is the second differential – the Hessian

Aliasing

- A column in the data matrix is linearly dependent on the others

				Selected base		
	Exposure: # Doors →	2	3	~~4~~	5	Unknown
	Colour ↓					
Selected base	~~Red~~	13,234	12,343	13,432	13,432	0
	Green	4,543	4,543	13,243	2,345	0
	Blue	6,544	5,443	15,654	4,565	0
	Black	4,643	1,235	14,565	4,545	0
Further aliasing	~~Unknown~~	0	0	0	0	3,242

Undefined parameters

- The model is trying to estimate log(0) for one or more parameters.

Exposure: # Doors →	2	3	Selected base ~~4~~	5	Unknown
Colour ↓					
Selected base ~~Red~~	13,234	12,343	13,432	13,432	0
Green	4,543	4,543	13,243	2,345	0
Blue	6,544	5,443	15,654	4,565	0
Black	4,643	1,235	14,565	4,545	2
Unknown	0	0	0	0	3,242

Do undefined parameters matter?

- NO
  - Because most of the fitted values are reasonable

- YES
  - Because some fitted values will have undefined or extreme values
  - If unfixed, some policyholders could have near zero premiums

Opening the Black Box

- Why GLMs go wrong
- Identifying the bottlenecks
- Expensive options
- Simulation
- Bootstrapping

Bottlenecks …
… or why it is unwise to go modelling with a laptop

- Three steps to model a GLM:
  - Construct the data matrix (with some optimisations)
  - Calculate the differential matrices and add them up
  - Invert and calculate new parameter estimates

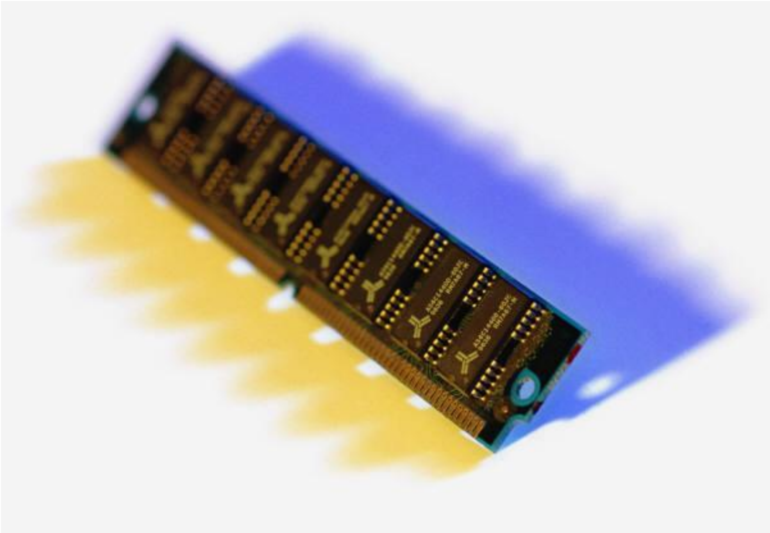
- This uses:
  - Lots of CPU time
  - Lots of memory
  - … and there may also be large files lying around

Multithreading

- Adding up the differentials is a commutative, distributive operation across records
  - So you can multithread it
- The latest desktop PCs now have up to 6 cores / 12 threads

Memory management

- The data matrix is large, but can be condensed.
- So it is vital to use software which optimises the storage of the data matrix, and has its own memory management to handle very large storage requirements.



Large data files

- Storage of data is often overlooked in system design
- Standard corporate networking is insufficient
- Magnetic storage for PCs has highly variable performance – optimal performance requires multiple disks

Opening the Black Box

- Why GLMs go wrong
- Identifying the bottlenecks
- Expensive options
- Simulation
- Bootstrapping

Expensive options …
… and cheats

- Some of the statistical significance tests require the fitting of multiple models (sometimes many models)

- Some explanatory variables require more CPU / more memory to fit in the GLM
  - What’s in a variate?

Cheats …
… what if I don’t actually need an accurate answer?

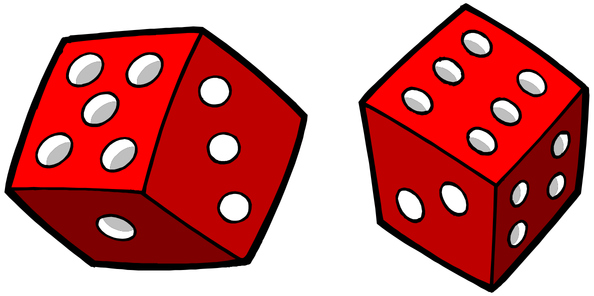
- Non-MLE GLMs
- Approximate significance tests
- Fitting models to fitted values

Opening the Black Box

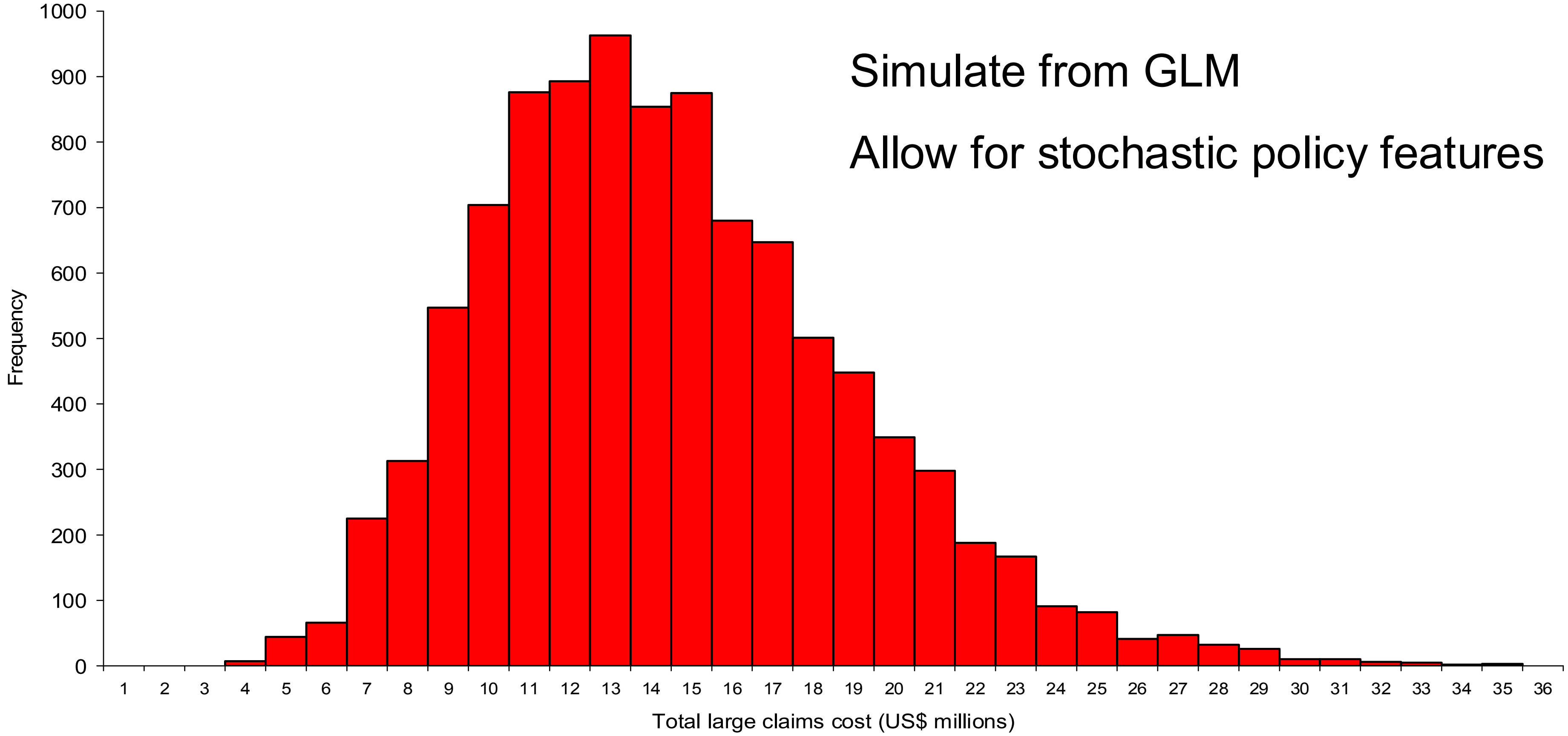
- Why GLMs go wrong
- Identifying the bottlenecks
- Expensive options
- Simulation
- Bootstrapping

Simulation

- Maths is hard, but random numbers are easy!
- Interesting but hard problems:
  - Aggregate deductibles
  - Percentile pricing
  - Effect of reinsurance
  - Proportional excess (with collar and cap)



Simulation from GLMs

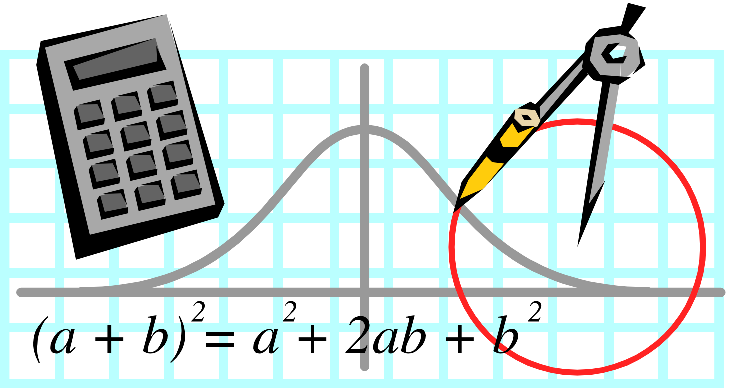


Simulate from GLM

Allow for stochastic policy features

Fit for purpose?

- Have you modelled the mean or the whole distribution?
  - Over/under dispersion & the scale parameter
  - Estimates and Maximum Likelihood Estimates
- GIGO
- If you have a hammer…
  - The exponential family
- Bayes and MCMC…



Opening the Black Box

- Why GLMs go wrong
- Identifying the bottlenecks
- Expensive options
- Simulation
- Bootstrapping

Bootstrapping

- Fantastic theory!
  - IID
  - Resample with replacement
  - Infinite amounts of data available?
- Original use to estimate variation in the mean
  - Samples same size as original sample
- Useful where you have
  - Lots of data!
  - Unknown or non-standard distribution

How far can you levitate?

- Are samples really IID?
- Only get out what you started with
  - Mean, variance, …
- Interesting but hard problems (again)

Questions or comments?

Expressions of individual views by members of The Actuarial Profession and its staff are encouraged.

The views expressed in this presentation are those of the presenter.